Estadística para Biología y Ciencias de la Salud

J. SUSAN MILTON

3.ª edición ampliada

McGRAW-HILL • INTERAMERICANA

ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD

ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD

3.ª edición ampliada

J. Susan Milton
Universidad de Radford

Incluye: **MÉTODOS ESTADÍSTICOS CON STATGRAPfflCS Y SPSS**

Agustín Turrero y Pilar Zuluaga
Universidad Complutense de Madrid

McGRAW-HILL • INTERAMERICANA

MADRID • BUENOS AIRES • CARACAS • GUATEMALA • LISBOA • MÉXICO
NUEVA YORK • PANAMÁ • SAN JUAN • BOGOTÁ • SANTIAGO • SAO PAUBLO
AUCKLAND • HAMBURGO • LONDRES • MILÁN • MONTREAL • NUEVA DELI • PARIS
SAN FRANCISCO • SYDNEY • SINGAPUR • ST. LOUIS • TOKIO • TORONTO

Traducción

DIEGO DELGADO CRESPO
JUAN LLOVET VERDUGO
JULIÁN MARTÍNEZ VALERO

Profesores del Departamento de Matemáticas
Facultad de Ciencias
Universidad de Alcalá, Madrid

Apéndices C y D: Métodos estadísticos con STATGRAPHICS y SPSS

AGUSTÍN TURRERO NOGUÉS
PILAR ZULUAGA ARIAS

Doctores en Ciencias Matemáticas
Profesores Titulares del Departamento de Estadística e Investigación Operativa
Facultad de Medicina
Universidad Complutense de Madrid

ESTADÍSTICA PARA BIOLOGÍA Y CIENCIAS DE LA SALUD

Tercera edición, 2001
Primera reimpresión, 2002
Segunda reimpresión, 2004
Tercera reimpresión, 2006
Tercera edición 2001, ampliada en 2007

ISBN: 978-84-481-5996-2
Depósito legal: M-7176-2007

Traducido de la tercera edición del inglés de la obra:
STATISTICAL METHODS IN THE BIOLOGICAL AND HEALTH SCIENCES
de J. SUSAN MILTON

ISBN: 0-07-290148-9 (Edición original)

Preimpresión: MonoComp, S. A. Cartagena, 43. 28028 Madrid
Impreso en: Inmagrag.S.L.
IMPRESO EN ESPAÑA - PRINTED IN SPAIN

PRÓLOGO A LA 3.ª EDICIÓN EN ESPAÑOL AMPLIADA

La creciente utilización de recursos informáticos, tanto por estudiantes como por investigadores, hace interesante conocer la resolución de problemas estadísticos con un *software* adecuado. Este conocimiento resulta imprescindible para el análisis de las complejas bases de datos que se manejan habitualmente en la investigación de Ciencias de la Salud. Por ello, al final de cada capítulo del libro se pueden encontrar las sentencias correspondientes del paquete estadístico SAS que realizan lo allí expuesto.

No obstante, hemos creído que los paquetes estadísticos ejecutables por medio de menús son más accesibles para un usuario del entorno Windows. Por esta razón, hemos añadido los Apéndices C y D, que incorporan el análisis y la interpretación de problemas estadísticos con STATGRAPHICS Plus, versión 5.1 (www.statgraphics.com), y SPSS, versión 12.0 (www.addlink.es), respectivamente. Ambas versiones son de fácil manejo y muy versátiles desde el punto de vista de los resultados, numéricos y gráficos, que podemos obtener.

Los Apéndices C y D están divididos en 9 bloques que abarcan todas las técnicas estadísticas desarrolladas en los 13 capítulos del libro, incluyendo, en algunos casos, la presentación de métodos alternativos de análisis, no contenidos en dichos capítulos. En cada uno de los 9 bloques se hace referencia a los capítulos donde puede encontrarse con detalle la base teórica de los métodos estadísticos allí referidos. Los contenidos de los bloques se corresponden entre ambos Apéndices. Esta estructura permite comparar el tratamiento que de un mismo problema hacen sendos paquetes. En algunas ocasiones resultará indiferente resolver un problema con uno u otro *software,* pero en otras dispondremos de técnicas complementarias, numéricas o gráficas, que serán de utilidad para la mejor y más completa resolución de los problemas.

La exposición de una técnica estadística en cualquiera de los bloques obedece al siguiente itinerario:

- Se presenta una base de datos y sobre ésta un problema a resolver.
- Se identifica la técnica estadística adecuada.
- Se elige el procedimiento del paquete estadístico que ejecuta dicha técnica.
- Se enseña el manejo de dicho procedimiento con la ayuda de ventanas y pantallas del propio programa.
- Se obtienen los resultados, que se presentan mediante tablas y gráficas.
- Finalmente, se interpretan dichos resultados añadiendo sugerencias sobre cómo proseguir el análisis o llamadas de atención cuando se vulnera alguna hipótesis relevante de la técnica utilizada.

La estructura formal de los Apéndices C y D está concebida para un uso independiente del texto principal. El lector con conocimientos previos de estadística no necesitará acudir a dicho texto para acometer el análisis estadístico deseado; le bastará con identificar la técnica adecuada, buscar inmediatamente el apartado correspondiente del STATGRAPHICS (Apéndice C) o SPSS (Apéndice D) y seguir la secuencia de procedimientos que allí se detallan Para aquellos lectores sin conocimientos previos de estadística, el camino puede ser similar las bases de datos suministradas sugieren diferentes análisis estadísticos que pueden servir de guía para identificar el análisis adecuado o sugerir otros nuevos. Los propios comentarios y la interpretación de resultados ayudarán, a ambos tipos de lectores, a clarificar y entender el método de análisis.

Por último, los problemas estadísticos tratados son de tres tipos:

- Ejemplos resueltos en el texto que se replican con los paquetes. El objetivo en estos casos es identificar los procedimientos adecuados y ver las posibles ampliaciones que dichos programas ofrecen.
- Ejercicios propuestos en el texto. Además de los objetivos anteriores, se aportan las soluciones.
- Cuestiones relativas a dos nuevas bases de datos que se presentan en los bloques C2 *y* C6, la primera referida a la supervivencia de mujeres con cáncer cervicouterino, y la segunda al crecimiento fetal. Estas bases, por el amplio número de datos y variables que contienen, permiten enfocar los problemas desde un punto de vista más realista. Generaímente, para abordar la resolución de un problema real es necesario emplear varias técnicas estadísticas, dependiendo la elección de algunas de ellas de los resultados obte nidos por las precedentes. Por ello, cada una de estas dos bases de datos se utilizarán para ilustrar los procedimientos contenidos en varios bloques.

AGUSTÍN TURRERO y PILAR ZULUAGA

ACERCA DEL AUTOR

J. Susan Milton es profesora de estadística en la Universidad de Radford. La Dra. Milton obtuvo el grado de Bachelor of Science en la Western Carolina University, el de Master of Arts de la University of North Carolina en Chapel Hill y el Ph.D. en estadística en el Virginia Polytechnic Institute y la State University. Ostenta el cargo de Danforth Associate y ha recibido el Radford University Foundation Award for Excellence in Teaching. Ha publicado *Introduction to Probability and Statistics: Principies and Applications for Engineering and the Computing Sciences,* así como *Introduction to Statistics, Probability with the Essential Analysis, Applied Statistics with Probability,* y *A First Course in the Theory of Linear Statistical Models.*

A mi familia:

Joan y Tom Savage
Enid Milton
Stephanie y David Savage
Deborah Savage y Tim Woolf

CONTENIDO

PROLOGO

Se ha hecho ya evidente que la interpretación de muchas de las investigaciones en las Ciencias Biológicas y de la Salud dependen en gran parte de los métodos estadísticos. Por esta razón, es esencial que los estudiantes de estas áreas se familiaricen lo antes posible, en sus carreras, con los razonamientos estadísticos. Este libro se entiende como un *primer* curso sobre los métodos estadísticos para estudiantes de Biología y Ciencias de la Salud, aunque también puede ser empleado de forma ventajosa por estudiantes ya licenciados, con escasa o ninguna experiencia en métodos estadísticos.

El libro no es un recetario estadístico ni tampoco un manual para investigadores. Pretendemos encontrar un camino intermedio que proporcione al estudiante una comprensión de la lógica empleada en las técnicas estadísticas así como su puesta en práctica. No se requieren conocimientos previos de matemáticas. El lector con una base adecuada de álgebra elemental será capaz de seguir los argumentos presentados.

Hemos elegido ejemplos y ejercicios específicamente pensados para estudiantes de Ciencias Biológicas y de la Salud. Se han tomado éstos de la Genética, la Biología general, la Ecología y la Medicina. Y, excepto donde se indique, los datos son simulados. En todo caso, la simulación está hecha con cuidado, de modo que los métodos de análisis sean consistentes con lo puesto de manifiesto por investigaciones recientes. De esta forma, el estudiante se hará una idea de los tipos de problemas que interesan en los trabajos actuales propios de las Ciencias Biológicas. Muchos ejercicios se dejan incompletos con la esperanza de estimular algunas discusiones en clase.

Se supone que el estudiante tiene acceso a algún tipo de calculadora electrónica. En el mercado existen muchas marcas y modelos, y la mayor parte tiene incorporadas funciones estadísticas. Recomendamos el uso de estas calculadoras, dado que con ello se permite al estudiante concentrarse en la interpretación del análisis, más que en los cálculos aritméticos. En el texto se dan las instrucciones para utilizar la calculadora TI83. Ésta, que es relativamente nueva en el mercado, permite realizar la mayoría de las técnicas presentadas en el libro. Pueden obtenerse, además, muchos de los intervalos de confianza descritos y la mayor parte de las tablas estadísticas mostradas en el manuscrito.

Queremos hacer hincapié en el hecho de que muchos de los conjuntos de datos aquí presentados son más bien *pequeños,* para que el estudiante no se abrume por el aspecto puramente operativo del análisis estadístico. Ello no implica que las muestras pequeñas de

datos sean aceptables en la investigación biológica. De hecho, la mayor parte de los principales programas de investigación implican una tremenda inversión de tiempo y dinero, y el resultado es un número elevado de datos. Tales datos invitan por sí mismos al análisis por medio del ordenador. Por esta razón, incluimos algunas instrucciones en la interpretación de las *salidas* o *outputs* del ordenador. El paquete elegido con fines ilustrativos es el SAS *(Statistical Analysis System:* SAS Institute, Inc., Raleigh, North Carolina). Ello se debe a su popularidad y fácil manejo. No pretendemos suponer que sea superior a otros productos bien conocidos, tales como el SPSS *(Statistical Package for the Social Sciences),* el **BMD (***Biomedical Computer Programs,* University of California Press) o el MINITAB (Duxbury Press). A final de algunos capítulos, en la sección Herramientas computacionales, se incluye una introducción al SAS, junto con el código de programa necesario para generar la salida dada en el texto.

Esta es una revisión sustancial de la segunda edición del libro. En muchos apartados del texto, se han incorporado los comentarios sugeridos por distintos revisores del mismo, para reforzar las exposiciones presentadas. Se han añadido, igualmente, nuevos ejercicios. Al final de muchos capítulos se ha incluido una sección, Herramientas computacionales, para introducir la programación en SAS y la calculadora gráfica TI83. También se han incluido nuevas aportaciones como los diagramas de tallos y hojas adosados, una comparación de varianzas efectuada de forma muy simple y una tabla *T* ampliada. En el texto continúa teniendo un papel importante el modo de hallar e interpretar los valores *P.*

A partir de este libro se pueden impartir distintos cursos. Su extensión en el tiempo puede variar desde un semestre hasta un año. Es difícil determinar exactamente la materia que puede ser cubierta en un tiempo dado, puesto que ello está en función del tamaño de la clase, de la madurez académica de los estudiantes y de las inclinaciones del profesor. En todo caso, ofrecemos algunos criterios para el uso del texto en los resúmenes de los capítulos.

Capítulo 1. Este capítulo es una introducción a la Estadística descriptiva. Se presentan, y pronto se diferencian, las nociones de población y muestra, en las que se hace especial hincapié. Se han añadido, además, los temas de análisis exploratorio de datos (EDA), los diagramas de tallo y hojas, así como los de cajas. También se remarca la importancia que tiene el hecho de evaluar la forma, la posición y la variabilidad.

Capítulo 2. En él se introduce la probabilidad desde un punto de vista intuitivo. Se hace hincapié en los diagramas de árbol y su utilización en la resolución de problemas de Genética. Se dan técnicas de conteo (combinatoria) en relación con problemas de cálculo de probabilidades mediante el método clásico. Si el tiempo es insuficiente para tratar todo el capítulo, sugerimos que se vean las Secciones 2.1. y 2.2.

Capítulo 3. Este capítulo comprende los axiomas de la probabilidad, además de los teoremas que se deducen de los axiomas. También se encuentran en el capítulo los temas de independencia, probabilidad condicionada y teorema de Bayes. Una sección titulada «Tests de diagnóstico y riesgo relativo», ofrece aplicaciones de la probabilidad condicionada, parti - cularmente interesantes para los estudiantes de Medicina (y disciplinas afines). Este capítulo puede «saltarse» si el tiempo no permite su estudio.

Capítulo 4. Este capítulo desarrolla únicamente las variables aleatorias discretas, introduciendo los conceptos de densidad, distribución acumulada y esperanza.

Capítulo 5. En este capítulo se exponen, de forma paralela, las ideas presentadas en el Capítulo 4, pero aplicadas a variables aleatorias continuas. También se incluye un subapartado con la regla de probabilidad normal y sus aplicaciones en gráficos médicos.

Capítulo 6. En él hablamos de la estimación de la media, puntual y por intervalos, así como el contraste de hipótesis respecto a valores de este parámetro. Se incluye una sección sobre muestreo aleatorio y aleatorización. Además, se explica pormenorizadamente el uso del

valor *P,* algo en lo que se incide en el resto del texto. Finalmente, se añade una sección sobre el efecto del tamaño de la muestra en la amplitud del intervalo de confianza y en la potencia de un contraste.

Capítulo 7. Este es un capítulo breve sobre inferencias sobre la varianza y la desviación típica de una variable aleatoria. Se ha simplificado la exposición sobre la comparación de varianzas, incluyendo una regla práctica para comprobar la igualdad. El contraste formal *F* se incluye todavía en el texto.

Capítulo 8. En el Capítulo 8 se comentan las inferencias sobre una proporción y las comparaciones entre dos proporciones, con el Teorema central del límite, utilizado con el fin de justificar las técnicas ya empleadas.

Capítulo 9. En este capítulo comparamos dos medias, mediante estimación puntual y estimación por intervalos, y mediante contrastes de hipótesis. Se exponen los contrastes preliminares *F* para comparar varianzas. Se explican los métodos para comparar medias de muestras independientes: el de varianza conjunta y el de Smith-Satterthwaite. Se incluye una exposición sobre el modo de utilizar un paquete informático comercial para realizar estos contrastes. El capítulo concluye con una sección sobre datos emparejados.

Capítulo 10. En este capítulo, se introducen las técnicas utilizadas al comparar las medias de más de dos poblaciones, incluyendo comentarios sobre el modelo de clasificación de una vía, los bloques aleatorizados y el modelo de clasificación de dos vías. Se incluye una exposición sobre la eficacia de la construcción de bloques y el contraste *T* de Bonferroni, para hacer comparaciones por parejas. A lo largo de todo el capítulo, se incluyen notas sobre los cálculos.

Capítulo 11. Este capítulo explica con cierto detalle la regresión lineal simple y la correlación. Se ha añadido una nueva sección sobre regresión múltiple.

Capítulo 12. Se consideran en él los problemas relativos a los datos categóricos, haciendo especial hincapié en las pruebas de independencia y pruebas de homogeneidad en las tablas 2 x 2 y *r* x c.

Capítulo 13. Este capítulo presenta alternativas de distribución libre a los métodos clásicos ya vistos en capítulos anteriores. Incluye nuevos apartados sobre la prueba de Lilliefors de normalidad, el de Bartlett para la igualdad de varianzas, así como un contraste binomial para proporciones, en el caso de muestras pequeñas.

Ya que muchos cursos de este nivel duran un semestre, y es difícil abarcar el texto entero en ese tiempo, pueden omitirse las secciones etiquetadas como «opcional».

Deseo agradecer a Maggie Rogers, Bill Barter y Cathy Smith su aliento y consejo durante la revisión de este texto, y a Joann Fisher el mecanografiado del manuscrito. Mi agradecimiento también a Tonya Porter, por su ayuda en la preparación de las soluciones del manual. Quiero igualmente reconocer a Joan Savage y Charlene Lutes su ayuda como asesoras biológicas. Finalmente, gracias muy especiales, por sus numerosas sugerencias a lo largo de la revisión del original, a las siguientes personas: Charles M. Biles, Ph.D., Humboldt State University; John E. Boyer, Jr., Kansas State University; Annette Bucher, Colorado State University; Christiana Drake, University of California; Dr. R. K. Elswick, Jr., Medical College of Virginia, Virginia Commonwealth University; Thomas J. Glover, Hobart and William Smith Colleges; Golde I. Holtzman, Virginia Tech (VPI); Mark Krailo, University of Southern California; Benny Lo, NW Polytechnic University; Christopher Morrell, Loyola College; Lisa Sullivan, Boston University; Andrew Jay Tierman, Saginaw Valley State University; Mark S. West, Auburn University; y Robert F. Woolson, Ph.D., The University of Iowa.

J. SUSAN MILTON

Métodos descriptivos

La estadística se ha convertido en una herramienta indispensable para la mayoría de los científicos. ¿Qué es la estadística y cómo pueden utilizarse las técnicas estadísticas para responder a las cuestiones prácticas propuestas por los científicos?

Se ha definido la estadística como el arte de la decisión frente a la incertidumbre. Comenzaremos describiendo un problema típico que requiere una solución estadística y utilizaremos este ejemplo para introducir parte del lenguaje subyacente al campo de la estadística. Los términos se usan aquí a nivel intuitivo. Se definirán después, más rigurosamente, cuando surja la necesidad.

Un investigador, estudiando una enfermedad cardíaca en personas de 18 años o mayores, ha identificado cuatro factores potencialmente asociados con el desarrollo de la misma: la edad, el peso, el número de cigarrillos fumados por día y los antecedentes familiares de enfermedad cardíaca. El investigador quiere acumular pruebas que confirmen estos factores como contribuyentes al desarrollo de la enfermedad, o demuestren que no son importantes. ¿Cómo debe proceder?

Aquí se plantea un problema estadístico. ¿Qué características lo identifican como tal? Simplemente éstas:

1. El problema se asocia a un grupo grande de objetos (en este caso, personas) acerca de los cuales van a hacerse inferencias. Este grupo de objetos se llama *población.*
2. Ciertas características de los miembros de la población son de particular interés. El valor de cada una de esas características puede cambiar de objeto a objeto dentro de la población. Estas características se llaman *variables aleatorias:* variables porque cambian de valor; aleatorias porque su comportamiento depende del azar y es impredecible.
3. La población es demasiado grande para ser estudiada en su totalidad. Por tanto, debemos hacer inferencias sobre la población basadas en lo observado estudiando sólo una porción, o *muestra,* de objetos de la población.

En el estudio de factores que afectan a la enfermedad cardíaca, la población es el conjunto de todas las personas que padecen la enfermedad. Las variables aleatorias de interés son la edad y el peso del paciente, el número de cigarrillos fumados por día y la historia familiar. Es imposible identificar y estudiar a cada persona con enfermedad cardíaca. De este modo,

cualesquiera que sean las conclusiones, deben basarse solamente en el estudio de una porción o muestra de esas personas.

Las variables aleatorias se agrupan en dos categorías: continuas y discretas. Una *variable aleatoria continua* es una variable que puede tomar cualquier valor en algún intervalo o porción continua de los números reales. En el estudio de la enfermedad cardíaca, la variab le edad es continua, como lo es también la variable peso. Por ejemplo, la edad de una persona puede tomar cualquier valor entre 18 y, digamos, 110 años, intervalo continuo de tiempo. Y el peso de una persona puede situarse en cualquier lugar, digamos entre 40 y, quizá, 270 kg. Una *variable aleatoria discreta* es una variable que toma su valor en puntos aislados. De este modo, el conjunto de los posibles valores es finito o infinito numerable. Con frecuencia, las variables aleatorias discretas surgen en la práctica en conexión con las variables de conteo. El número de cigarrillos fumados por día es discreto. Si contamos la parte de un cigarro fumado como un cigarro entero, entonces su conjunto de posibles valores es {0, 1, 2, 3, 4, 5,...}, una colección infinita numerable. Si el historial familiar se estudia registrando el número de padres y abuelos que experimentaron dolencias cardíacas, entonces esta variable es también discreta. El conjunto de sus posibles valores es {0, 1, 2, 3, 4, 5, 6}, una colección finita. Generalmente, las variables aleatorias se indican con letras mayúsculas.

Una medida descriptiva relacionada con una variable aleatoria, cuando la variable se considera sobre toda la población, se denomina *parámetro*. Los parámetros se indican generalmente con letras griegas. Para recordar que los parámetros describen poblaciones sólo hay que observar que ambos empiezan por *p*. Un parámetro con el que es frecuente encontrarse es el valor promedio de la población o media de la población. Este parámetro se indica mediante la letra griega μ. Por ejemplo, en el estudio de las enfermedades cardíacas, el investigador estaría interesado en determinar el valor promedio de cigarros fumados al día por los miembros de la población. No es posible obtener el valor exacto de este parámetro, salvo que sean estudiados todos los miembros de la población. Puesto que es imposible hacerlo, el valor exacto de μ seguirá siendo desconocido incluso tras haber finalizado nuestro estudio. Sin embargo, podremos utilizar métodos estadísticos para aproximarnos a su valor basándonos en los datos obtenidos a partir de la muestra de pacientes extraída de la población.

Una medida descriptiva relacionada con una variable aleatoria, cuando la variable sólo se considera sobre una muestra, se denomina *estadístico*. Los estadísticos tienen dos fines. Por un lado, describen la muestra que está disponible y, por otro, sirven como aproximación a los parámetros correspondientes a la población. Por ejemplo, la media de cigarros fumados diariamente por los miembros de una muestra de pacientes con enfermedades cardíacas es un *estadístico*. Se le denomina promedio de la muestra o media muestral. Su valor para una muestra dada, probablemente, no será exactamente igual a la media μ de la población. Sin embargo, se espera que al menos su valor se aproxime a μ.

Un estadístico, o usuario estadístico, siempre está trabajando en dos mundos. El mundo ideal está al nivel de la población y es de naturaleza teórica. Es el mundo que desearíamos ver. El mundo de la realidad es el mundo de la muestra. Este es el nivel en el que realmer te operamos. Esperamos que las características de nuestra muestra reflejen bien las características de la población. Es decir, tratamos nuestra muestra como un microcosmos que refleja a toda la población. La idea se ilustra en la Figura 1.1.

Nos interesamos principalmente por tres cuestiones concernientes al comportamiento de la variable aleatoria. Son éstas:

1. ¿Cuál es la posición de la variable? Es decir, ¿alrededor de qué valor fluctúa la variable?
2. ¿Qué cantidad de variación existe? Es decir, los valores de la variable observados, ¿tienden a agruparse o se encuentran muy dispersos?

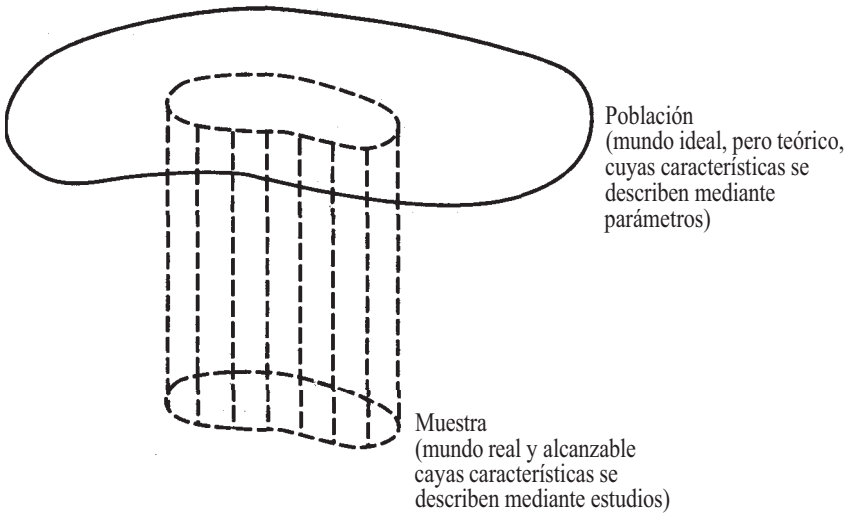


Figura 1.1. La muestra se visualiza como una población en miniatura. Se espera que el comportamiento de la variable aleatoria estudiada en la muestra proporcione una imagen precisa de su comportamiento en la población.

3. ¿Cuál es la forma de la distribución? Es decir, ¿tienden los valores a adoptar forma de campana, plana, en U u otra diferente?

En este capítulo introducimos algunas técnicas gráficas y analíticas que pueden utilizarse para responder a estas cuestiones.

1.1. TABLAS DE DISTRIBUCIÓN: DATOS DISCRETOS

Recuérdese que una variable aleatoria discreta es una variable aleatoria que puede tomar como máximo un número finito o un número infinito numerable de valores posibles. Las variables aleatorias discretas surgen frecuentemente en los datos de cualquier estudio. Por ejemplo, quizá queramos comparar la opinión de las mujeres con la opinión de los hombres sobre el tema del aborto. De ahí que una variable del estudio sea el «sexo». Esta variable es discreta pues sólo toma los dos valores que aparecen de forma natural, «masculino» o «femenino». Podríamos preguntarnos: ¿Está usted a favor de la legalización del aborto si se solicita durante el primer trimestre del embarazo? Dado que la respuesta a esta pregunta varía de una persona a otra, nos encontramos ante una variable aleatoria. El investigador puede decidir registrar cada respuesta como «sí», «no», «indeciso» o «no sabe, no contesta». De esta forma, se crea una variable aleatoria discreta con cuatro valores posibles. Para comprender y resumir estos datos, es útil presentarlos en una tabla o gráfico, en la que aparezcan los valores posibles de la variable aleatoria y el número de veces que cada valor se repite. A este número se le denomina *frecuencia absoluta* o simplemente *frecuencia.* El Ejemplo 1.1.1 recoge esta idea.

Ejemplo 1.1.1. Se realiza un estudio comparativo de dos geriátricos del oeste de Virginia. El objetivo del estudio es determinar el tipo de pacientes a los que se atiende y averiguar dónde van los pacientes cuando dejan el geriátrico. Están implicadas cuatro variables aleatorias discretas: sexo (codificado por el investigador como F = femenino o M = masculino), diagnóstico (codificado como RM = retrasado mental, EM = enfermo mental, FE = físicamen-

te enfermo), edad y destino al dejar el geriátrico (codificado como 1 = fallecido, 2 = hogar de familiares, 3 = hospital, 4 = calle, 5 = otro geriátrico, 6 = sanatorio particular, 7 = no ha dejado el geriátrico). (Los datos presentados son de un geriátrico real y han sido recogidos de un estudio más amplio realizado por el laboratorio estadístico y Debbie Thompson, Departamento de Trabajos Sociales, Radford University, 1990.)

Sexo	Diagnóstico	Edad	Destino	Sexo	Diagnóstico	Edad	Destino
M	EM	29	2	F	EM	72	6
M	RM	35	7	M	EM	52	7
F	FE	34	7	F	FE	31	7
M	EM	36	7	M	FE	35	7
F	RM	25	7	M	FE	42	7
F	EM	20	7	F	EM	29	2
F	FE	31	7	F	RM	61	7
F	FE	89	1	F	EM	18	3
M	RM	42	7	F	RM	64	7
M	EM	41	7	M	FE	51	7
F	FE	47	7	F	FE	30	7
M	FE	41	2	F	RM	35	7
M	EM	87	7	M	FE	40	6
F	RM	56	1	M	RM	76	3
F	RM	50	7	M	FE	59	7
F	FE	28	7	F	EM	71	6
M	RM	35	7	F	EM	62	7
F	FE	23	7	F	EM	65	3
F	RM	39	3	M	RM	51	7
M	FE	42	7	F	RM	18	7

La distribución de frecuencias para la variable *diagnóstico* se muestra en la Tabla 1.1. Obsérvese que la tabla relaciona la categoría en la que se ubica la respuesta junto con la cantidad de observaciones por categoría.

En la mayoría de estudios se obtienen recuentos de frecuencias, los cuales proporcionan una valiosa idea del comportamiento de la variable aleatoria objeto del estudio. Sin embargo, los recuentos de frecuencia por sí solos pueden causar confusión. Por ejemplo, supongamos que nos dicen que se han diagnosticado 10 nuevos casos de síndrome de inmunodeficiencia adquirida (SIDA) en un hospital particular durante el mes de junio. ¿Es ello motivo de alarma? Quizá sí o quizá no. Naturalmente, depende del número de personas que hayan pasado a prueba de la enfermedad. Diez casos descubiertos entre 20 personas analizadas describen un panorama completamente diferente a 10 casos hallados entre 1000 personas analizadas. Para dar una perspectiva de un recuento de frecuencias, consideramos el recuento relativo al total,

Tabla 1.1. Distribución de frecuencias de la variable *diagnóstico* del Ejemplo 1.1.1

Categoría	Frecuencia
EM (enfermo mental)	12
RM (retrasado mental)	13
FE (físicamente enfermo)	15

formando así *una frecuencia relativa.* La Tabla 1.2 proporciona las distribuciones de frecuencias y de frecuencias relativas de la variable *diagnóstico* del Ejemplo 1.1.1. Las frecuencias relativas pueden multiplicarse por 100 para obtener el porcentaje de observaciones que corresponden a cada categoría. Esta información es útil puesto que los porcentajes son rápidamente comprendidos por todos. La Tabla 1.3 muestra el resumen completo de la variable *diagnóstico.*

La Tabla 1.4 es el resumen completo de los datos tal y como lo presentaría el SAS, iniciales de *Statistical Analysis System,* paquete informático de amplio uso entre analistas de datos, estadísticos e investigadores. Algunas nociones básicas del SAS se explican en la sección de Herramientas Computacionales de este libro. Obsérvese que el SAS ha listado las variables de diagnóstico por orden alfabético. También ha incluido una columna llamada «frecuencia acumulada» y otra llamada «porcentaje acumulado». La palabra *acumulado/a* significa que los valores se suman acumulándose. Así, la frecuencia acumulada 25 se obtiene al sumar el número de pacientes con retraso mental (13), que se encuentra en la segunda fila, al número de pacientes con enfermedades mentales (12), que se encuentra en la primera fila; la frecuencia acumulada 40 es la suma de todos los valores en la columna de frecuencias (40 = 12 + 13 + 15). Obsérvese que si los datos han sido introducidos correctamente, el último número de la columna de frecuencias acumuladas debe ser el tamaño de la muestra.

La columna de porcentaje acumulado se obtiene sumando la columna de porcentaje; su último valor debe ser siempre 100 %. No obstante, en algunas tablas los porcentajes pueden no sumar 100 % exactamente, debido a diferencias en el redondeo. Debemos señalar que, cuando los valores de las variables no son numéricos o tienen un orden lineal no natural, la distribución acumulada puede no ser significativa. El código del SAS usado para hacer esta tabla se proporciona en la sección de Herramientas Computacionales al final de este capítulo.

Tabla 1.2. Distribución de frecuencias y de frecuencias relativas de la variable *diagnóstico* del Ejemplo 1.1.1

Categoría	Frecuencia	Frecuencia relativa
EM (enfermo mental)	12	12/40 = 0.300

Tabla 1.3. Distribución completa de la variable *diagnóstico* del Ejemplo 1.1.1

Categoría	Frecuencia	Frecuencia relativa	Porcentaje
EM (enfermo mental)	12	12/40 = 0.300	30.0
RM (retrasado mental)	13	13/40 = 0.325	32.5
FE (físicamente enfermo)	15	15/40 = 0.375	37.5

Tabla 1.4. Frecuencias y porcentajes para la variable diagnóstico del Ejemplo 1.1.1

Diagnóstico	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
EM	12	30.0	12	30.0
RM	13	32.5	25	62.5
FE	15	37.5	40	100.0

Gráficos de barras

Para transmitir visualmente la información contenida en una tabla de distribución, se puede utilizar un gráfico de barras verticales. Cada categoría está representada por una barra verti cal, todas de la misma anchura. Las alturas de las barras dependen del número de observacio nes por categoría. El eje vertical del gráfico puede representar frecuencias, frecuencias relativas o porcentajes. Cada tipo de gráfico es informativo, y los dos últimos tienen la ventaja de que sus escalas verticales no dependen de los datos. En el caso de un gráfico de barras de frecuencias relativas, varían de 0 a 1 y, en el caso de un gráfico de porcentajes, de 0 % 100%. La Figura 1.2 muestra todos estos gráficos para la variable *diagnóstico* del Ejemplo 1.1.1. Si se desea, las barras pueden colocarse horizontalmente. De hecho, los gráficos de barras horizontales son algunas veces preferibles al escribir informes, puesto que requieran

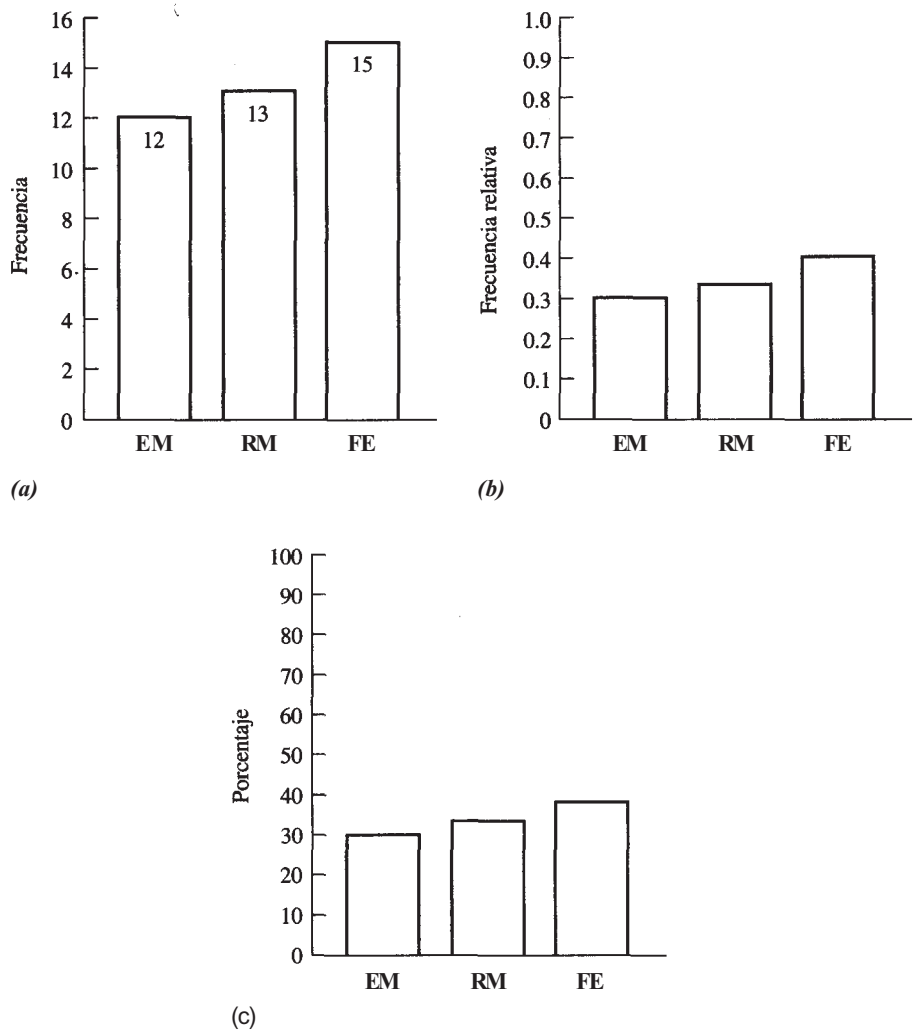


Figura 1.2. (a) Gráfico de barras de frecuencias para la variable *diagnóstico* del Ejemplo 1.1. *(b)* gráfico de barras de frecuencias relativas para la variable *diagnóstico; (c)* gráfico de barras de porcentajes para la variable *diagnóstico.*

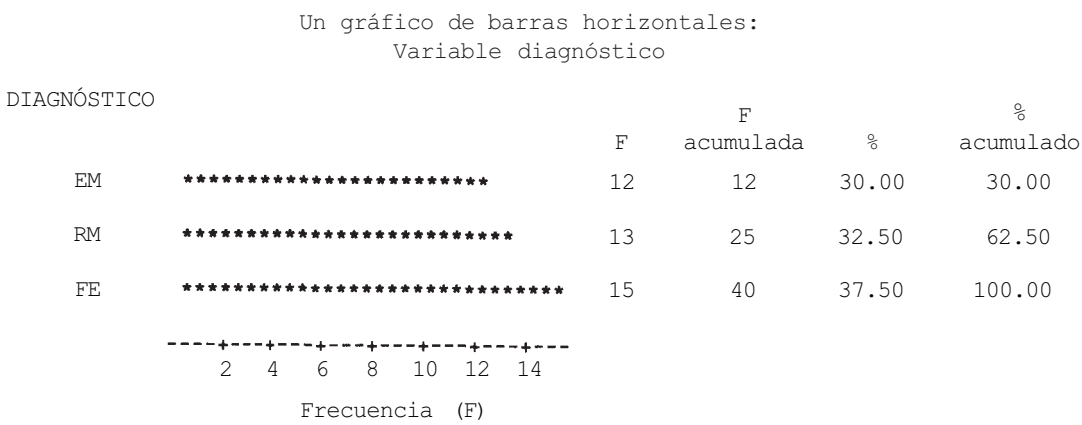


Figura 1.3. Un gráfico de barras horizontales, obtenido con el SAS, para la variable diagnóstico del Ejemplo 1.1.1.

menos espacio que los gráficos de barras verticales. La Figura 1.3 muestra el gráfico de barras horizontal para la variable *diagnóstico* hecho por el SAS. Nótese que este gráfico también muestra la información contenida en la Tabla 1.4.

Datos con dos variables: tablas de doble entrada

Algunas veces deseamos estudiar simultáneamente dos variables aleatorias discretas. Por ejemplo, es posible que queramos utilizar datos del Ejemplo 1.1.1 para investigar una posible relación entre el sexo de un paciente y el diagnóstico efectuado. Para iniciar este estudio, construimos una tabla de doble entrada. Esta tabla contiene r filas, donde r es el número de respuestas posibles de la primera variable, y c columnas, donde c es el número de respuestas asociadas a la segunda variable. De esta forma, una tabla de doble entrada tiene $r \cdot c$ categorías o celdas. Generalmente, en la tabla se incluye la información concerniente a las frecuencias, frecuencias relativas y los porcentajes. En los márgenes de la tabla también se incluye la distribución marginal de cada variable.

Ejemplo 1.1.2. Consideremos los datos del Ejemplo 1.1.1. La variable aleatoria *sexo* tiene dos respuestas posibles. Si utilizamos esta variable para formar las filas de nuestra tabla, $r = 2$. La variable aleatoria *diagnóstico* tiene tres respuestas y, por lo tanto, la tabla tiene $c = 3$ columnas. Esto da como resultado una tabla de doble entrada con $r \cdot c = 2 \cdot 3 = 6$ celdas, las cuales se muestran en la Tabla 1.5. En la Tabla 1.6 se muestra la distribución de los 40 pacientes en las 6 celdas.

Tabla 1.5. Tabla de distribución de doble entrada para el estudio del Ejemplo 1.1.1 con $r = 2$, $c = 3$ y $r \cdot c = 6$ celdas

	Diagnóstico		
Sexo	**EM**	**RM**	FE
F	Mujeres enfermas mentales	Mujeres retrasadas mentales	Mujeres físicamente enfermas
M	Varones enfermos mentales	Varones retrasados mentales	Varones físicamente enfermos

Tabla 1.6. Tabla de distribución de doble entrada: *sexo* y *diagnóstico,* para los datos del Ejemplo 1.1.1 *

	Diagnóstico			
Sexo	EM	RM	FE	**Distribución del *sexo***
F	7 0.175 17.5 %	8 0.200 20.0%	8 0.200 20.0%	23 0.575 57.5%
M	5 0.125 12.5 %	5 0.125 12.5 %	7 0.175 17.5%	17 0.425 42.5 %
Distribución del *diagnóstico*	12 0.300 30%	13 0.325 32.5 %	15 0.375 37.5 %	40

* En cada celda, el primer número nos da la frecuencia, le sigue la frecuencia relativa y, finalmente, el porcentaje.

Las tablas de doble entrada pueden construirse de manera que permitan una rápida comparación de un grupo de individuos de una muestra con otra. Por ejemplo, en el estudio del Ejemplo 1.1.1 sería interesante comparar el diagnóstico de los varones con el de las mujeres. Ya que las frecuencias, las frecuencias relativas y los porcentajes de la Tabla 1.6 se refieren a las celdas, se requieren algunos cálculos adicionales para hacer la comparación. Obsérvese que en la muestra hay 17 varones. De éstos, 5 son retrasados mentales, 5 son enfermos mentales y 7 están físicamente enfermos. Esta información puede utilizarse para obtener la distribución de los varones en la muestra, la cual se presenta en la Tabla *1.7 a.* En la Tabla *1.7 b* se incluye la distribución de las 23 mujeres de la muestra. Obsérvese que existen algunas diferencias entre las distribuciones de los dos grupos. La más sobresaliente es que, en el momento actual, el porcentaje de varones físicamente enfermos (41.18 %) es algo superior al porcentaje de mujeres físicamente enfermas (34.78 %).

El SAS incluye un procedimiento para obtener tablas de doble entrada de manera rápida. La Tabla 1.8 es la versión del SAS para la tabla de doble entrada de los datos del Ejemplo 1.1.1. Obsérvese que el SAS lista automáticamente los encabezamientos de las filas y las columnas en orden alfabético. La esquina superior izquierda de la salida impresa del SAS comenta el significado de los números contenidos en cada celda. El primer número es la frecuencia. Por ejemplo, en la tabla podemos ver que hay 7 mujeres con enfermedades mentales en la muestra. El segundo número de cada celda proporciona el porcentaje que dicha frecuencia repru-

Tabla 1.7a. Distribución de la variable *diagnóstico* para los hombres del Ejemplo 1.1.1

	Diagnostico			
Sexo	**EM**	**RM**	**FE**	
F				
M	5 $\frac{5}{17} = 0.2941$ 29.41 %	5 $\frac{5}{17} = 0.2941$ 29.41 %	7 $\frac{7}{17} = 0.4118$ 41.18%	17

Tabla 1.76. Distribución conjunta de la variable *diagnóstico* para los hombres y mujeres

	Diagnóstico			
Sexo	**EM**	**RM**	**FE**	
F	**7** $\frac{7}{23} = 0.3043$ **30.43%**	**8** $\frac{8}{23} = 0.3478$ **34.78%**	**8** $\frac{8}{23} = 0.3478$ **34.78%**	**23**
M	**5** $\frac{5}{17} = 0.2941$ **29.41 %**	**5** $\frac{5}{17} = 0.2941$ **29.41%**	**7** $\frac{7}{17} = 0.4118$ **41.18%**	**17**

senta respecto al total. Podemos ver que 7 mujeres con enfermedades mentales constituyen un 17.5 % del total de la muestra. El tercer número da el porcentaje que la frecuencia representa en la fila. En el ejemplo podemos ver que el 30.43 % de las *mujeres* de la muestra estaban mentalmente enfermas. Obsérvese que los porcentajes por fila coinciden con los mostrados en la Tabla 1 .*7b.* La Tabla 1.8 nos permite ver que, de los pacientes con enfermedades mentales, el 58.33 % eran mujeres. Advierta que la distribución de la variable *sexo* se encuentra en el SAS en los totales y los porcentajes que aparecen en la última columna y que coincide con los valores de la Tabla 1.6. Además, la distribución para la variable *diagnóstico* se muestra en los totales y porcentajes de la última fila, y coincide también con los datos de la Tabla 1.6. El código para este procedimiento en el SAS se encuentra en la sección de Herramientas Computacionales al final del capítulo.

Tabla 1.8. Tabla de doble entrada usada para investigar la asociación entre *sexo* y *diagnóstico* en el Ejemplo 1.1.1, tal y como la muestra el SAS.

Tabla de doble entrada usada para investigar la relación entre sexo y diagnóstico

Tabla de Sexo frente a DIAGNÓSTICO

Sexo	DIAGNÓSTICO			
Frecuencia Porcentaje Pct fila Pct col	EM	RM	FE	Total
F	7 17.50 30.43 58.33	8 20.00 34.78 61.54	8 20.00 34.78 53.33	23 57.50
M	5 12.50 29.41 41.67	5 12.50 29.41 38.46	7 17.50 41.18 46.67	17 42.50
Total	12 30.00	13 32.50	15 37.50	40 100.00

Recuerde que lo único que hemos hecho en los ejemplos de esta sección es describir una muestra de pacientes de un geriátrico concreto. Las técnicas para llegar a conclusiones sobre la población de pacientes atendidos por este geriátrico, a partir de la muestra, se presentarán en los capítulos siguientes.

EJERCICIOS 1.1

1. Los datos siguientes proceden de un segundo geriátrico y están tomados del estudio descrito en el Ejemplo 1.1.1.

Sexo	Diagnóstico	Edad	Destino	Sexo	Diagnóstico	Edad	Destino
F	EM	67	6	M	RM	80	7
M	FE	71	7	M	EM	83	2
F	FE	54	1	M	FE	49	3
F	EM	63	7	F	FE	78	6
F	EM	48	7	M	EM	57	7
M	EM	56	7	M	RM	69	3
M	FE	62	3	F	EM	83	7
F	RM	57	2	F	FE	92	1
F	FE	81	7	F	FE	55	3
F	FE	36	7	F	FE	63	6
F	FE	72	3	F	FE	64	4
F	FE	65	3	M	FE	89	7

   a) Construir una tabla de distribución para la variable *diagnóstic.,*
   b) Construir un gráfico de barras de porcentajes para la variable *diagnóstico.* Comentar las diferencias que se observan entre este gráfico y el de la Figura 1. 2c.
2. *a)* Utilizar los datos del Ejemplo 1.1.1 y del Ejercicio 1 para construir gráficos de barras de frecuencias para la variable *destino* en los dos geriátricos. ¿Por qué no es adecuada una comparación de esta variable, basadas sólo en estos gráficos, para las dos muestras?
   b) Construir un gráfico de barras de frecuencias relativas para la variable *destino* de cada geriátrico. Comentar cualquier diferencia aparente que se observe.
   c) Construir una tabla de doble entrada de *sexo* frente a *destino* para cada grupo.
   d) Construir una tabla de doble entrada de *diagnóstico* frente a *destino* para cada grupo.
3. Construir una tabla de doble entrada de *sexo* frente a *diagnóstico* para los datos del Ejercicio 1. Comparar esta tabla con la Tabla 1.6. Comentar cualquier diferencia aparente que se observe.
4. Construir una tabla de doble entrada de *sexo* frente a *diagnóstico* que muestre la distribución de la variable *diagnóstico* para cada sexo según los datos del Ejercicio 1. Comentar cualquier diferencia notable entre los varones y las mujeres de esta muestra.
5. Se realizó un estudio para investigar la relación entre la dieta y la aparición de dolores de cabeza. Se identificaron dos grupos de personas que sufrían jaqueca crónica. Estos se codificaron como V = jaquecas vasculares y T = jaquecas por tensión. En este estudio también se incluyó un grupo de control consistente en personas que decían sufrir dolores de cabeza poco frecuentes. Estas personas se codificaron como C. Cada individuo también fue identificado por su sexo. El grupo de estudio es el siguiente: (Basado en un estudio recogi-

do en Patricia Guarnieri, Cynthia Radnitz y Edward Blanchard, «Assessment of Dietary Risk Factors in Chronic Headache», *Biofeedback and Self-Regulation,* vol. 15, marzo de 1990, págs. 15-25.)

Sexo	Diagnóstico	Sexo	Diagnóstico	Sexo	Diagnóstico	Sexo	Diagnóstico
M	V	F	V	M	V	M	V
F	V	F	T	M	V	F	T
F	V	M	C	M	V	M	V
M	T	F	V	M	C	F	C
F	T	M	V	F	V	M	T
M	V	M	C	F	T	M	V
M	V	F	C	F	V	M	T
M	C	F	V	F	V	M	C
F	C	F	V	F	V	F	V
F	V	F	C	F	V	F	V
F	C	M	V	F	V	F	V
M	V	F	T	F	V	F	T
F	C	F	C	F	T	F	V
F	C	F	C	F	T	F	T
M	V	F	V	F	C	F	V
F	C	F	C	F	V	F	C
F	V	F	C	F	T	F	T
M	V	M	V	M	C	F	V
F	C	F	T	F	T	M	C
F	C	F	C	M	C	M	T
F	T	M	V	F	V	F	C
M	V	F	T	M	C	F	T
M	T	F	C	M	T	F	V
F	V	M	C	F	V	F	V
M	T	F	T	F	C	F	T
F	V	F	V	F	V	M	T
F	V	F	T	F	C	F	V
F	V	F	V	F	C	F	V
F	T	F	V	F	V	F	C
F	V	F	V	F	C	F	V
F	T	F	V	F	T	F	V
F	V	F	V	F	V	M	T
F	V	F	C	M	C	M	T
F	V	F	V	M	T	M	C
F	T	F	V	M	T	F	C
F	T	F	C	F	V	F	C
F	C	F	V	F	C	M	T
F	V	M	T	F	T	F	V
M	C	F	V	F	V	F	C
M	T	M	T	F	T	F	T
F	T	F	V	F	V	F	C
F	V	F	T	F	T	F	V
F	V	M	C	F	C	F	V
F	V	F	V	F	V		
F	V	F	C				

a) Construir una tabla de doble entrada de *sexo* frente a *diagnóstico,*
b) Construir una tabla de doble entrada que permita comparar el diagnóstico de los varones con el de las mujeres.

6. En muchas disciplinas se hacen preguntas para determinar la fuerza de una opinión mantenida por un grupo de gente con respecto a un tema determinado. Las respuestas se puntúan según la escala «Likert». Una escala típica de este tipo etiquetaría las respuestas como sigue:

1 = bastante en desacuerdo
2 = algo en desacuerdo
3 = neutral
4 = algo de acuerdo
5 = bastante de acuerdo
6 = no aplicable

En un estudio sobre la opinión de estudiantes acerca de la afirmación de que «El centro de salud de R.U. tiene horarios convenientes para los estudiantes», se extrajo una muestra de 246 estudiantes y cada uno de ellos se clasificó en función de su género y de su respuesta a la afirmación según la escala Likert. Utilizar el resultado impreso del SAS dado en la Tabla 1.9, para responder a las siguientes preguntas:

a) Sólo hay 245 respuestas de estudiantes en el recuento que aparece en la tabla. Hay varias razones para que haya una respuesta perdida. Enumerar alguna de ellas.
b) ¿Cuántos estudiantes hubo en la muestra que fueran mujeres y que estuvieran bastante de acuerdo?
c) ¿Qué porcentaje de la muestra eran mujeres y estaban bastante de acuerdo con la afirmación?
d) ¿Qué porcentaje de las mujeres estuvo bastante de acuerdo con la afirmación?
e) ¿Qué porcentaje de los varones estuvo bastante de acuerdo con la afirmación?
f) De los que estuvieron bastante de acuerdo, ¿qué porcentaje eran mujeres?

Tabla 1.9. Tabla completa de doble entrada para las variables género y horario del centro de salud

Tabla de género frente a respuesta

Género Respuesta

Frecuencia Porcentaje Pct fila Pct col	Bastante en desacuerdo	Algo en desacuerdo	Neutral	Algo de acuerdo	Bastante de acuerdo	No aplicable	Total
f	18	20	23	45	20	16	142
	7.35	8.16	9.39	18.37	8.16	6.53	57.96
	12.68	14.08	16.20	31.69	14.08	11.27	
	78.26	50.00	47.92	61.64	68.97	50.00	
m	5	20	25	28	9	16	103
	2.04	8.16	10.20	11.43	3.67	6.53	42.04
	4.85	19.42	24.27	27.18	8.74	15.53	
	21.74	50.00	52.08	38.36	31.03	50.00	
Total	23	40	48	73	29	32	245
	9.39	16.33	19.59	29.80	11.84	13.06	100.00

Frecuencias perdidas = 1

g) Desde un punto de vista intuitivo, ¿parece probable que varones y mujeres difieran sustancialmente en su opinión acerca de los horarios del centro de salud? Razone su respuesta.

h) Dé algunas razones de por qué un estudiante pueda marcar como «no aplicable» su respuesta a la afirmación.

(Basado en datos obtenidos en la Universidad de Radford, en marzo de 1997.)

7. En un estudio llevado a cabo para investigar la asociación entre las especies de arañas y varias actividades que éstas realizan, se obtuvieron los datos de la Tabla 1.10. Utilizar éstos para contestar a las siguientes cuestiones:

a) ¿Qué especies tienen el mayor porcentaje de sus miembros participando en la actividad del cortejo?

b) ¿Qué porcentaje de toda la muestra de 130 arañas participaron en el cortejo?

c) Basándose en estos datos, ¿cree que hay una diferencia sustancial en los hábitos de cortejo entre estas dos especies? Razónese la respuesta.

(Basado en un experimento llevado a cabo por Travis Alderman, Departamento de Biología, Universidad de Radford, en la primavera de 1997.)

1.2. UN VISTAZO RÁPIDO A LA DISTRIBUCIÓN: DIAGRAMA DE TALLO Y HOJAS

Antes de comenzar a analizar un conjunto de datos es importante comprender lo que representan los datos. En particular, es importante entender que cada número de un conjunto de datos es un valor observado de alguna variable aleatoria. Algunas veces tenemos datos de toda la población; habitualmente no es así. Cuando los datos disponibles son datos de población, cualquier pregunta pertinente puede responderse mediante observación directa. No existe incertidumbre en lo concerniente a las características de la población. Sin embargo,

Tabla 1.10. Tabla de doble entrada para las variables *especie* y *presencia de cortejo*

Tabla de especie frente a respuesta

Especie	Respuesta		
Frecuencia Porcentaje Pct fila Pct col	no	sí	Total
1	40	25	65
	30.77	19.23	50.00
	61.54	38.46	
	53.33	45.45	
2	35	30	65
	26.92	23.08	50.00
	53.85	46.15	
	46.67	54.55	
Total	75	55	130
	57.69	42.31	100.00

si los datos sólo representan una muestra de las observaciones extraídas de la población, entonces necesitaremos emplear métodos estadísticos para determinar la naturaleza de la población.

Consideremos una variable aleatoria cuantitativa discreta con un gran número de valores posibles o una variable aleatoria continua. Nuestra primera tarea será tener alguna idea de *a* distribución de la variable aleatoria. Es decir, deseamos determinar dónde se centran los valores, si se distribuyen de manera amplia o si encajan en un patrón característico. Para ello, emplearemos algunas de las herramientas del análisis exploratorio de datos (EDA). En palabras de John W. Tukey, un conocido analista de datos y autor de muchas de las técnicas EDA[16].

> Tendremos que trabajar con números. Es necesario que los manejemos con facilidad y los observemos de forma eficiente. Las técnicas para la manipulación y visualización —ya sea gráfica, aritmética o mixta— serán importantes. Cuanto más simples sean estas técnicas, siempre que funcionen, mejor trabajaremos.

Una técnica para la observación de la distribución, que funciona bien, es el diagrama de tallo y hojas. Es fácil de diseñar y puede hacerse rápidamente. Como se verá, en el diagrama de tallo y hojas, el conjunto de datos estará reproducido bastante fielmente. Así, creamos un diagrama en el que los datos puntuales se agrupan de tal modo que podemos visualizar la forma de la distribución mientras que mantenemos su individualidad. Un diagrama de tallo y hojas consiste en una serie de filas horizontales de números. El número utiliza do para designar una fila es su *tallo,* el resto de números de la fila se denominan *hojas.* El tallo es la mayor porción del número. Por ejemplo, en los números 3.1, 3.2, 3.7 y 3.5 lo primero que salta a la vista es que todo son «treses»: el tallo de cada número es tres. Las hojas dan una información secundaria acerca del número, en nuestro ejemplo sería la cifra decimal, que serviría para distinguir entre los «treses». No hay reglas exhaustivas sobre cómo diseñar este diagrama. En general, los pasos son los siguientes:

Construcción de un diagrama de tallo y hojas simple

1. Elija algunos números oportunos que puedan servir de tallos. Para facilitar la determinación de la forma se necesitan al menos 5 tallos. Los tallos elegidos generalmente son el primero o los dos primeros dígitos de los números del conjunto de datos.
2. Etiquete las filas con los tallos elegidos.
3. Reproduzca gráficamente los datos anotando el dígito que sigue al tallo, como hoja del tallo adecuado.
4. Gire el gráfico hacia un lado para ver cómo se distribuyen los números. En concreto, intente responder a preguntas como:

    a) ¿Tienden a agruparse los datos cerca de un tallo o tallos en particular, o se distribuyen de forma uniforme por el diagrama?
    b) ¿Tienden a estrecharse los datos hacia un extremo u otro del diagrama?
    c) Si se traza una curva a lo largo de la parte superior del diagrama, ¿forma más o menos una campana? ¿Es plana? ¿Es simétrica?

Un ejemplo aclarará la idea.

Ejemplo 1.2.1. Considere estas observaciones sobre la variable aleatoria *X,* magnitud de un terremoto en California según su medición en la escala de Richter:

1.0	8.3	3.1	1.1	5.1
1.2	1.0	4.1	1.1	4.0
2.0	1.9	6.3	1.4	1.3
3.3	2.2	2.3	2.1	2.1
1.4	2.7	2.4	3.0	.4.1
5.0	2.2	1.2	7.7	1.5

Los primeros dígitos de estos números son 1, 2, 3, 4, 5, 6, 7, 8. Estos dígitos servirán como nombres de los tallos y las filas. Véase la Figura *Aa.* A continuación, representamos los datos gráficamente anotando el número que aparece después del punto decimal, como hoja del tallo apropiado. En la Figura *Ab* se muestran los primeros cuatro datos puntuales. En la Figura 1.4c se visualiza todo el conjunto de datos. Para tener una idea de la forma, gire el libro hacia un lado y observe la curva que se ha trazado en la Figura 1.4d. De aquí puede deducirse que estos datos se aproximan al extremo inferior de la escala: muchos terremotos eran suaves. Si este ejemplo fuera una indicación precisa de la intensidad de los terremotos en California, sería bastante inusual observar un terremoto intenso. Obsérvese también que la visualización no es simétrica. Hay más bien una cola larga o ahusada hacia el extremo superior o derecho de la visualización. Se dice que los datos de este tipo están sesgados hacia la derecha. Si la cola larga estuviera hacia la izquierda, diríamos que los datos están sesgados hacia la izquierda. (Basado en los datos hallados en Robert Iacopi, *Earthquake Country,* Lane Books, Menlo Park, Calif., 1971.)

Algunas veces, la utilización del primero o de los dos primeros dígitos de los datos puntuales como tallos, no proporciona suficientes tallos como para permitirnos detectar la forma. Una manera de solucionar este problema es utilizar tallos dobles. Es decir, utilizar cada tallo dos veces: una vez para representar las hojas inferiores 0, 1, 2, 3, 4 y, a continuación, nuevamente para representar las hojas superiores 5, 6, 7, 8, 9. El Ejemplo 1.2.2 ilustra el diagrama de tallo doble.

Ejemplo 1.2.2. En un estudio sobre el crecimiento de los varones, se obtuvieron estas observaciones sobre *X,* perímetro craneal en centímetros, de un niño al nacer.

33.1	34.6	34.2	36.1	34.2	35.6
34.5	35.8	34.5	34.2	34.3	35.2
33.7	36.0	34.2	34.7	34.6	34.3
33.4	34.9	33.8	33.6	35.2	34.6
33.7	34.8	33.9	34.7	35.1	34.2
36.5	34.1	34.0	35.1	35.3	

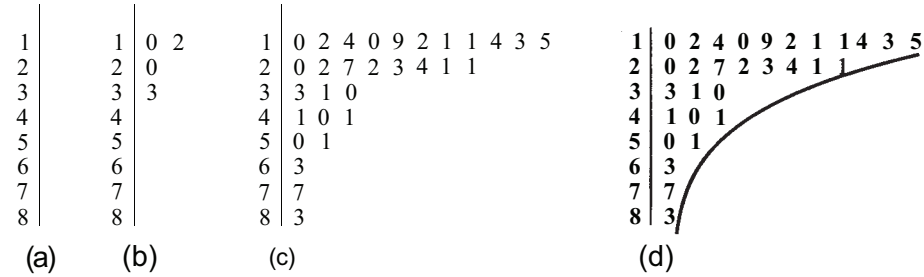


Figura 1.4. Visualización del diagrama de tallo y hojas para la variable aleatoria: magnitud de una muestra de terremotos en California, medidos en la escala de Richter: (a) elección de tallos, *(b)* registro de los primeros cuatro datos puntuales, *(c)* visualización de todo el conjunto de datos, (d) búsqueda de la forma.

```
33 | 1 4
33 | 7 7 8 9 6
34 | 1 2 2 0 2 2 3 3 2
34 | 5 6 9 8 5 7 7 6 6
35 | 1 2 1 3 2
35 | 8 6
36 | 0 1
36 | 5
```

Figura 1.5. Visualización de un diagrama de tallo y hojas doble dando el perímetro craneal, en centímetros, de un niño al nacer, basándose en los datos del Ejemplo 1.2.2.

Si utilizamos los primeros dos dígitos como tallos, sólo tendremos cuatro tallos, 33, 34, 35, 36. Puesto que no es suficiente para que podamos detectar la forma, utilizaremos dos veces cada uno de ellos y formaremos un diagrama de tallo doble. La visualización se muestra en la Figura 1.5. Obsérvese que, en cada caso, las hojas inferiores 0, 1, 2, 3,4 están representadas en el primer tallo seguidas por las hojas superiores 5,6,7,8,9. De aquí podemos observar que los datos tienden a agruparse alrededor de 34 centímetros. Aunque no hay una simetría perfecta, estos datos son más simétricos que los datos del terremoto del Ejemplo 1.2.1.

Los diagramas de tallo y hojas son útiles para comparar dos grupos de datos de naturales a similar. Por ejemplo, podríamos querer comparar los niveles de colesterol de varones y mujeres; o representar los resultados de dos programas de pérdida de peso, uno frente a otro; o bien querríamos una representación visual del crecimiento a lo largo del tiempo de una especie de árbol a dos alturas diferentes. Comparaciones de este tipo pueden realizarse por medio de los llamados diagramas de tallo y hojas «adosados». El Ejemplo 1.2.3 ilustra esta técnica.

Ejemplo 1.2.3. En un estudio llevado a cabo para comparar el crecimiento durante 10 años del roble americano a una altitud de 975 m y a otra de 675 m, la variable medida fue la longitud de muestras del núcleo, cubriendo los últimos 10 años de anillos de crecimiento, en centímetros. En la Figura 1.6 se muestran los datos obtenidos:

975 m			675 m		
3.8	2.8	6.0	1.8	2.3	1.0
1.3	3.8	1.7	2.3	1.1	2.9
2.6	1.5	1.9	2.0	1.1	0.8
2.2	4.0	2.5	2.2	2.6	1.6
2.0	1.7	0.7	2.4	2.1	1.7

Podríamos construir un diagrama de tallos y hojas para cada conjunto de datos, por separado. No obstante, puesto que el propósito es la comparación y hay solapamiento de tallos, ambos

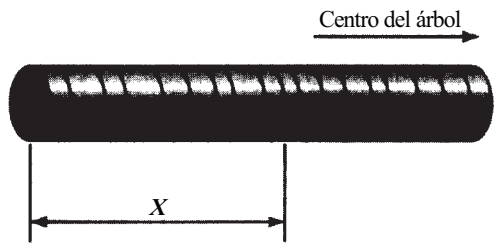


Figura 1.6. Una muestra de núcleo típica. Cada anillo representa un año de crecimiento. La variable X es la longitud, en centímetros, de los últimos 10 anillos.

conjuntos de datos pueden representarse convenientemente en el mismo conjunto de tallos. Para hacer esto, dejamos la parte entera de cada dato como tallo; el segundo dígito de cada dato sería la hoja. Los tallos se sitúan en el centro del diagrama. Las hojas para la altitud de 975 m se muestran en la izquierda y las de 675 m, a la derecha. La Figura 1.7 muestra el diagrama de tallos y hojas adosado. Hay varias cuestiones que comentar. Primero, aunque no tenemos observaciones con tallo 5, éste debe ser incluido en el diagrama. Esto es así para que el ojo pueda tener la perspectiva correcta acerca de la dispersión o distribución de los datos a 675 m. Segundo, parece que hay diferencias en la distribución de la variable crecimiento entre esas dos alturas. Los valores a 975 m están más dispersos que los de 675 m, como se deduce del hecho de que haya 4 valores mayores que cualquiera de los encontrados a 675 m. Parece también que el centro de los datos a 675 m está por debajo del de 975 m, puesto que el conjunto inferior no tiene datos que excedan de 2.9 mientras que el superior tiene varios. Téngase en cuenta que estas afirmaciones sólo son aproximaciones basadas en la figura. En capítulos posteriores se aprenderá cómo comparar posiciones y variabilidad de manera analítica. (Basado en datos obtenidos por Allison Field, Departamento de Biología, Universidad de Radford, en otoño de 1996.)

Muchos de los procedimientos estadísticos que se presentarán más adelante se desarrollan basándose en la suposición de que la variable aleatoria estudiada tiene al menos aproximadamente una distribución en forma de campana. El diagrama de tallo y hojas es una ayuda para determinar si esta suposición es razonable o no. Por ejemplo, nos sorprenderíamos si nos dijeran que la variable aleatoria X, magnitud de un terremoto en California medido en la escala de Richter, tiene una distribución en forma de campana. El diagrama de tallo y hojas de la Figura 1.4*d* no parece en absoluto una campana. Por otra parte, el diagrama de la Figura 1.5, aunque no es perfectamente simétrico, tiende a aproximarse a la forma de una campana. No nos sorprenderíamos si nos dijeran que X, perímetro craneal de un niño al nacer, tiene una distribución en forma de campana.

EJERCICIOS 1.2

1. Una importante variable usada para medir el estado de desarrollo del SIDA en pacientes infectados es la relación de linfocitos T colaboradores y linfocitos T supresores. El rango

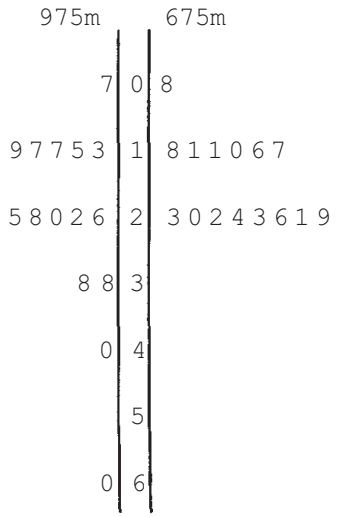


Figura 1.7. Diagrama de tallo y hojas adosado para la variable X, crecimiento en 10 años de robles americanos a dos altitudes diferentes.

normal para esta variable es de 1.0 a 2.9. Los siguientes datos se obtuvieron en pacientes afectados de SIDA seleccionados aleatoriamente: (Basado en información encontrada en *Interpretative Data Guide,* Laboratorios ARUP, 1996.)

0.45	0.52	0.98	0.62	0.62
0.40	0.78	0.53	0.71	0.71
0.66	0.71	0.68	0.70	0.51
0.79	0.61	0.67	0.81	0.84
0.82	0.91	0.90	1.00	1.10

a) Construir un diagrama de tallo y hojas para estos datos, usando como tallos 4, 5, 6, 7, 8, 9, 10 y 11. Representar 1.00 como 10/0 y 1.10 como 11/0. (Recuérdese que el punto decimal no aparece como tal en el diagrama como parte del tallo.)
b) Comentar la forma sugerida por estos datos.
c) ¿Parece probable que la relación linfocitos T colaboradores/T supresores sea inferior en pacientes con SIDA que en pacientes sanos? Expliquese.

2. Se considera como *gran derrumbamiento* aquél en el que los escombros han recorrido una distancia sobre el suelo plano, o ligeramente inclinado, varias veces mayor que su altura de caída. Se realizó un estudio del alcance (distancia recorrida por los escombros) de estos derrumbamientos y se obtuvieron los siguientes datos. (Basado en los datos publicados por Charles Campbell, «Self-Lubrication for Long Runout Landslides», *Journal of Geology,* noviembre de 1989, págs. 653-665.)

Alcance, km					
1.4	9.8	3.2	7.1	7.9	8.6
6.1	10.3	4.0	8.6	6.7	6.6
6.2	6.8	7.2	11.5	3.4	5.8
2.7	5.6	8.3	9.3	5.8	6.8

a) Construir un diagrama de tallo y hojas para estos datos. Utilícese la parte entera de cada número como tallo y el primer dígito tras el decimal como hoja.
b) ¿Piensa que en el futuro sería raro encontrar un derrumbamiento con un alcance dé 10 ó más kilómetros? Razónese.
c) Mediante una simple inspección, dé una aproximación del alcance medio de estos derrumbamietos.

3. Se ha desarrollado una nueva vacuna contra la difteria para aplicarla a niños. El nivel dé protección estándar obtenido por las antiguas vacunas era de 1 μg/mL, un mes después de la inmunización. Transcurrido un mes, se han obtenido estos datos del nivel de protección de la nueva vacuna: (Basado en un informe del *Journal of Family Practice,* enero de 1990, págs. 27-30.)

12.5	13.8	13.0	13.5	13.2
12.2	13.4	14.0	13.6	13.3
13.3	14.1	14.6	13.1	12.1
13.7	13.4	12.8	12.6	12.7

a) Construir un diagrama de tallo y hojas doble para estos datos.
b) ¿Se sorprendería si le dijeran que *X,* nivel de protección transcurrido un mes de la nueva vacuna, tiene una distribución en forma de campana?

c) Mediante la inspección del diagrama de tallos y hojas, haga un cálculo aproximado del nivel de protección medio utilizando la nueva vacuna. ¿Se sorprendería si le dijeran que la nueva vacuna tiende a proporcionar un mayor nivel de protección que la estándar?

4. En un estudio realizado a pacientes clínicos cardíacos varones, el objeto del mismo fue detectar las variables que contribuyen al estrés de estos pacientes. El estrés se midió mediante la puntuación de ansiedad de Hamilton. Estas marcas se encuentran en una escala de 1 a 25, donde el número 18 denota un estrés moderado y el 25, un estrés grave. Se trataba de comparar los dos grupos de pacientes. Se obtuvieron los siguientes datos: (Basado en datos publicados en Earl Burch, Jr., y Jeffery Brandenburg, «Variables Contributing to Distress in Male Cardiac Patients», *Journal of Family Practice,* enero de 1990, págs. 43-47.)

Viven solos				**Viven con otras personas**			
8.6	9.0	9.3	9.6	13.2	15.4	17.5	18.5
9.3	13.5	9.5	8.3	14.7	16.9	14.0	13.3
10.1	11.0	10.3	8.1	14.2	16.0	13.6	14.6
9.4	8.7	10.7	9.4	15.6	17.3	18.1	15.2
14.2	8.2	12.9	11.6	18.0	16.1	17.4	17.2

a) Construir diagramas de tallo y hojas para cada grupo.
b) ¿Alguna de estas distribuciones tiene forma de campana?
c) ¿Alguna de estas distribuciones parece sesgada? Si es así, ¿en qué dirección?
d) Construir un diagrama de tallo y hojas adosado para estos datos.
e) ¿Qué grupo tiende a tener una menor puntuación media de estrés?
f) Basándonos en estos datos, ¿podemos concluir que la puntuación media de estrés para todos los pacientes cardíacos varones que viven solos está por debajo de la puntuación de todos los pacientes cardíacos varones que viven con otras personas? Expliquese.

5. En un experimento se utilizaron saltamontes para estudiar la dirección durante el vuelo. El interés se centraba en la reacción del saltamontes a un estímulo acústico y visual. En cada caso, la variable de interés era la latencia, el tiempo que pasa entre la recepción del estímulo y el movimiento de la cabeza realizado por el saltamontes, que da como resultado una alteración de la marcha. Se obtuvieron estos datos: (Basado en los datos hallados en C. H. F. Rowell, «Descending Interneurones of the Locust Reporting Deviation from Flight Course: What Is Their Role in Steering?» *Journal of Experimental Biology,* vol 146, septiembre de 1989, págs. 177-194.)

Latencia, ms					
Acústico			**Visual**		
86	106	117	72	95	73
102	109	120	99	71	90
103	113	101	102	97	71
99	114	126	75	80	70
108	107	109	100	104	81
100	107	106	103	101	103
115			77	78	89

a) Construir un diagrama de tallo y hojas doble para cada conjunto de datos. Utiliza: los dos primeros dígitos de cada número como tallo. El tallo para un número de dos dígitos como 86 es 08.
b) ¿Se sorprendería si le dijeran que la latencia está simétricamente distribuida en ambos casos?
c) ¿Se sorprendería si le dijeran que la latencia sigue una distribución en forma de campana en ambos casos?
d) ¿Bajo qué estímulo es más dispersa la latencia?

6. En circunstancias normales, en los alimentos existen pequeñas cantidades de cinc y cobre. Estos elementos pueden ser tóxicos y causar problemas al interactuar entre sí e impedir, de esta forma, su absorción por el organismo. Se realizó un estudio sobre los niveles de estos elementos en preparados infantiles. Cada dato puntual representa el nivel medio, en miligramos por litro, para muestras de igual tamaño seleccionadas entre las 16 principales marcas del mercado. (Basado en los datos hallados en B. Lonnerdal, «Trace Element Absorption in Infants as a Foundation to Setting Upper Limits for Trace Elements in Infant Formulas», *Journal of Nutrition,* diciembre de 1989, págs. 1839-1844.)

Cinc				Cobre			
3.0	5.8	5.6	4.8	0.40	0.51	0.47	0.55
5.1	3.6	5.5	4.7	0.56	0.41	0.60	0.46
5.7	5.0	5.9	5.7	0.60	0.61	0.48	0.63
4.4	5.4	4.2	5.3	0.50	0.45	0.62	0.57

a) Construir un diagrama de tallo y hojas doble para cada conjunto de datos.
b) ¿Alguno de los conjuntos de datos muestra cierto sesgo? Si es así, ¿en qué dirección?
c) Se fabrica un nuevo preparado y su nivel medio de cobre se estima en 0.53. ¿Es esta cifra excesivamente alta comparada con las de los preparados que actualmente se encuentran en el mercado? Razónese la respuesta.
d) ¿Sería raro observar en el nuevo preparado un nivel medio de cinc estimado inferior a 4.0? Justifíquese.

7. Construir diagramas de tallo y hojas para la variable edad del Ejemplo 1.1.1 y del Ejercicio 1 de la Sección 1.1. ¿Parecen similares las distribuciones de edad de los dos geriátricos en cuanto a forma y localización? Expliquese.
8. Se realiza un estudio para ayudar a comprender el efecto que tiene el hábito de fumar en los patrones del sueño. La variable aleatoria considerada es *X,* tiempo en minutos que se tarda en quedar dormido. Las muestras de fumadores y no fumadores producen estas observaciones sobre *X.*

No fumadores						Fumadores					
17.2	19.7	18.1	15.1	18.3	17.6	15.1	20.5	17.7	21.3	16.0	24.8
16.2	19.9	19.8	23.6	24.9	20.1	16.8	21.2	18.1	22.1	15.9	25.2
19.8	22.6	20.0	24.1	25.0	21.4	22.8	22.4	19.4	25.2	18.3	25.0
21.2	18.9	22.1	20.6	23.3	20.2	25.8	24.1	15.0	24.1	21.6	16.3
21.1	16.9	23.0	20.1	17.5	21.3	24.3	25.7	15.2	18.0	23.8	17.9
21.8	22.1	21.1	20.5	20.4	20.7	23.2	25.1	16.1	17.2	24.9	19.9
19.5	18.8	19.2	22.4	19.3	17.4	15.7	15.3	19.9	23.1	23.0	25.1

a) Construir un diagrama de tallo y hojas adosado de estos conjuntos de datos. Utilizar los enteros del 15 al 25 inclusive como tallos.
b) ¿Se sorprendería si alguien le dijera que no existe diferencia en cuanto a la distribución de X en los dos grupos? Expliquese.

9. Los incendios de vegetación en pradera, matorral y bosque son un fenómeno común. Algunos son accidentales, pero otros son provocados con el fin de crear hábitats post-fuego que beneficien a plantas y animales. No obstante, el suelo que ha sido expuesto a un alto calentamiento puede esterilizarse. Se realizó un estudio para determinar el efecto de esta esterilización en el crecimiento de plantas, en concreto rábanos. La variable medida fue el peso seco de la planta al cabo de 4 semanas. (Basado en un estudio de Joy Burcham, Departamento de Biología, Universidad de Radford, otoño de 1996.)

Suelo estéril (peso seco en gramos)			**Suelo no** estéril **(peso seco en gramos**		
9	**28**	**26**	**16**	**19**	**13**
10	**18**	**17**	**15**	**14**	**2**
10	**28**	**10**	**7**	**11**	**6**
30	**30**	**11**	**9**	**6**	**3**
25	**35**	**34**	**18**	**14**	**11**
9	**15**		**20**		

a) Construir un diagrama de tallo y hojas doble para cada uno de los conjuntos de datos. ¿Parece tener cada diagrama forma aproximada de campana? ¿Cuál parece estar más disperso? ¿Cuál parece tener la menor tendencia central?
b) Construir un diagrama de tallo y hojas adosado doble para estos datos. Comentar qué reflejan estos datos acerca de la capacidad de crecimiento de los rábanos en suelo estéril.

1.3. DISTRIBUCIONES DE FRECUENCIA: HISTOGRAMAS

En la Sección 1.2, presentamos el diagrama de tallo y hojas, que es una técnica gráfica rápida para organizar conjuntos de datos numéricos en los que hay un gran número de valores distintos. El diagrama de tallo y hojas nos da una idea aproximada de la forma de la distribución, así como de su localización. La técnica funciona bien para los conjuntos de datos que no tienen una dispersión muy grande. Sin embargo, si los datos puntuales cubren una amplia gama de valores, es difícil escoger tallos adecuados. En este caso, necesitamos un sistema alternativo para agrupar los datos de manera que podamos determinar la forma. Los gráficos construidos al efecto para detectar la forma se denominarán *histogramas*. Utilizaremos tres tipos de histogramas (de frecuencias, de frecuencias relativas y de porcentajes). Esta técnica se ha utilizado durante muchos años, atribuyéndose el origen del término *histograma* a Karl Pearson en 1895.

Un histograma de frecuencias es un gráfico de barras verticales u horizontales. Describe la distribución de valores de tal forma que el área de cada barra es proporcional al número de objetos en la categoría o clase representada por la barra. Así, un histograma de datos continuos sirve para el mismo fin que los gráficos de barras presentados en la Sección 1.1. Dado que el conjunto de datos con gran cantidad de valores numéricos distintos no tiene clases naturales obvias, debemos diseñar un método para definirlas. Queremos definir clases de igual tamaño, de tal forma que cada observación corresponda clara y exactamente a una de

ellas. A lo largo de los años se han ido creando muchos de estos métodos. La técnica que se ilustra aquí es una de las que funcionan bien. Se utilizarán estas reglas para la creación de clases. En el Ejemplo 1.3.1 se describirán paso a paso.

Reglas para agrupar datos en categorías o clases

1. Decidir el número de clases deseado. El número elegido depende de la cantidad de observaciones disponibles. La Tabla 1.11 ofrece algunas sugerencias para el número de clases a utilizar en función del tamaño de la muestra. Está basada en la *regla de Sturges,* fórmula desarrollada por H. A. Sturges en 1926. Esta regla afirma que k, número de clases, viene dada por $k = 1 + 3.322 \log_{10} n$ donde n es el tamaño de la muestra. Se utilizó esta fórmula para obtener los números de clase que aparecen en la Tabla 1.11. Puede verificar alguno de estos valores por usted mismo. (H. A. Sturges, «The Choice of a Class lnterval», *Journal of the American Statistical Association,* vol. 21, 1926, págs. 65-66.)
2. Localizar la observación mayor y la menor. Hallar la diferencia entre estas dos observaciones. Restar en el orden «mayor menos menor». A esta diferencia se la denomina *rango* de los datos.
3. Hallar la amplitud (ancho) mínima de la clase requerida para cubrir este rango, dividiendo el rango por el número de clases deseado. Este valor es el mínimo requerido para cubrir el rango, si se toma el límite inferior de la primera clase como el dato menor. Sin embargo, para asegurarse de que ningún dato caiga en un límite, definiremos los límites de tal forma que tengan un decimal más que los datos. Así pues, comenzaremos la primera clase ligeramente por *debajo* del primer dato puntual. Haciéndolo, la amplitud mínima requerida de la clase, necesaria para cubrir el conjunto de datos, no es lo suficientemente grande para atrapar el dato mayor en la última clase. Por esta razón, el ancho real utilizado deberá ser un poco mayor que el mínimo. En concreto, el ancho real de la clase a utilizar se halla redondeando el ancho mínimo hasta la misma cantidad de decimales que los datos. Si, por casualidad, el ancho mínimo ya tiene la misma cantidad de decimales que los datos, también redondearemos hasta una unidad. Por ejemplo, si tenemos datos registrados con un decimal de precisión, y la amplitud mínima requerida para cubrir los datos es 1.7, la elevaremos hasta 1.8 para obtener la amplitud real de la clase a utilizar.

Tabla 1.11. Número de clases sugerido para subdividir datos numéricos en función del tamaño de la muestra

Tamaño muestral	Número de clases
Menos de 16	Datos insuficientes
16-31	5
32-63	6
64-127	7
128-255	8
256-511	9
512-1023	10
1024-2047	11
2048-4095	12
4096-8190	13

4. El límite inferior para la primera clase está \ unidad por debajo de la observación menor. La Tabla 1.12 muestra unidades y medias unidades para diferentes tipos de conjuntos de datos. Los límites para las restantes categorías se hallan añadiendo la amplitud de la clase al valor del límite precedente.

Ejemplo 1.3.1. Se realizó un estudio de llamadas ultrasónicas en jerbos de Mongolia jóve¬ nes. A cada animal se le aisló durante un minuto, cada uno de los 14 primeros días de su vida, y se grabaron los sonidos que produjo. Una variable de interés fue el número de llamadas emitidas. Existen factores asociados al manejo diario de los animales de experimentación que pueden influenciar su comportamiento. Para detectar esta posible fuente de error, se dispuso un grupo de animales de control, animales no manipulados en absoluto, y que se estudiaron el quinto día. Los datos de este día, para los dos grupos, fueron los siguientes:

Número de llamadas por animal					
Experimental			**Control**		
135	149	130 (el menor)	123	112	112
137	151	151	109	105	121
148	143	139	118	106	100
152	154	151	116	115	115
144	146	137	96	120	112
138	145	156 (el mayor)	88	112	122
142	136	138	102	123	128
145	150	144	117	110	124
147	151	142	119	98	109
147	138	155	101	111	90

Consideremos primero los datos de los animales de experimentación. Nuestra tarea es separar estos datos en clases. Véase que hay 30 datos puntuales. Las directrices de la Ta¬ bla 1.11 sugieren que dividamos los datos en cinco clases. Localizamos el dato mayor (156) y el menor (130), que nos servirán para calcular la amplitud de la clase o longitud del intervalo que contiene todos los datos puntuales. En este caso, los datos cubren un intervalo de longitud 156 - 130 = 26 unidades. Para encontrar la amplitud mínima requerida para cada clase, dividimos este valor por el número de clases deseado. Así, la amplitud mínima de la clase será 26/5 = 5.2 unidades. La amplitud de clase que usaremos en la práctica para separar los datos la obtendremos redondeando por exceso la amplitud mínima, hasta obtener un valor con el mismo número de cifras decimales que los datos. Como los datos vienen dados en números enteros, redondearemos la amplitud mínima, 5.2, por exceso, hasta el número entero más

Tabla 1.12. Unidades y medias unidades para los datos registrados en el grado de precisión establecido

Datos mínimos registrados	**Unidad**	2 **unidad**
Número entero	1	0.5
Décimas (1 decimal)	0.1	0.05
Centésimas (2 decimales)	0.01	0.005
Milésimas (3 decimales)	0.001	0.0005
Diezmilésimas (4 decimales)	0.0001	0.00005

próximo, 6. Las clases que usaremos serán pues de amplitud 6. La primera clase comienza 1/2 unidad por debajo de la observación más pequeña. Puesto que los datos tienen valores enteros, vemos en la Tabla 1.12 que una unidad es 1 y media unidad es 0.5. Empezamos la primera clase 0.5 por debajo de la observación más pequeña. Esto es, el límite inferior para la primera clase es 130 - 0.5 = 129.5. Los límites para las restantes clases se encuentran sumando sucesivamente la amplitud de la clase (6) al límite superior precedente, hasta que se hayan cubierto todos los puntos. De esta forma, obtenemos las siguientes cinco clases finitas para los animales en experimentación: 129.5 a 135.5, 135.5 a 141.5, 141.5 a 147.5, 147.5 a 153.5 y 153.5 a 159.5. Obsérvese que, puesto que los límites tienen una cifra decimal más que los datos, ningún dato puntual puede coincidir con uno de ellos, esto es, cada dato debe pertenecer estrictamente a una sola clase. Ahora podemos resumir los datos en una tabla, contando el número de observaciones en cada clase (véase la columna 3 de la Tabla 1.13).

En la Figura 1.8 se muestra el gráfico de la distribución de frecuencias. A este gráfico se le llama *histograma* de frecuencias. Obsérvese que, puesto que las barras tienen la misma anchura, el área de cada barra es directamente proporcional a su altura. Como la altura es igual al número de observaciones de la clase representada por la barra, el área es también directamente proporcional al número de observaciones en su clase, como era de esperar. Esta propiedad de los histogramas es útil puesto que es fácil comparar áreas visualmente. Los histogramas representan un esquema visual de la distribución de frecuencias de los números en el conjunto de datos.

Para los animales del grupo de control, la observación menor es 88, la mayor es 128 y el rango es 40. La amplitud mínima necesaria de la clase, para dividir los datos en cinco clases, es 8. Obsérvese que, aunque éste ya es un número entero, lo aumentamos una unidad hasta 9 para obtener la amplitud real de la clase. Esto se hace para explicar el hecho de que el límite inferior de la primera categoría cae ligeramente por debajo del dato menor. En este caso, este límite es 87.5. La Tabla 1.14 muestra la tabla de frecuencias para los animales de control, y la Figura 1.9, el histograma de frecuencias correspondiente. Obsérvese que el histograma para los animales de control tiene una forma un tanto diferente al de los experimentales. Los histogramas también se sitúan en lugares distintos a lo largo del eje horizontal. Esto implica que, de hecho, pueden existir algunas diferencias básicas en el comportamiento de los dos grupos de animales.

La Figura 1.10 da la versión SAS del histograma mostrado en la Figura 1.8. Obsérvese que las barras en el SAS están etiquetadas por el punto medio de la clase, en lugar de los límites de clase. Por ejemplo, los límites para la primera clase son 129.5 y 135.5. El punto medio es (129.5 + 135.5)/2 = 132.5. El código usado para hacer este histograma en el SAS se halla en la sección de Herramientas Computacionales al final de este capítulo.

Como se mencionó en la Sección 1.1, los recuentos de frecuencia son importantes, pero no explican la verdadera naturaleza concerniente a la distribución de una variable aleatoria. Para situar la frecuencia en perspectiva, también registramos el recuento relativo al total que forma la distribución de frecuencias relativas de la variable. Cuando multiplicamos por 100 la frecuencia relativa, obtenemos el porcentaje relativo. Las Tablas 1.15 y 1.16 resumen lo que sabemos hasta el momento en relación con la distribución del *número de llamadas* de la variable aleatoria para los grupos experimental y de control, respectivamente.

Tabla 1.13. Animales experimentales: distribución de frecuencias

Clase	Límites	Frecuencia
1	129.5 a 135.5	2
2	135.5 a 141.5	7
3	141.5 a 147.5	10
4	147.5 a 153.5	8
5	153.5 a 159.5	3

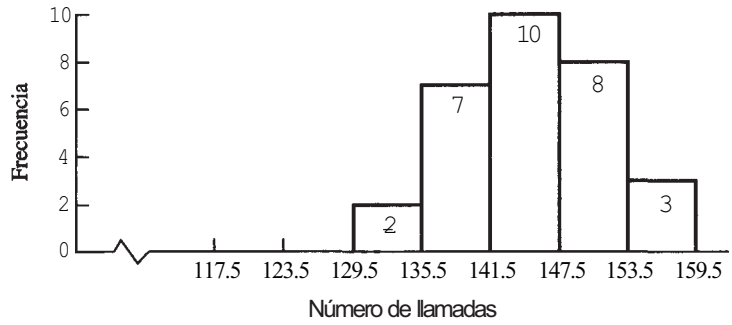


Figura 1.8. Histograma de frecuencias (experimental).

Tabla 1.14. Distribución de frecuencias de los animales de control

Clase	Límites	Frecuencia
1	87.5 a 96.5	3
2	96.5 a 105.5	5
3	105.5 a 114.5	9
4	114.5 a 123.5	11
5	123.5 a 132.5	2

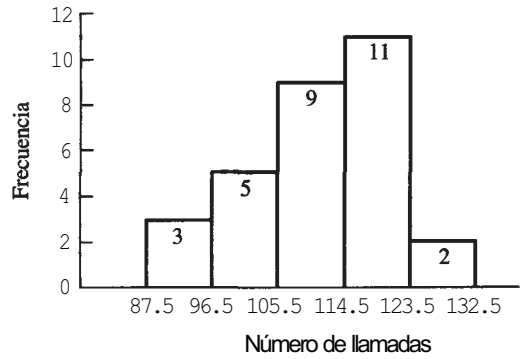


Figura 1.9. Histograma de frecuencias (control).

Otro punto más a tener en cuenta: el procedimiento aquí presentado para construir un histograma funciona bien si los datos no presentan valores extremadamente alejados de la línea general del resto de los datos. Por ejemplo, supóngase que el conjunto de datos de animales experimentales del Ejemplo 1.3.1 incluye el valor 201. Este número es mucho mayor que el resto de los datos. Esto tendrá un gran impacto en los valores de cada rango; de hecho, el rango pasa de 26 a 71. Esto, a su vez, cambia la amplitud de la clase de 5.2 a 14.2. ¿Qué efecto tiene esto en el histograma? Para apreciarlo, véase la Figura 1.11. Lo más importante a destacar es que, al expandir el ancho de cada clase, el grueso de los datos se concentra en dos clases muy grandes. Se pierde el sentido acerca de la distribución de los datos. Es evidente que, cuando un conjunto de datos contiene valores inusuales, el procedimiento citado ha de cambiarse. El Ejercicio 10 (de la Sección 1.3) muestra dos posibles soluciones para este problema.

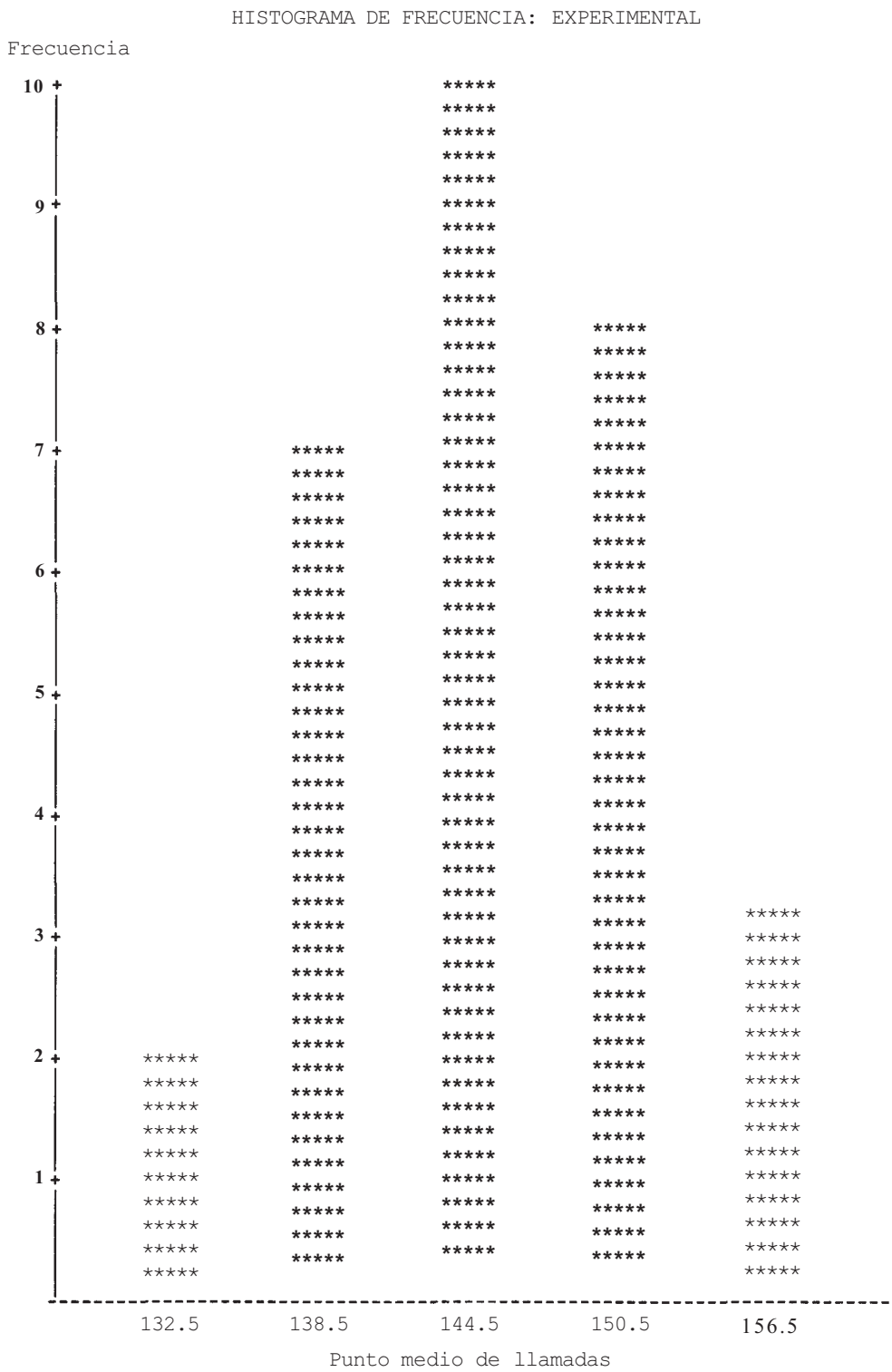


Figura 1.10. Versión SAS del histograma de los datos experimentales del Ejemplo 1.3.1.

Tabla 1.15. Animales experimentales

Clase	Límites	Frecuencia	Frecuencia relativa	Porcentaje
1	129.5 a 135.5	2	0.0667	6.67
2	135.5 a 141.5	7	0.2333	23.33
3	141.5 a 147.5	10	0.3333	33.33
4	147.5 a 153.5	8	0.2667	26.67
5	153.5 a 159.5	2	0.1000	10.00

Tabla 1.16. Animales de control

Clase	Límites	Frecuencia	Frecuencia relativa	Porcentaje
1	87.5 a 96.5	3	0.1000	10.00
2	96.5 a 105.5	5	0.1667	16.67
3	105.5 a 114.5	9	0.3000	30.00
4	114.5 a 123.5	11	0.3667	36.67
5	123.5 a 132.5	2	0.0667	6.67

Distribuciones acumuladas

Además de las distribuciones de frecuencias, frecuencias relativas y porcentajes para las clases, es interesante considerar las distribuciones de frecuencias acumuladas, frecuencias relativas acumuladas y porcentajes acumulados de las variables numéricas. Como se vio en la exposición acerca de datos discretos, los valores acumulados se obtenían sumando. Así pues, la frecuencia acumulada de una clase es el número de observaciones incluidas en o por debajo de la clase; la frecuencia relativa acumulada es la fracción de observaciones incluidas en la clase o por debajo de ella, y el porcentaje acumulado es el porcentaje de observaciones incluidas en o por debajo de la clase. En las Tablas 1.17 y 1.18, se presentan estas frecuencias para los datos del Ejemplo 1.3.1.

La Figura 1.12 muestra un histograma horizontal para los datos de animales experimentales dados en el Ejemplo 1.3.1. Obsérvese que la figura también incluye información sobre las frecuencias, porcentajes, frecuencias acumuladas y porcentajes acumulados mostrados en la Tabla 1.17.

El Ejemplo 1.3.2 ilustra la distribución de un conjunto de datos en el que éstos no son números enteros.

Ejemplo 1.3.2. Mucha gente manifiesta reacciones de alergia sistémica a las picaduras de insectos. Estas reacciones varían de paciente a paciente, no sólo en cuanto a gravedad, sino también en el tiempo transcurrido hasta que se inicia la reacción. Los datos siguientes representan este «tiempo de inicio hasta la reacción» en 40 pacientes que experimentaron una reacción sistémica a la picadura de abeja.

(Datos en minutos.)

10.5	11.2	9.9	15.0	11.4	12.7	16.5	10.1
12.7	11.4	11.6	6.2	7.9	8.3	10.9	8.1
3.8	10.5	11.7	8.4	12.5	11.2	9.1	10.4
9.1	13.4	12.3	5.9	11.4	8.8	7.4	8.6
13.6	14.7	11.5	11.5	10.9	9.8	12.9	9.9

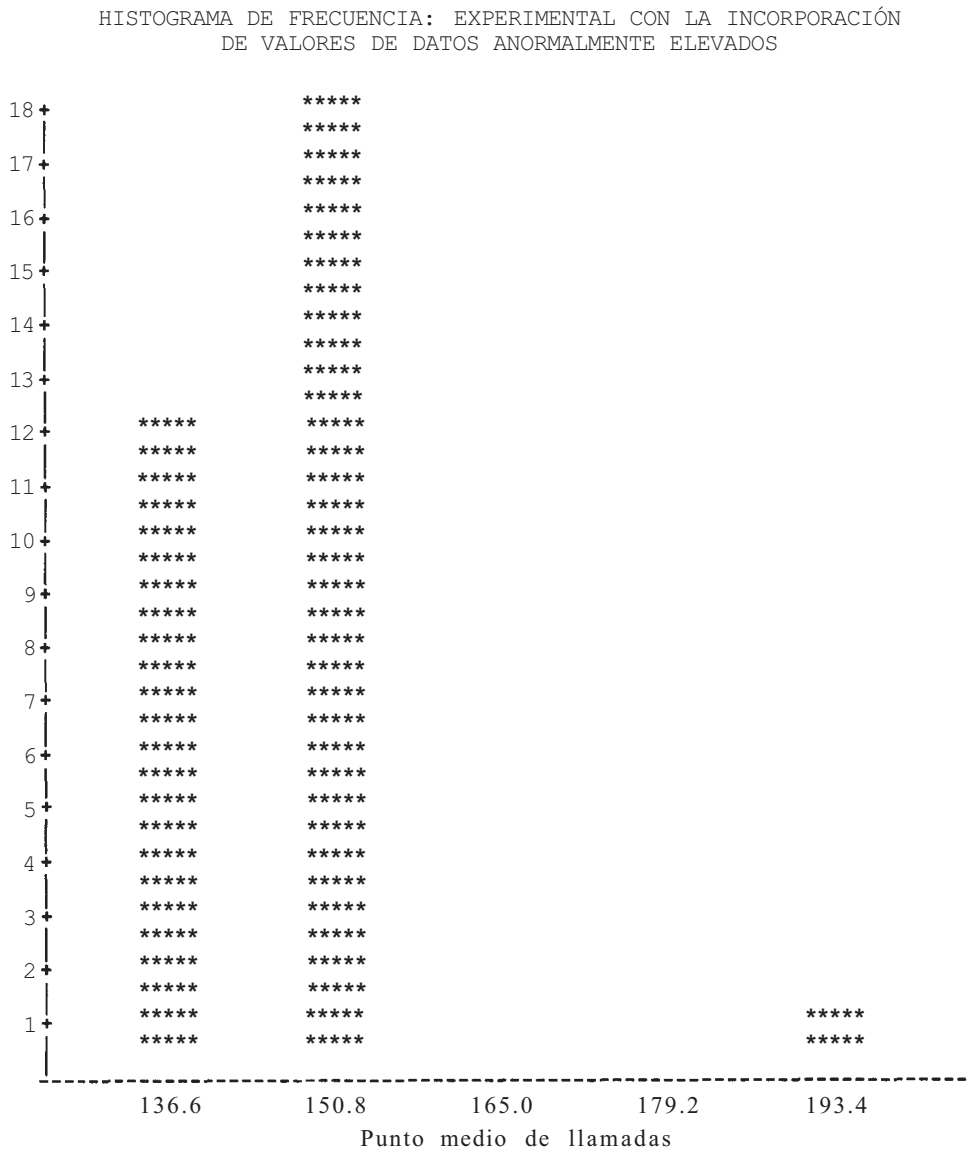


Figura 1.11

En la Tabla 1.11, vemos que es adecuado dividir estos datos en seis clases. La mayor observación es 16.5, la menor es 3.8 y el rango es 12.7. La amplitud mínima de la clase requerida para cubrir el rango es 12.7/6 = 2.12. Puesto que los datos están registrados con precisión de un decimal, redondeamos 2.12 hasta 2.2 para obtener la amplitud efectiva de la clase. En la Tabla 1.12 vemos que $^1/_2$ unidad para datos con una cifra decimal es 0.05. El límite inferior para la primera clase es 0.05 por debajo de la observación menor, o 3.80 - 0.05 = 3.75. En la Tabla 1.19 se muestra la distribución de frecuencias completa, así como las de la frecuencias acumuladas de los datos.

Tabla 1.17. Animales experimentales: distribuciones acumuladas

Clase	Límites	Frecuencia acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
1	129.5 a 135.5	2	2/30 = 0.0667	6.67
2	135.5 a 141.5	9	9/30 = 0.3000	30.00
3	141.5 a 147.5	19	19/30 = 0.6333	63.33
4	147.5 a 153.5	27	27/30 = 0.900	90.00
5	153.5 a 159.5	30	30/30 = 1.000	100.00

Tabla 1.18. Animales de control: distribuciones acumuladas

Clase	Límites	Frecuencia acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
1	87.5 a 96.5	3	3/30 = 0.1000	10.00
2	96.5 a 105.5	8	8/30 = 0.2667	26.67
3	105.5 a 114.5	17	17/30 = 0.5667	56.67
4	114.5 a 123.5	28	28/30 = 0.9334	93.34
5	123.5 a 132.5	30	30/30=1.0000	100.00

TABLA RESUMEN E HISTOGRAMA HORIZONTAL: EXPERIMENTAL

Punto medio de LLAMADAS		F	**F acumulada**	%	**% acumulado**
132.5	*******	2	2	6.45	6.45
138.5	**********	7	9	22.58	29.03
144.5	**********	10	19	32.26	61.29
150.5	**********	8	27	25.81	87.10
156.5	**********	4	31	12.90	100.0

1 2 3 4 5 6 7 8 9 1 0

Frecuencia (F)

Figura 1.12. Histograma horizontal mostrando la distribución de frecuencias, porcentajes, frecuencias acumuladas y porcentajes acumulados para los datos de animales experimentales del Ejemplo 1.3.1.

Tabla 1.19. Distribución de la variable tiempo de inicio de la reacción

Clase	Límites	Frecuencia	Frecuencia relativa	Porcentaje	Frecuencia acumulada	Frecuencia relativa acumulada	Porcentaje acumulado
1	3.75 a 5.95	2	2/40 = 0.500	5	2	2/40 = 0.050	5
2	5.95 a 8.15	4	4/40 = 0.100	10	6	6/40 = 0.150	15
3	8.15 a 10.35	10	10/40 = 0.250	25	16	16/40 = 0.400	40
4	10.35 a 12.55	16	16/40 = 0.400	40	32	32/40 = 0.800	80
5	12.55 a 14.75	6	6/40 = 0.150	15	38	38/40 = 0.950	95
6	14.75 a 16.95	2	2/40 = 0.050	5	40	40/40 = 1.000	100

EJERCICIOS 1.3

1. Una variable de interés en un estudio del cangrejo *Xanthidae* (pequeño cangrejo que habita en las proximidades de Gloucester Point, Virginia) es el número de huevos pues tos por individuo. Las siguientes son las observaciones obtenidas para 45 cangrejos:

1959	4534	7020	6725	6964	7428	9359	9166
2802	2462	4000	3378	7343	4189	8973	4327
2412	7624	1548	4801	737	5321	849	5749
6837	8639	7417	6082	10241	962	3894	1801
5099	6627	4484	5633	4148	6588	5847	4632
6472	8372	8225	6142	12130			

   a) Determinar la observación mayor y la más pequeña.
   b) Determinar el rango.
   c) Si se utilizan las directrices dadas en la Tabla 1.11, ¿cuántas clases se necesitan para subdividir estos datos?
   d) ¿Cuál es la amplitud mínima necesaria por clase para cubrir el intervalo, si se emplean 6 clases?
   e) Determinar la amplitud efectiva de la clase que se usará para la partición del cor junto de datos.
   f) Determinar el límite inferior para la primera clase.
   g) Determinar la tabla de distribución para el conjunto de datos y construir un histc grama de frecuencias relativas.
   h) ¿Se sorprendería si le dijeran que la variable aleatoria *número de huevos desovadas* presenta una distribución en forma de campana? Expliqúese.

2. En el estudio de pautas de crecimiento en niños, una variable importante es la edad dí 1 niño cuando comienza el crecimiento rápido de la adolescencia. Las siguientes observa ciones se obtuvieron en un estudio de 35 chicos y 40 chicas (edad en años).

Chicos						
16.0	14.9	14.1	14.8	14.4	14.0	14.6
15.2	14.7	13.6	14.6	16.1	13.2	13.2
14.9	14.1	15.4	15.3	14.4	14.8	14.8
13.5	15.1	13.5	15.0	14.6	15.4	15.9
13.7	15.9	14.7	14.5	14.4	13.8	15.3

Chicas							
12.2	13.7	13.3	12.3	12.5	12.9	11.9	11.6
13.4	12.4	12.6	13.5	12.5	13.4	11.7	13.5
13.7	12.1	14.1	11.8	12.8	12.9	11.6	14.3
13.1	13.3	13.5	14.7	12.3	11.6	13.1	12.6
12.7	12.7	12.0	11.4	13.5	12.4	12.1	12.1

a) Dividir cada conjunto de datos en el número de clases apropiado según la Tabla 1.11, utilizando el método expuesto en esta sección.
b) Construir un histograma de frecuencias relativas para cada conjunto de datos. Comentar las semejanzas o diferencias llamativas entre los histogramas.

3. En pacientes con distrofia muscular progresiva (enfermedad de Duchenne) la actividad de la creatina cinasa sérica se eleva llamativamente sobre el valor normal de menos de 50 unidades/litro. Los siguientes datos son niveles séricos de creatina cinasa medidos en 24 pacientes jóvenes con la enfermedad de Duchenne (en unidades por litro):

3720	3796	3340	5600	3802	3580
5500	2000	1571	2360	1500	1840
3723	3790	3345	3805	5595	3577
1995	5504	2055	1573	1835	1505

a) Dividir el conjunto de datos en la cantidad de clases sugerida en la Tabla 1.11.
b) Determinar la tabla de distribución para el conjunto de datos.
c) Construir un histograma de porcentajes y describir su forma y localización.

4. *Polígono de frecuencias relativas acumuladas (ojiva).* Si la variable aleatoria que se está estudiando es continua, es posible utilizar un gráfico de la distribución acumulada para así conseguir una valiosa información. El gráfico se obtiene trazando el límite superior de cada clase sobre el eje horizontal y la frecuencia relativa acumulada sobre el eje vertical. A continuación, los puntos obtenidos se unen con segmentos rectilíneos. El gráfico se completa trazando un segmento de forma que pase por el punto 0 situado en el límite inferior de la primera clase. Este gráfico se denomina polígono de frecuencias relativas acumuladas u ojiva. Por ejemplo, considere los datos del Ejemplo 1.3.2. La variable «tiempo transcurrido hasta la aparición de alguna reacción a una picadura de insecto», es continua. El polígono de frecuencias relativas acumuladas se construye a partir de la información de la Tabla 1.19 y se muestra en la Figura 1.13. Ahora, haremos dos preguntas:

   1. ¿Qué proporción aproximada de pacientes ha experimentado una reacción en los 10 primeros minutos?
   2. ¿En qué instante se ha producido la reacción para la mitad de los pacientes?

   La primera pregunta puede responderse gráficamente localizando el 10 sobre el eje horizontal, proyectando una línea vertical hasta la ojiva y, a continuación, proyectando una línea horizontal sobre el eje vertical (Fig. 1.14). La proporción buscada es aproximadamente 0.37. La segunda pregunta se responde colocando 0.5 sobre el eje vertical y, a continuación, invirtiendo el proceso. La respuesta es de aproximadamente 11 minutos (Fig. 1.14).

   a) Construir un polígono de frecuencias relativas acumuladas para cada conjunto de datos del Ejercicio 2. Utilizar la ojiva para calcular aproximadamente la edad en la

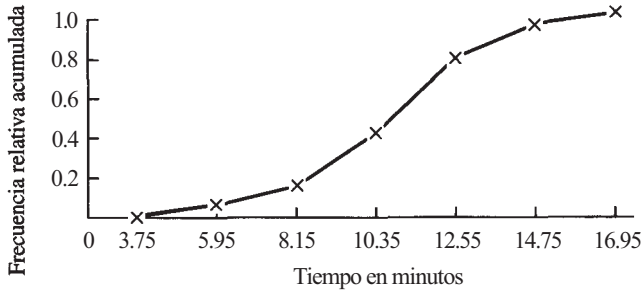


Figura 1.13. Polígono de frecuencias relativas acumuladas.

que el 50% de los chicos ha iniciado el crecimiento rápido de la adolescencia; hacer lo mismo para las chicas. ¿Existe una diferencia notable entre los dos valores?

b) A la edad de 12 años, ¿qué porcentaje aproximado de chicas ha experimentado el inicio del crecimiento rápido de la adolescencia? A la edad de 14 años, ¿qué porcentaje aproximado de chicos ha experimentado el comienzo del crecimiento rápido de la adolescencia?

5. *a)* Dibujar un polígono de frecuencias relativas acumuladas para el conjunto de datos del Ejercicio 3.

   b) Basándose en estos datos, ¿qué proporción de pacientes con la enfermedad de Duchenne tienen niveles séricos de creatina cinasa de al menos 50 veces por encima del valor normal?

6. En un estudio diseñado para correlacionar el cambio estacional de testosterona en plasma con el ciclo reproductor de los saurios, se obtuvieron los siguientes datos de una muestra de 33 saurios de una especie particular, capturada en el monte August. (Los niveles de testosterona están en nanogramos por mililitro.)

7.5	7.2	3.0	12.1	15.1	12.1	11.5	11.8	7.2	13.2	13.6
8.2	9.5	8.4	13.3	12.5	12.4	2.1	10.7	9.4	6.7	6.8
6.1	8.3	7.9	6.0	7.6	13.2	4.5	9.3	8.1	3.5	9.0

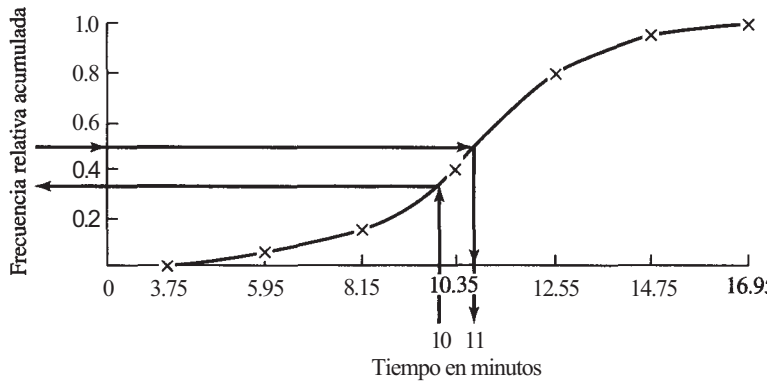


Figura 1.14. Método de proyección para estimar porcentajes.

En octubre, una muestra de 40 saurios de la misma especie reveló los siguientes niveles de testosterona en plasma:

43.7	37.2	29.0	31.6	47.5	48.3	38.3	29.7
32.5	45.2	36.1	30.5	37.2	50.5	36.9	44.5
35.9	28.7	37.5	30.2	36.9	43.2	27.0	26.2
41.8	26.4	34.3	28.6	35.9	22.0	45.4	30.3
29.8	46.1	42.7	31.5	37.4	25.1	27.2	45.0

a) Hallar la tabla de distribución para cada conjunto de datos, utilizando el número de clases sugerido en la Tabla 1.11.
b) Construir histogramas de frecuencias relativas para ambos conjuntos de datos.
c) Comparar los histogramas, teniendo en cuenta la observación adicional de que las características sexuales secundarias se desarrollan en saurios jóvenes en el invierno y el apareamiento tiene lugar en verano.

7. Actualmente se realizan esfuerzos para elaborar fibras textiles de fibra de turba. Esto creará una fuente de materiales económicos para las industrias textil y papelera. Una variable estudiada es *X,* porcentaje del contenido en ceniza de una determinada turbera. Supongamos que una muestra aleatoria de 50 turberas produce estas observaciones: (Basado en los datos hallados en «Peat Fibre: Is There Scope in Textiles?» *Textile Horizons,* vol. 2, n.° 10, octubre de 1982, pág. 24.)

0.5	1.8	4.0	1.0	2.0
1.1	1.6	2.3	3.5	2.2
2.0	3.8	3.0	2.3	1.8
3.6	2.4	0.8	3.4	1.4
1.9	2.3	1.2	1.9	2.3
2.6	3.1	2.5	1.7	5.0
1.3	3.0	2.7	1.2	1.5
3.2	2.4	2.5	1.9	3.1
2.4	2.8	2.7	4.5	2.1
1.5	0.7	3.7	1.8	1.7

a) Construir un diagrama de tallo y hojas para estos datos. Utilizar los números 0, 1, 2, 3, 4, 5 como tallos.
b) ¿Existe alguna razón para sospechar que *X* no sigue una distribución en forma de campana?
c) Utilizar el método descrito en esta sección para subdividir los datos. Utilizar la Tabla 1.11 para determinar el número adecuado de clases.
d) Construir la tabla de distribución para estos datos.

8. La exposición intensa al cadmio produce dificultad respiratoria, daños en los ríñones y el hígado, y puede ocasionar la muerte. Por esta razón, se controla el nivel de polvo de cadmio y de humo de óxido de cadmio en el aire. Este nivel se mide en miligramos de cadmio por metro cúbico de aire. Una muestra de 35 lecturas proporciona estos datos: (Basado en un informe de *Environmental Management,* septiembre de 1981, pág. 414.)

0.044	0.030	0.052	0.044	0.046
0.020	0.066	0.052	0.049	0.030
0.040	0.045	0.039	0.039	0.039
0.057	0.050	0.056	0.061	0.042
0.055	0.037	0.062	0.062	0.070
0.061	0.061	0.058	0.053	0.060
0.047	0.051	0.054	0.042	0.051

a) Construir un diagrama de tallo y hojas para estos datos. Utilizar los números 02,03, 04, 05, 06 y 07 como tallos.
b) ¿Se sorprendería si le dijeran que la variable aleatoria *X,* nivel de cadmio del aire, sigue una distribución en forma de campana?
c) Utilizar el método descrito en esta sección para distribuir estos datos en seis clases.
d) Construir la tabla de distribución para estos datos.
e) Construir un histograma de frecuencias relativas para estos datos. ¿Tiene el histograma forma de campana?
f) Construir un polígono de frecuencias relativas acumuladas para estos datos. Utilizar la ojiva para obtener por aproximación el punto al que le correspondería el 50 % de las lecturas.

9. Se lleva a cabo un estudio para comparar la diversidad de plantas hallada en una porción incendiada y otra no incendiada de un bosque nacional. Para cada zona, la variable medida fue el índice de Comparación Secuencial (ICS). Un alto valor de ICS indica que se encontraron especies muy diferentes en ese sitio; un valor bajo de ICS indica la presencia de sólo unas pocas especies. Los siguientes datos se obtuvieron en muestras de 35 sitios incendiados y 35 no incendiados. Basado en información encontrada por Jackie Cummings, Departamento de Biología, Universidad de Radford, otoño de 1996.)

Incendiados				
0.155	1.317	0.196	1.753	0.503
0.303	1.564	1.795	2.017	0.901
1.686	0.591	2.527	0.733	1.555
1.055	0.109	1.000	2.377	0.729
1.214	1.523	0.459	1.192	1.377
0.713	1.269	1.418	1.368	1.469
2.067	2.479	1.423	2.179	0.141
No incendiados				
1.856	0.892	1.662	0.804	0.998
1.518	1.507	2.122	0.380	1.234
0.382	1.187	2.203	0.648	0.517
0.498	0.029	0.383	0.489	0.010
1.044	0.935	0.374	0.423	1.483
1.624	0.559	0.939	0.171	0.805
1.282	0.544	1.505	0.635	0.777

a) Una versión de SAS del histograma y de la tabla resumen de cada grupo se muestra en la Figura 1.15. Se permitió al SAS elegir sus propias clases para construir estos gráficos. Es decir, no dimos al SAS los puntos medios obtenidos por el procedi miento indicado en el texto. ¿Utiliza al SAS por defecto el algoritmo especifica en el texto para determinar el número de clases, límites y puntos medios? Razone su respuesta.
b) ¿Qué amplitud de clase usa el SAS en cada caso?

Grupo = incendiado

Punto medio de ICS		F	F acumulada	%	% acumulado
0.2	**********	5	5	14.29	14.29
0.6	**********	6	11	17.14	31.43
1.0	********	4	15	11.43	42.86
1.4	**********	11	26	31.43	74.29
1.8	******	3	29	8.57	82.86
2.2	********	4	33	11.43	94.29
2.6	****	2	35	5.71	100.00

2 4 8 10

Frecuencia (F)

Grupo = no incendiado

Punto medio de ICS		F	F acumulada	%	% acumulado
0.2	**********	7	7	20.00	20.00
0.6	**********	9	16	25.71	45.71
1.0	**********	8	24	22.86	68.57
1.4	**********	6	30	17.14	85.71
1.8	**********	3	33	8.57	94.29
2.2	********	2	35	5.71	100.00

1 2 3 4 5 6 7 8 9

Frecuencia (F)

Figura 1.15. Análisis SAS de los datos del Ejercicio 9.

c) ¿Cuál es el límite superior para la primera clase del SAS? ¿Cae algún punto en ese límite? ¿Podría caer en este límite un futuro dato, con una precisión de tres decimales?

d) Si usamos el algoritmo proporcionado en el texto para encontrar los límites, ¿cuáles serían los de la primera clase en cada caso? ¿Podría caer en este límite un futuro dato, con una precisión de tres decimales?

e) Si se dispone del SAS, construya los histogramas usando el algoritmo del texto y compare los resultados con los aquí mostrados.

f) Si se dispone del SAS, cambie el dato 0.459 en los sitios incendiados por 0.400. Construya de nuevo el gráfico y vea cómo maneja el SAS un punto que cae en un límite de clase.

10. *Histogramas: valores atípicos.* Considere los datos experimentales del Ejemplo 1.3.1 con el dato extra, 201, añadido. Presentamos dos soluciones para representar los datos en un histograma. En cada caso, el área de la barra de una clase es proporcional al número de observaciones en esa clase.

    Añadir una clase. Mantenemos los límites entre clases para el grueso de los datos como se definieron en el texto. Añadimos una clase más para cubrir el punto atípicp. Necesariamente, el ancho de esta clase ha de ser mayor que las otras. Así que, para que el área de la barra sea proporcional al número de observaciones, la altura de la misma no podrá ser igual al número de observaciones en esa clase.

    a) Definir los límites de la clase añadida como 159.5 y 201.5. ¿Cuál es la amplitud de la clase?

    b) ¿Cuál es el área total del histograma de frecuencias para la totalidad de los datos? ¿Con qué área está representada cada observación?

    c) ¿Cuál debería ser el área de la barra que representa la clase añadida?

    d) ¿Cuál debería ser la altura de la barra de la clase añadida para que su área sea 1a encontrada en c?

    Añadir más clases. Mantenemos los límites para el volumen de los datos como Se definieron en el texto. Añadimos más clases con la misma amplitud, hasta que el valor atípico esté cubierto. Normalmente, esto originará algunas clases vacías y algunos huecos en el histograma, pero preserva la idea de que las clases sean de la misma anchura y de que la altura de la barra en un histograma de frecuencias sea la frecuencia de la clase.

    e) Considérese el resultado del SAS mostrado en la Figura 1.16. ¿Cuántas clases se han añadido para cubrir el punto 201?

    f) Si se añadiese el punto 185 a los datos de control del Ejemplo 1.3.1, ¿cuántas clases adicionales se necesitarían para cubrirlo? ¿Cuántas deberían estar vacías?

11. Construir, a partir de los datos del Ejemplo 1.3.1, un diagrama de tallo y hojas adosado doble. Es decir, representar los datos experimentales y los de control en el mismo diagrama de tallos y hojas usando como tallos 08, 08, 09, 09, 10, 10, 11, 11, 12, 12, 13, 13, 14, 14, 15, 15. Estos diagramas señalan claramente una diferencia principal entre estos dos conjuntos de datos. ¿Cuál es esta diferencia? ¿Qué técnica, diagrama de tallo y hojas, o histogramas da una mejor impresión de la forma potencial de la población?

1.4. MEDIDAS DE POSICIÓN O TENDENCIA CENTRAL

Recordemos que el objetivo principal de la estadística es extraer conclusiones sobre una población basándonos en la información obtenida de una muestra. Ya hemos visto dos técnicas que pueden utilizarse para determinar la forma de la distribución, como son el diagrama de tallos y hojas y el histograma. Estos gráficos también dan una idea aproximada de la localización. En esta sección consideraremos dos medidas de localización: la media y la mediana. La media poblacional es el valor medio de la variable aleatoria sobre toda la población. Su valor se indica con μ. La mediana de la población es el punto M que tiene la propiedad de que aproximadamente el 50% de los valores de la población está en o por debajo de M y el resto, por encima de M. Tanto μ como M son parámetros. Sus valores son desconocidos y no pueden hallarse a partir de una muestra. Sin embargo, es posible estimarlos a partir de una muestra mediante la media muestral y la mediana muestral. Estos estadísticos, que definiremos en esta sección, se simbolizan mediante $\bar{x}$ y $\tilde{x}$, respectivamente. La Figura 1.17 ilustra la idea.

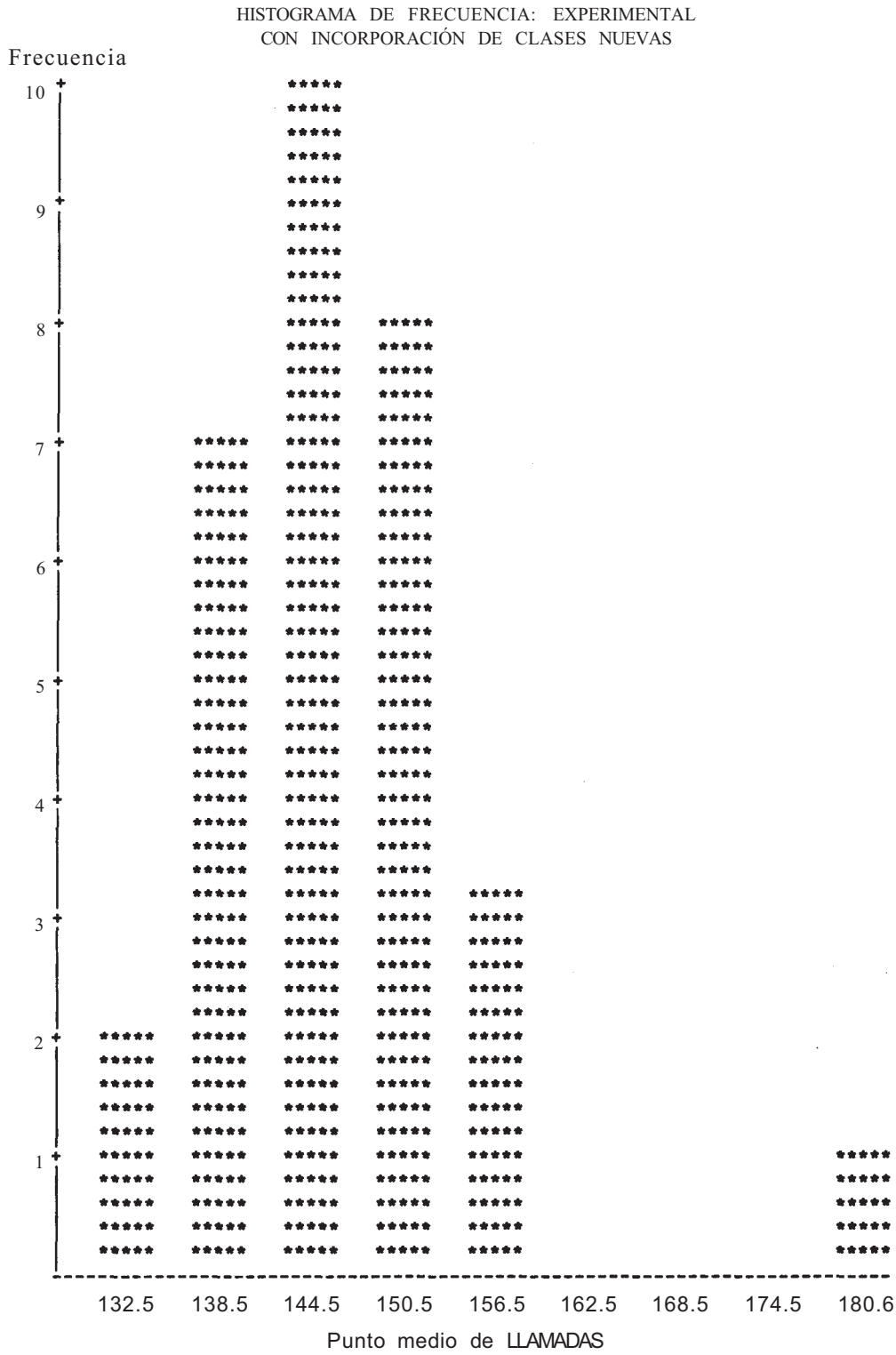


Figura 1.16. Adición de clases extras para cubrir datos atípicos.

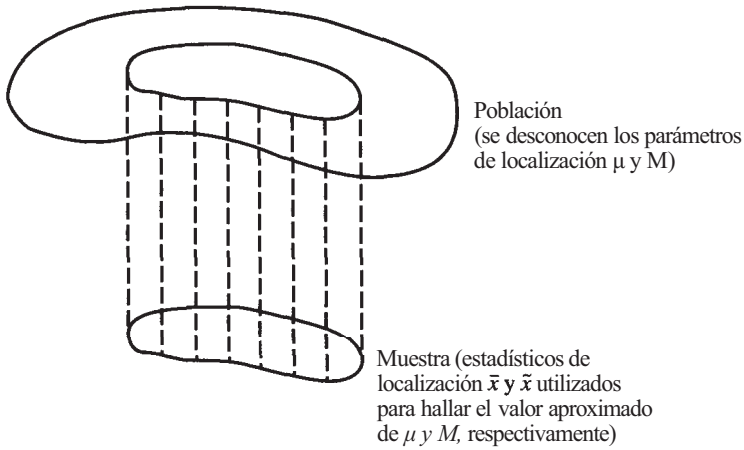


Figura 1.17. Los estadísticos $\bar{x}$ y $\tilde{x}$ se utilizan para describir la posición en la muestra y para hallar los valores aproximados de μ y *M,* respectivamente.

Media muestral

La medida de localización más común es la media o valor medio. La media de una muestra es el promedio aritmético de las observaciones. La definición 1.4.1 se da mediante el uso de la notación sumatoria. En esta notación, se utiliza la letra griega Σ (sigma mayúscula) para indicar adición. Por ejemplo, Σ x significa sumar los datos observados. Si no está familiarizado con esta notación, consulte el Apéndice A.

Definición 1.4.1. Media muestral. Sean $x_1, x_2, \ldots, x_n$ un conjunto de n observaciones de la variable X. A la media aritmética de estos valores se le llama *media muestral* y se representa por $\bar{x}$ (enunciado «x barra»). La fórmula es:

$$\text{Media muestral} = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

Siempre que se hace un cálculo, surge la pregunta, ¿cuántas cifras decimales deben darse en la respuesta final? Esto es especialmente cierto ahora que la mayoría de cálculos se efectúa con calculadoras o mediante programas de ordenador, que, habitualmente, nos proporcionan resultados con una precisión de hasta 8 cifras decimales. A menos que los mismos datos se expresen con este nivel de precisión, no es necesario ni deseable escribir todos los valores. No existen normas estrictas concernientes al número de decimales que deben emplearse. Utilizaremos la convención de que, al calcular medias, el resultado se expresará con hasta una cifra decimal más que los datos. La última cifra decimal se hallará mejor por redondeo que por truncamiento. Por ejemplo, si promediamos los valores 23.72, 89.10 y 112.56 mediante una calculadora científica estándar, obtendremos el número 75.12666667. Esta media se escribirá como 75.127. Una excepción a esta convención se hará cuando x tenga una expansión decimal finita. Por ejemplo, si promediamos 1, 2, 5 y 1, la media obtenida será 9/4 = 2.25. *Advertencia:* Al hacer una serie de cálculos, no redondee los resultados hasta que finalicen todos. De lo contrario, se acumularán los errores de redondeo.

Ejemplo 1.4.1. En un estudio de contaminación de agua, se tomaron muestras de mejillones de dos localidades de Suecia. Las variables de interés fueron *X e Y,* concentración de plomo en

el mejillón (medida en miligramos por gramo de peso seco) en las áreas de Smygehuk y el canal de Falsterbo, respectivamente. Los datos obtenidos fueron:

Smygehuk	Canal de Falsterbo
106.3	113.0
209.3	140.5
246.5	163.3
252.3	185.7
294.4	202.5
	207.2

Dado que estamos haciendo un muestreo de dos poblaciones distintas, tenemos dos medias muestrales $\bar{x}$ e $\bar{y}$, con valores:

$$\bar{x} = \frac{106.3 + 209.3 + \cdots + 294.4}{5} = 221.76$$

$$\bar{y} = \frac{113.0 + 140.5 + \cdots + 207.2}{6} = 168.70$$

¿Podemos deducir, de estos estadísticos, que la concentración media de plomo hallada en los mejillones de Smygehuk, μ_x, sobrepasa a la media en el Canal de Falsterbo, μ_y? La respuesta a esta pregunta es no. No tenemos los datos de la población; sólo tenemos pequeñas muestras extraídas de estas poblaciones. Las medias muestrales difieren bastante, lo que nos lleva a sospechar que μ_x no es igual a μ_y. Sin embargo, no podemos afirmarlo con un 100 % de certeza.

Obsérvese que $\bar{X}$ es una variable aleatoria. Su valor variará de una muestra a otra. Por ejemplo, si hemos tomado 10 muestras de la población de mejillones de Smygehuk, cada una de tamaño 5, y hemos calculado la media muestral de cada muestra, no podemos esperar que se obtenga el mismo valor en cada caso. Los valores $\bar{x}$ diferirán entre sí debido a los factores aleatorios del proceso de muestreo. En el Capítulo 7 se estudiará la distribución de $\bar{X}$.

La mayoría de calculadoras científicas posee funciones estadísticas, de manera que calculan automáticamente $\bar{x}$ según los datos introducidos en el modo «estadístico». Ya que los pasos a seguir para hallar $\bar{x}$ difieren de una calculadora a otra, no podemos dar aquí instrucciones específicas para una determinada calculadora. Consulte el manual de su calculadora para aprender a calcular la media muestral. Los pasos requeridos por la TI83 se dan en la sección Herramientas Computacionales, al final del capítulo.

Mediana muestral

La segunda medida de localización es la mediana. En términos coloquiales, la mediana es el número situado en el «medio» del conjunto de datos ordenados.

Definición 1.4.2. Supongamos que $x_1, x_2, \ldots, x_n$ es una muestra de observaciones dispuestas de menor a mayor. La mediana muestral es la observación que ocupa el lugar central si n es impar. Si n es par, es el promedio de las dos observaciones centrales. Representaremos la mediana muestral con $\tilde{x}$ (enunciado «x tilde»).

Si n es pequeña, es fácil detectar el centro del conjunto de datos. Sin embargo, si n es grande, es útil tener una fórmula que detecte la situación de la observación u observaciones centrales. A continuación se presenta la fórmula (su uso se ilustra en el Ejemplo 1.4.2).

$$\text{Situación de la mediana} = \frac{n+1}{2}$$

Ejemplo 1.4.2. Consideremos que los datos del Ejemplo 1.4.1 anterior ya han sido ordenados. Dado que $n = 5$ para la muestra de Smygehuk, la situación de la mediana debe ser $(n + 1)/2 = 6/2 = 3$. La mediana muestral es la tercera observación. Es decir, $\tilde{x} = 246.5$. Para los datos del Canal de Falsterbo, $n = 6$ y $(n + 1)/2 = 7/2 = 3.5$. Interpretamos que esto signifij que la mediana es el promedio de las observaciones tercera y cuarta. De donde:

$$\tilde{y} = \frac{163.3 + 185.7}{2} = 174.5$$

Al resumir los datos, es útil considerar tanto la media muestral como la mediana muestral. Ambas miden la localización, pero de forma ligeramente diferente.

La mediana muestral tiene una ventaja sobre la media muestral como medida de localiz ción: es *resistente.* Esto significa que su valor sólo cambia ligeramente cuando se borra o sustituye una pequeña parte de los datos por nuevos números que pueden ser muy diferentes de los originales. Esta propiedad es deseable cuando un conjunto de datos contiene uno que queda lejos del resto de datos puntuales. Un dato de este tipo se denomina *outlier o* «atípico»: Su presencia puede afectar drásticamente al valor de la media muestral; tiene poco o ningún efecto sobre el valor de la mediana muestral. Los valores atípicos surgen por dos razones: l) son observaciones legítimas cuyos valores son inusualmente grandes o pequeños, o 2) son el resultado de un error de medición, de una deficiente técnica experimental o de un error al guardar los datos. En el primer caso, se sugiere que se considere su presencia y que se calculen los estadísticos de la muestra con y sin el valor atípico. En el segundo caso, el dato puede corregirse si fuera posible o eliminarse del conjunto de datos. En la Sección 1.6, se presenta un método gráfico para detectar valores atípicos.

EJERCICIOS 1.4

1. Considérense los siguientes conjuntos de datos:

Conjunto de datos I		
2	4	0
1	4	3
3	1	1

Conjunto de datos II		
1	1	5
3	4	6
2	1	5

Hallar, para cada uno de ellos, la media muestral y la mediana muestral.

2. En un estudio sobre parásitos, se consideró la distribución de la garrapata *Ixodes triangu-liceps* en el cuerpo de los ratones. Se obtuvieron las siguientes observaciones del número de garrapatas encontradas sobre 44 ratones.

0	2	0	0	2	2	0	0	1
1	3	0	0	1	0	0	1	0
1	4	0	0	1	4	2	0	0
1	0	0	2	2	1	1	0	6
0	5	1	3	0	1	0	1	

a) Diseñar un gráfico de barras de frecuencias para estos datos y estimar la media muestral mediante su observación.
b) Calcular la media muestral y comparar este valor con su estimación del apartado *a)*.
c) Determinar la mediana muestral.

3. *«Outliers» o datos atípicos.* Los estudios sobre pájaros suelen realizarse mediante captura, anillamiento y puesta en libertad, de manera que puedan seguirse después sus movimientos. Una variable estudiada fue la distancia de vuelo desde el punto en que se soltó un pájaro recién anillado hasta su primera posada. Los siguientes datos corresponden a dos tipos de pájaros, el petirrojo y la paloma de Carolina (la distancia está dada en pies).

Petirrojo (I)			Paloma de Carolina (II)			
128.8	57.2	48.2	40.0	381.7	358.9	1200.0*
160.0	65.2	69.2	80.0	266.8	13.9	
192.1	68.9	117.3	313.9	162.7	165.5	
163.4	24.7	36.5	175.7	76.0	317.2	
186.4	37.4	140.8	55.5	22.1	300.6	
156.2	99.7	59.3	44.7	170.0	197.7	
70.0	265.0	71.3	166.7	263.7	288.1	
10.0	78.7	105.3	83.4	369.7	102.0	

a) Calcular la media y la mediana muestral para cada conjunto de datos. ¿Son semejantes los conjuntos con respecto a alguna de las medidas?
b) Nótese que la observación con asterisco en los datos correspondientes a la paloma de Carolina es muy diferente del resto. Es lo que se llama un *outlier* o dato atípico. Para comprobar su efecto, eliminarlo de los datos y calcular la media y mediana para las 24 observaciones restantes. ¿Qué medida está menos afectada por la presencia del *dato atípico?* ¿Se percibe la razón por la que es deseable conocer tanto la media como la mediana en un conjunto de datos? ¿Aparece algún valor atípico en el conjunto de datos I?

4. Determinar la media y la mediana muestral para cada uno de los conjuntos de datos del Ejercicio 9 de la Sección 1.3.
5. Determinar la media y la mediana muestral de los datos del Ejercicio 1 de la Sección 1.2.
6. Determinar la media y la mediana muestral de los datos del Ejercicio 8 de la Sección 1.3.
7. Algunas veces, los datos del recuento se registran en diarios científicos en forma de tabla. Por ejemplo, el estudio del número de especies halladas en muestras de agua tomadas de un río cercano a un foco de contaminación, puede arrojar estos datos:

Número de especies x	Número de muestras/
0	1
1	3
2	2
4	8
5	2

En realidad, el conjunto completo de datos es el siguiente:

0 2 4 4
1 2 4 4
1 4 4 5
1 4 4 5

La media de esta muestra puede calcularse de la forma habitual. En este caso, $\bar{x} = 3.1$. Esta media puede hallarse más rápidamente multiplicando cada uno de los valores de x por su frecuencia correspondiente, sumando estos términos y, a continuación, dividiendo por la suma de las frecuencias. Para estos datos,

$$\bar{x} = \frac{0(1) + 1(3) + 2(2) + 4(8) + 5(2)}{16}$$

Utilizar esta técnica y el gráfico de barras de frecuencias hallado en el Ejercicio 2, para verificar la respuesta al apartado *b* de dicho ejercicio.

1.5. MEDIDAS DE VARIABILIDAD O DE DISPERSIÓN

Recuérdese que el comportamiento de una variable aleatoria está determinado por el azar. Así pues, los valores observados de una variable aleatoria difieren entre sí en cierta medida. En algunos casos, las diferencias son pequeñas; en otros, son pronunciadas. Puesto que esperemos que las características de nuestra muestra reflejen bien las características de la población correspondiente, medimos la variabilidad en la muestra para comprender el grado de variación que existe en la población. En esta sección consideramos cuatro medidas de variación: (el rango muestral, la varianza muestral s^2, la desviación típica muestral s y el rango intercuartíli - co *iqr*. Estos estadísticos describen la variabilidad en la muestra y se utilizan para aproximar el rango de la población, la varianza de la población σ^2, la desviación típica de la población σ y el rango intercuartílico de la población, IQR, respectivamente. Véase la Figura 1.18. Un ejemplo nos mostrará la necesidad de estas medidas, como complemento de la de localización introducidas en la última sección.

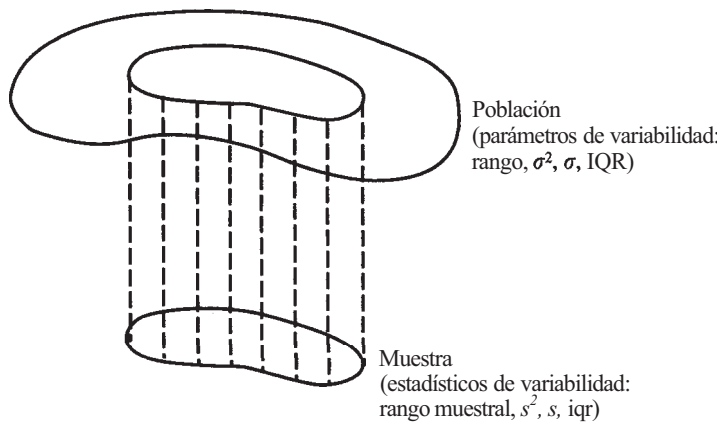


Figura 1.18. El rango muestral de los estadísticos, s^2, se *iqr se* utilizan para describir la variabilidad en la muestra y para aproximar los parámetros de variabilidad de la población.

Ejemplo 1.5.1. En una investigación sobre lesiones producidas por práctica de deportes escolares, se seleccionaron y estudiaron 25 distritos escolares dentro de una misma región. Se obtuvieron los siguientes datos sobre el número de lesiones graves causadas a deportistas masculinos mientras practicaban baloncesto y fútbol:

Baloncesto					**Fútbol**				
1	**2**	**4**	**4**	**7**	**1**	**7**	**7**	**6**	**1**
3	**3**	**2**	**4**	**5**	**2**	**6**	**1**	**7**	**2**
2	**4**	**3**	**5**	**3**	**1**	**3**	**2**	**7**	**5**
4	**4**	**3**	**6**	**5**	**6**	**1**	**7**	**4**	**1**
5	**6**	**4**	**6**	**5**	**5**	**7**	**6**	**3**	**2**

La Tabla 1.20 es la distribución de frecuencias para estos datos, y en la Figura 1.19 se muestra el histograma de frecuencias relativas para cada conjunto de datos. Es evidente que, aunque ambos histogramas se centran en 4, sus formas son bastante diferentes. Representaremos con una X el número de lesiones sufridas en el baloncesto por distrito, y con una Y el número de lesiones sufridas en el fútbol. Un cálculo rápido nos indica que $\bar{x} = \bar{y} = 4$. Obsérvese que las medianas muestrales son idénticas a las medias muestrales. Si hubiéramos utilizado sólo la localización para comparar los dos conjuntos de datos, hubiéramos deducido erróneamente que no existe diferencia entre ellos.

Una característica que no está siendo detectada por la media o la mediana muestral es la *variabilidad.* Hay alguna fluctuación en las observaciones y no es siempre la misma. Algunas están próximas a la media; otras no. Necesitamos una medida que cuantifique esta variabilidad. Queremos un estadístico con la propiedad de que, cuando los datos puntuales estén agrupados cerca de la media, su valor sea pequeño; y que, cuando los datos puntuales estén distribuidos de manera más amplia, muchos de ellos bastante alejados de la media, su valor sea grande. La Figura 1.20 ilustra esta idea. Para esta cuestión, se han propuesto varios estadísticos. Quizás la forma más lógica de tratar de medir la variabilidad respecto de la media es determinar la distancia de cada dato puntual desde la media y sumar estas distancias. Cada distancia se determina restando la media de los datos puntuales. Es decir, se determinan formando la diferencia $x\text{-}\bar{x}$ para cada observación. Un rápido ejemplo mostrará que, si bien este propósito es intuitivamente atrayente: ¡no funciona!

Tabla 1.20. Distribución de frecuencias de lesiones en deportes escolares

Lesiones en baloncesto			**Lesiones en fútbol**		
Número de casos	**Frecuencia**	**Frecuencia relativa**	**Número de casos**	**Frecuencia**	**Frecuencia relativa**
1	**1**	**0.040**	**1**	**6**	**0.240**
2	**3**	**0.120**	**2**	**4**	**0.160**
3	**5**	**0.200**	**3**	**2**	**0.080**
4	**7**	**0.280**	**4**	**1**	**0.040**
5	**5**	**0.200**	**5**	**2**	**0.080**
6	**3**	**0.120**	**6**	**4**	**0.160**
7	**1**	**0.040**	**7**	**6**	**0.240**

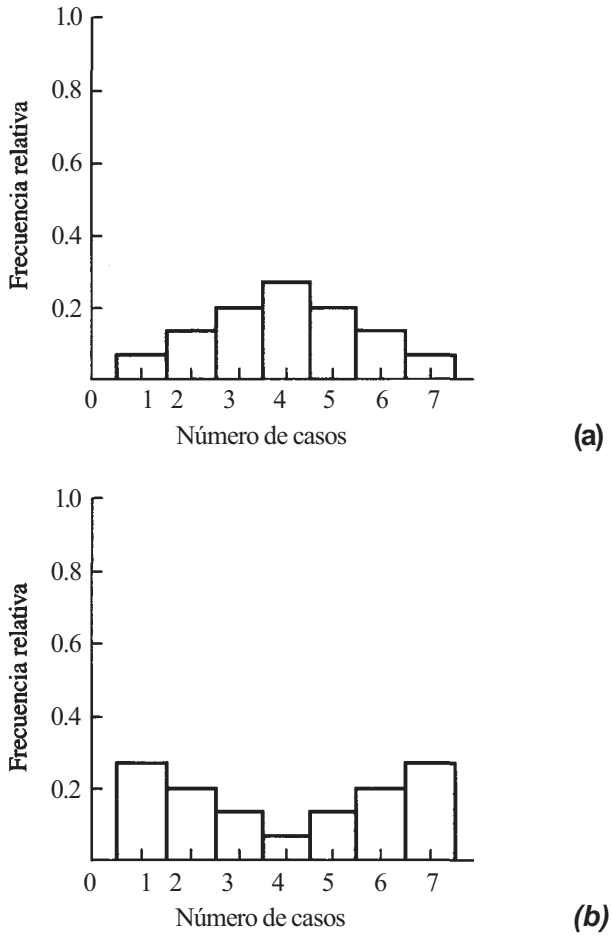


Figura 1.19. Histogramas de frecuencia relativa: (a) baloncesto y *(b)* fútbol

Ejemplo 1.5.2. Considérese el siguiente conjunto de observaciones:

$$x_1 = 2 \quad x_3 = 1 \quad x_5 = 4$$

La media muestral para este conjunto de datos es:

$$\bar{x} = \frac{\sum x}{5} = \frac{2+5+1+3+4}{5} = 3$$

La suma de las diferencias entre las observaciones y la media muestral es:

$$(2 - 3) + (5 - 3) + (1 - 3) + (3 - 3) + (4 - 3) = (-1) + 2 + (-2) + 0 + 1 = 0$$

El método propuesto para medir la variabilidad parece indicar que no hay fluctuación en los datos. El problema es evidente. Estamos permitiendo a las diferencias negativas asociadas con los datos puntuales que están por debajo de $\bar{x}$ cancelar las positivas que aparecen cuando un dato puntual está por encima de $\bar{x}$.

Este problema debe ser corregido para obtener una medida satisfactoria de variabilidad en torno a la media.

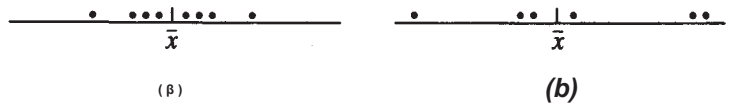


Figura 1.20. (a) Los datos puntuales se agrupan cerca de la media muestral. La variabilidad en torno a la media es pequeña. *(b)* Los datos puntuales están más dispersos. Nuestro estadístico medidor de la variabilidad debe ser mayor en el caso *(b)* que en el caso (a).

Varianza muestral

Hay dos formas de evitar el problema de que las diferencias negativas anulen a las positivas. Podríamos simplemente ignorar los signos negativos y trabajar con los valores absolutos de las diferencias implicadas o podríamos elevar al cuadrado las diferencias. Este último procedimiento es la técnica utilizada habitualmente. Así, nuestra medida de variabilidad hará uso de la suma de los cuadrados de las diferencias entre los datos puntuales y la media muestral $\sum(x - \bar{x})^2$. ¿Es suficiente esta suma para conseguir el objetivo perseguido? Para responder a esta pregunta, consideremos dos ejemplos extraídos de la misma población. Supongamos que una muestra es de tamaño 5 y la otra de tamaño 5000. ¿Cuál de ellas tendrá un valor mayor para el estadístico $\sum(x - \bar{x})^2$? La segunda, naturalmente, pero debido fundamentalmente a la diferencia de tamaño. Para asegurarnos de que las diferencias en el tamaño de las muestras no influyen en nuestra medida de variabilidad, no trabajaremos con la suma de las diferencias elevadas al cuadrado, sino con el promedio de las diferencias al cuadrado. De ahí que la medida de variabilidad sobre la media más lógica sea la media aritmética de estas diferencias al cuadrado,

$$\frac{\sum(x - \bar{x})^2}{n}$$

Esta medida es aceptable y es la preferida por muchos. Sin embargo, se puede demostrar que este estimador, en promedio, tiende a subestimar la varianza de la población σ^2. Es decir, si dibujamos en una línea los valores obtenidos con esta fórmula para varias muestras de un mismo tamaño, obtenemos un diseño similar a la Figura 1.21. Esta situación puede remediarse dividiendo la suma de las diferencias al cuadrado por n - 1 en lugar de n. El estadístico así obtenido se llama *varianza muestral.* El patrón de los valores obtenidos al estimar σ^2 por

$$\frac{\sum(x - \bar{x})^2}{n - 1}$$

se muestra en la Figura 1.22. La definición se formaliza a continuación.

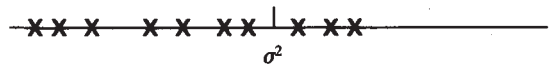


Figura 1.21. Cuando usamos $\sum(x - \bar{x})^2/n$ para estimar σ^2, la mayor parte de las estimaciones cae por debajo de σ^2. Es decir, el estadístico tiende a infraestimar σ^2

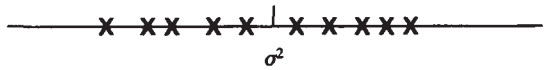


Figura 1.22. Cuando usamos $\sum(x - \bar{x})^2/(n - 1)$ para estimar σ^2, los valores varían, por término medio, alrededor de σ^2

Definición 1.5.1. Varianza muestral. Sea $x_1, x_2, \ldots, x_n$ un conjunto de n observaciones sobre una variable X, con media muestral $\bar{x}$. La *varianza muestral* se denota por s^2 y viene dada por

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Tomaremos las varianzas muestrales hasta con dos cifras decimales más que los datos.

Ejemplo 1.5.3. El conjunto de datos

2 1 4
5 3

tiene una media muestral igual a 3. Su varianza muestral viene dada por:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$= \frac{(2-3)^2 + (5-3)^2 + (1-3)^2 + (3-3)^2 + (4-3)^2}{4}$$

$$= \frac{(-1)^2 + 2^2 + (-2)^2 + 0^2 + 1^2}{4}$$

$$= \frac{1+4+4+0+1}{4} = \frac{10}{4} = 2.50$$

La fórmula para la varianza muestral dada en la Definición 1.5.1 implica una buena cantidad de cálculos, especialmente si el conjunto de datos es grande. Si se tiene una calculadora con funciones estadísticas incorporadas, la fórmula para s^2 está programada internamente. Para determinar s^2, en lugar de la definición, es preferible utilizar esta característica de la calculadora.

Recuérdese la interpretación práctica de s^2. La medida está definida de tal forma que r o puede ser negativa. Además, si la mayor parte de las observaciones caen cerca de la media, la varianza será pequeña y si los datos muestran una variabilidad importante —es decir, si los valores se alejan a menudo de la media—, la varianza será grande.

Ahora podemos investigar más a fondo las diferencias que existen entre los dos conjuntos de datos del Ejemplo 1.5.1.

Ejemplo 1.5.4. Considérense los histogramas de la Figura 1.19. La mayor parte de las observaciones obtenidas sobre lesiones de baloncesto están cerca del valor medio 4, como revela la elevación central del histograma *a.* Sin embargo, muy pocas de las observaciones de lesiones de fútbol caen cerca del 4, como demuestra la depresión central del diagrama *b.* Intuitivamente, cabe esperar que la varianza muestral para los datos sobre lesiones en baloncesto sea más pequeña que la de los datos sobre lesiones de fútbol. Para verificarlo, calculemos s^2 para cada conjunto de datos. El cálculo está hecho en una calculadora TI83. Puede comprobar los valores dados a continuación.

Baloncesto	Fútbol
$\bar{x} = 4$	$\bar{y} = 4$
$\tilde{x} = 4$	$\tilde{y} = 4$
$s_X^2 = 2.17$	$s_Y^2 = 6.00$

(Obsérvese que hemos redondeado $s^2 = 2.166666$ a 2.17, de forma que s^2 se anota hasta con dos cifras decimales más que los datos). Como se esperaba, $s^2_x < s^2_Y$.

Desviación típica muestral

Una segunda medida de variabilidad es la desviación típica muestral. La definimos a continuación.

> **Definición 1.5.2. Desviación típica muestral.** Sea $x_1, x_2, \ldots x_n$ un conjunto de n observaciones sobre una variable X con varianza muestral s^2. La *desviación típica muestral* se simboliza por s y se define por:
>
> $$s = \sqrt{s^2}$$

Las desviaciones típicas se registrarán con un decimal más que los datos.

Obsérvese que la desviación típica muestral es simplemente la raíz cuadrada no negativa de la varianza muestral. Puesto que estas dos medidas de variabilidad están tan íntimamente relacionadas, la pregunta natural que surge es: ¿por qué plantearse las dos? Hay una razón muy práctica para querer medir la variabilidad utilizando la desviación típica del conjunto de datos. Considérense los datos sobre el número de lesiones de fútbol y baloncesto recogidos en 25 distritos escolares. La unidad asociada con cada dato puntual y con cada media muestral es una «lesión». Cuando se calcula la varianza muestral, las diferencias entre los valores observados y la media muestral están *elevadas al cuadrado.* La unidad asociada con la varianza muestral es, por tanto, una «lesión al cuadrado». Esto no tiene ningún sentido. En todo caso, puesto que la desviación típica muestral es la raíz cuadrada de la varianza muestral, la unidad asociada a s es nuevamente una «lesión». Sucede a menudo que la unidad original pierde sentido cuando se eleva al cuadrado. Por esta razón, generalmente no se asocia unidad a una varianza. Por el contrario, la unidad asociada con una desviación típica será siempre la misma que la asociada con los datos originales y, por tanto, físicamente tendrá sentido.

> **Ejemplo 1.5.5.** La varianza muestral para las lesiones relativas al baloncesto es 2.17 (véase el Ejemplo 1.5.4). La desviación típica muestral es $s = \sqrt{s^2} = \sqrt{2.166666} = 1.5$. Obsérvese que, cuando hallamos la desviación típica, no redondeamos la varianza antes de hacer la raíz cuadrada.

Rango muestral

La siguiente medida de variabilidad a considerar es el rango o recorrido muestral. Esta medida es la más fácil de calcular, y se define a continuación.

> **Definición 1.5.3. Rango muestral.** Sea $x_1, x_2, \ldots, x_n$ un conjunto de n observaciones correspondientes a una variable X. *El rango muestral* es la diferencia, en este orden, entre el mayor y el menor de los valores observados.

Esta medida de variabilidad se utilizó en la Sección 1.3 para construir histogramas. Un rango grande implica que los datos están distribuidos en un intervalo amplio; un rango pequeño garantiza que los datos están concentrados en un pequeño segmento de la recta real.

Rango intercuartílico

La varianza, la desviación típica y el rango son las medidas de variabilidad que encontraremos más a menudo. Sin embargo, todas se ven seriamente afectadas por los datos atípicos.

Así pues, un solo valor atípico puede inflar su valor y dar una impresión un tanto confusa de la variación que existe en el grueso de los datos. Es útil tener una medida de variabilidad resistente a los datos atípicos. Nuestra cuarta medida de variación, el rango intercuartílico, es una de ellas. El rango muestral intercuartílico, *iqr,* representa la longitud del intervalo que contiene, aproximadamente, el 50 % de datos situados en el medio. Si el *iqr* es pequeño, gran parte de los datos se encuentran cerca del centro de la distribución; si es grande, los datos tienden a distribuirse ampliamente. Los pasos para calcular el *iqr* son los siguientes:

Determinación del rango intercuartílico muestral

1. Determinar la posición de la mediana, $(n + 1)/2$, donde n es el tamaño de la muestra. Si el tamaño muestral es un número impar, entonces este lugar será un entero. En otro caso, será un número a mitad de camino entre dos enteros. Por ejemplo, si $n = 17$, a localización de la mediana es $(17 + 1)/2 = 9$, un entero. Si $n = 18$, su localización es $(18 + 1)/2 = 9.5$, el número a mitad de camino entre los enteros 9 y 10.
2. Si fuese necesario, se truncará la ubicación de la mediana ignorando el 0.5. Por ejemplo, si está localizada en el 9.5, se trunca, es decir, se toma el valor 9. Si está localizada ya en un número entero, no es necesario truncar.
3. Determinar la posición del cuartil q mediante:

$$1 = \frac{\text{posición truncada de la mediana} + 1}{2}$$

4. Determinar q_1 contando desde el dato puntual más pequeño hasta la posición q. Si q es un entero, q_1 es el dato puntual en la posición q. Si q no es un entero, q_1 es el promedio de los datos puntuales en las posiciones q - 0.5 y q + 0.5. Aproximadamente el 25 % de los datos caerán en q_1 o por debajo de q_1.
5. Determinar q_3 contando hacia abajo desde el dato puntual más grande hasta la posición q, como en el punto 4. Aproximadamente el 75 % de los datos caerán en q_3 o por debajo de q_3.
6. Definir *iqr* mediante $iqr = q_3 - q_1$.

Ejemplo 1.5.6. Se ha realizado un estudio sobre el tipo de sedimentos hallados en dos lugares de perforación diferentes, a gran profundidad. La variable aleatoria de interés es el porcentaje en volumen de cemento hallado en las muestras del sondeo. Por cemento queremos decir carbonates disueltos y precipitados de nuevo. Se obtuvieron los siguientes datos: (Basado en la información hallada en Andreas Wetzel, «Influence of Heat Flow on Ooze/Chalk Cementation: Quantification from Consolidation Parameters in DSDP Site 504 and 505 Seciments», *Journal of Sedimentary Petrology,* julio de 1989, págs. 539-547.)

Lugar I, % de cemento			
10	21	12	12
20	13	24	36
31	18	17	16
37	16	32	13
14	49	25	19
13	32	27	

Lugar II, % de cemento		
1	10	14
9	21	19
15	17	13
25	22	20
24	12	23
15	20	18

```
1 | 0 4 3 3 2 2 3
1 | 8 6 7 6 9
2 | 0 1 4
2 | 5 7
3 | 1 2 2
3 | 7 6
4 |
4 | 9
```

Figura 1.23. Diagrama de tallo y hojas doble para el porcentaje, en volumen de cemento de las muestras del sondeo, tomadas en el lugar de perforación I.

En la Figura 1.23 se muestra el diagrama de tallo y hojas doble para los datos del lugar I. La muestra tiene un tamaño de $n = 23$. La localización de la mediana es $(n + 1)/2 = 12$. La posición del cuartil $q = (12+1)/2 = 6.5$. Para hallar q_1, utilizamos el diagrama de tallo y hojas para localizar los datos sexto y séptimo, contando desde los números más pequeños hacia arriba. Estos valores son 13 y 14, respectivamente. Por lo tanto $q_1 = (13 + 14)/2 = 13.5$. Para hallar q_3 determinamos los datos sexto y séptimo contando desde los números más altos hacia abajo. Estos puntos son 31 y 27, respectivamente, resultando $q_3 = (31 + 27)/2 = 29$. El rango muestral intercuartílico es $q_3 - q_1 = 29 - 13.5 = 15.5$.

Conjuntos de datos múltiples (opcional)

Las medidas de tendencia central, variabilidad y rango proporcionan una información valiosa sobre un conjunto de datos. A menudo, el problema no viene dado por un único conjunto de datos, sino por la forma en que un conjunto particular de medidas se comporta como función del tiempo, de la dosis de fármaco o de alguna otra variable. Para cada instante de tiempo, dosis de fármaco y demás, se obtiene un conjunto de datos diferente. Esto puede implicar un gran número de datos puntuales; por ello, es deseable disponerlos de forma conveniente, de modo que proporcionen la máxima información. Con este objeto se describe un método en el Ejemplo 1.5.7.

Ejemplo 1.5.7. Se midió la concentración de lactato en sangre arterial, en una muestra de seis varones jóvenes antes y varias veces durante un ejercicio controlado y en el periodo de recuperación siguiente. Se obtuvieron los datos de la Tabla 1.21. Pueden calcularse la media y la desviación típica del conjunto de datos en cada instante de tiempo; datos que se muestran en la Tabla 1.22. Después, podemos construir un gráfico en el que la concentración media de lactato, en milimoles por litro, se represente en función del tiempo. La desviación **típica** en cada instante de tiempo del conjunto de datos se representa por segmentos que se extienden por encima y por debajo de los puntos correspondientes al valor medio, tal como se indica en la Figura 1.24. Una ojeada a este gráfico da una idea de las variaciones sanguíneas de lactato durante el ejercicio y la recuperación, y permite una rápida valoración de la variabilidad en instantes de tiempo diferentes.

Tabla 1.21. Concentraciones arteriales de lactato (milimoles/litro)

Reposo minutos	Ejercicio, minutos				Recuperación, minutos			
	5	10	20	30	5	20	35	65
0.93	7.60	8.25	6.70	6.49	4.35	2.05	1.35	0.85
0.55	3.95	4.31	8.85	4.38	6.22	2.98	2.02	0.58
0.87	6.54	6.52	4.56	8.75	2.45	1.17	0.76	0.95
0.62	4.27	5.15	7.29	6.72	3.57	2.71	1.21	1.10
0.72	5.85	7.31	5.88	4.88	3.81	1.84	1.58	0.62
0.79	5.41	6.30	6.96	7.70	5.62	2.36	1.29	0.79

Tabla 1.22. Concentración de lactato

Reposo minutos	Ejercicio, minutos				Recuperación, minutos			
	5	10	20	30	5	20	35	65
$\bar{x}$ 0.747	5.603	6.307	6.707	6.487	4.337	2.185	1.368	0.815
s 0.146	1.377	1.425	1.435	1.653	1.387	0.649	0.417	0.197

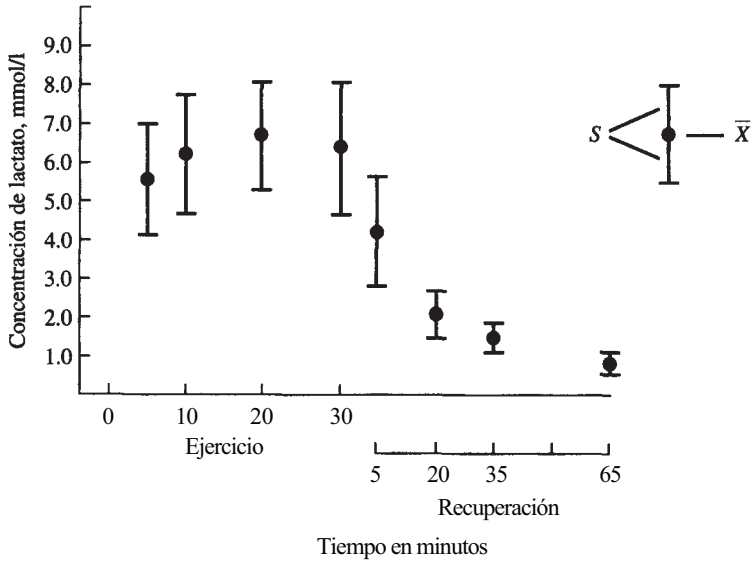


Figura 1.24. Concentración de lactato con respecto al tiempo.

Hay un detalle que conviene tener en cuenta. Gráficos similares a los descritos en el Ejemplo 1.5.7 son muy frecuentes en los libros sobre estadística. También se utilizan otros gráficos de aspecto similar, en los que se emplea una medida de variabilidad conocida como el *error estándar de la media* (SE), que determina la longitud de los segmentos colocados por encima y por debajo de la media. El concepto de error estándar se introduce en el Capítulo 6. Debe señalarse con claridad qué tipo de gráfico y qué medida de variabilidad se está usando.

Hay otras medidas estadísticas para describir los conjuntos de datos. Aquí hemos presentado las más habituales. En los ejercicios se mencionan otras, con fines informativos. La media muestral, la varianza y la desviación típica se emplean profusamente a lo largo de todo el texto; la mediana muestral desempeña un papel importante en la exposición de los métodos no paramétricos del Capítulo 13.

EJERCICIOS 1.5

1. Determinar el rango muestral, s^2, y s para los datos del Ejercicio 1 de la Sección 1.4.
2. Determinar el rango muestral, s^2, y s para los datos del Ejercicio 2 de la Sección 1.4.
3. *a)* Determinar s^2 y s para cada uno de los conjuntos de datos del Ejercicio 3 de la Sección 1.4.
   b) Extraer el dato atípico de los datos sobre las palomas de Carolina y volver a calcular s^2 y s. ¿Son estas medidas de variabilidad resistentes al dato atípico?

c) Determinar el *iqr* para cada uno de los conjuntos de datos del Ejercicio 3 de la Sección 1.4.

d) Extraer el dato atípico de los datos sobre las palomas de Carolina y volver a calcular el *iqr*. ¿Es resistente al dato atípico esta medida de variabilidad?

e) Hallar el rango para cada conjunto de datos del Ejercicio 3 de la Sección 1.4.

f) Extraer el dato atípico de los datos sobre las palomas de Carolina y volver a calcular el rango. ¿Resiste al dato atípico esta medida de variabilidad?

4. Basándose en los histogramas de los datos del Ejercicio 6 de la Sección 1.3, ¿qué conjunto de datos piensa usted que tiene una varianza mayor? Calcular s^2 de cada grupo para verificar su respuesta. ¿Qué unidad física de medida está asociada con s?
5. Determinar s^2, *s,* el rango muestral y el *iqr* para los datos del Ejercicio 7 de la Sección 1.3.
6. Determinar s^2, *s,* el rango muestral y el *iqr* para los datos del Ejercicio 8 de la Sección 1.3. ¿Qué unidad física de medida está asociada con s?
7. a*)* Construir un diagrama de tallo y hojas doble para los datos del lugar II del Ejemplo 1.5.6. Comparar la forma de esta distribución con la del lugar I.

   b) Determinar la media y la mediana muestral de cada muestra.

   c) Determinar s^2 y *s* en cada muestra.
8. Basándose en los datos del Ejercicio 1 de la Sección 1.2.

   a*)* Determinar $\bar{x}$, $\tilde{x}$, s^2, *s,* el rango y el *iqr* para estos datos.

   b) Suponer que se obtiene una observación adicional de 0.72. ¿Se verá afectado alguno de los estadísticos del apartado *a* por la adición de este punto? Expliquese.

   c) Añadir ahora la observación 2.8 al conjunto de datos original. ¿Tendrá este punto un gran impacto en el valor de alguno de los estadísticos del apartado *al* Si así fuera, ¿cuáles se verán afectados? Compruebe su respuesta calculando $\bar{x}$, $\tilde{x}$, s^2, *s,* el rango y el *iqr* para este conjunto de datos ampliado.
9. Considerar los datos del Ejercicio 9 de la Sección 1.3.

   a) Dar un ejemplo de una lectura ICS adicional que cambiase notoriamente el rango de los datos correspondientes a lugares no incendiados. ¿Cambiará el valor de la mediana en gran medida?

   b) Dar un ejemplo de una lectura de ICS adicional que tuviese un pequeño efecto en la media muestral, desviación típica y varianza para los datos correspondientes a lugares incendiados.
10. *Coeficiente de variación.* El coeficiente de variación es una medida para comparar la variabilidad en un conjunto de datos con la de otro, en situaciones en las que una comparación directa de desviaciones típicas no es conveniente o suficientemente realista. Por ejemplo, en un estudio del consumo de leche en Estados Unidos, se obtuvo que el número medio de galones de leche consumida por unidad familiar por semana fue 8, con una desviación típica muestral de 3 galones. Un estudio semejante en Canadá dio un consumo medio de 12 litros con una desviación de 4 litros. No tiene sentido comparar estas desviaciones típicas directamente porque están dadas en unidades diferentes. Una forma rápida de comparar la variabilidad es con el *coeficiente de variación* (CV) dado por:

$$\text{CV} = \frac{s}{\bar{x}}\,(100)$$

Los coeficientes de variación de las dos muestras son (3/8) • 100 = 37.5 y (4/12) • 100 = 33.3, respectivamente. Los datos de Estados Unidos presentan más variabilidad que los de Canadá.

a) Se realiza un experimento para investigar el efecto de una nueva comida para perros, sobre la ganancia de peso de los cachorros durante las primeras 8 semanas de vida. En un grupo de cachorros de Gran Danés, se obtiene una ganancia media de 30 libras con una desviación típica de 10 libras; en un grupo de cachorros de Chihuahua la ganancia media de peso es de 3 libras, con una desviación típica de 1.5 libras. Calcúlese el coeficiente de variación para cada grupo. ¿Qué grupo posee la variabilidad más grande? ¿Por qué es equívoca en este caso una comparación directa de desviaciones típicas?

b) En un estudio de los pesos de niñas de dos años de Gran Bretaña, se obtuvo una media muestral de 12.74 kilogramos con una desviación típica muestral de 1.60 kilogramos. Un estudio semejante en Estados Unidos dio una media muestral de 29.2 libras con una desviación típica muestral de 2 libras. Encontrar el coeficiente de variación para cada grupo. ¿Qué grupo posee mayor variabilidad?

11. Comprobar los valores de la Tabla 1.22.
12. En un estudio sobre dos anestésicos, utilizando ratas conscientes moviéndose libremente, la respuesta medida fue el porcentaje de cambio en la presión de CO_2 en la sangre arterial, tras la administración de dosis idénticas del medicamento. Se obtuvieron los datos siguientes: (Basado en la información publicada en Linas V. Kudzma et al., «A Novel Class of Analgesic and Anesthetic Agents», *Journal of Medicinal Chemistry,* diciembre de 1989, págs. 2534-2542.)

Porcentaje de cambio, mm Hg				
Compuesto I		**Compuesto II**		
27.2	31.7	55.1	65.8	63.6
30.1	32.0	56.3	58.3	64.0
30.5	28.6	60.0	57.1	65.3
28.4	29.2	63.5	55.4	62.8
30.7	33.0	64.9	56.5	59.5
31.3	31.7	62.7	55.1	
30.5	32.6	60.5	57.0	
30.1	28.2	59.2	59.3	
29.6	29.1	63.7	60.7	
30.2	30.7	64.1	62.1	

a) Construir un diagrama de tallo y hojas para cada conjunto de datos. ¿Qué conjunto de datos parece más disperso?

b) Calcular la media y la mediana muestrales de cada conjunto de datos.

c) Calcular la varianza muestral y la desviación típica muestral de cada conjunto de datos. Compruebe su respuesta al apartado *a.* ¿Qué unidad de medida está relacionada con la desviación típica muestral?

d) Calcular el rango muestral y el rango intercuartílico muestral de cada conjunto de datos.

e) Basándose en las características observadas en estas muestras, ¿se sorprendería si le dijeran que no existe diferencia en la forma en que reaccionan las ratas a estos compuestos? Expliquese.

f) Construir un diagrama similar al de la Figura 1.24 para comparar los dos conjuntos de datos.

13. Puede utilizarse la técnica mostrada en el Ejercicio 7 de la Sección 1.4 para calcular s^2, a partir de los datos anotados en la distribución de frecuencias. Para hacerlo emplearemos un atajo:

$$s^2 = \frac{n\sum x^2 - \left(\sum x\right)^2}{n(n-1)}$$

En primer lugar, para obtener $\sum x$, multiplicamos cada valor de x por su frecuencia correspondiente y sumamos estos valores; después, multiplicamos el cuadrado de cada x por la frecuencia correspondiente y lo sumamos para obtener $\sum x^2$. Finalmente, estos valores se sustituyen en la fórmula del atajo para obtener s^2. En el caso de los datos del Ejercicio 7 de la Sección 1.4:

$$\sum x = 0(1) + 1(3) + 2(2) + 4(8) + 5(2) = 49$$
$$\sum x^2 = 0^2(1) + 1^2(3) + 2^2(2) + 4^2(8) + 5^2(2) = 189$$

La varianza muestral para estos datos es:

$$s^2 = \frac{16(189) - (49)^2}{16(15)} = 2.60$$

a) Utilizar esta técnica para comprobar los valores del Ejemplo 1.5.4, y la Tabla 1.20.
b) Utilizar esta técnica para determinar la varianza muestral de los datos del Ejercicio 2 de la Sección 1.4.

1.6. DIAGRAMA DE CAJAS (OPCIONAL)

El diagrama de cajas es una representación gráfica de un conjunto de datos que facilita la percepción visual de su localización, extensión, y del grado y la dirección del sesgo. También permite identificar los datos atípicos. Es especialmente útil cuando se desean comparar dos o más conjuntos de datos. El método de diseño que aquí se muestra es el de Lambert H. Koopmans [7].

Construcción de un diagrama de cajas

1. Se construye una escala de referencia horizontal o vertical.
2. Determinar la mediana muestral, q_1, q_3, e iqr tal como se ha explicado en la Sección 1.5.
3. Determinar dos puntos f_1, yf_3, denominados «separadores interiores», mediante:

$$f_1 = q_1 - 1.5(iqr)$$
$$f_3 = q_3 + 1.5(iqr)$$

Los puntos por debajo de f_1 o por encima de f_3 se considerarán atípicos.

4. Determinar dos puntos a_1 y a_3 denominados «valores adyacentes». El punto a_1 es el dato más cercano a f_1 sin que su valor esté por debajo de f_3. El punto a_3 es el dato más cercano av f_3, sin que su valor esté por encima de f_3.

5. Determinar dos puntos F_1 y F_3 denominados «separadores exteriores», mediante:

$$F_1 = q_1 - 2(1.5)(iqr)$$
$$F_3 = q_3 + 2(1.5)(iqr)$$

6. Situar los puntos hallados hasta ahora sobre la escala horizontal o vertical. Sus posiciones relativas se muestran en la Figura 1.25a.
7. Construir una caja con los extremos en q_1 y q_3 con una línea interior dibujada en la mediana, tal como se muestra en la Figura 1.256.
8. Indicar los valores adyacentes con el símbolo *x,* y conectarlos a la caja con líneas punteadas. Estas líneas punteadas se llaman «patillas» o «bigotes». Situar los datos; puntuales que estén entre separadores interiores y exteriores y representarlos mediante círculos abiertos. Se considera que estos puntos son datos atípicos moderados Indicar los datos puntuales que caen fuera de separadores exteriores mediante asteriscos. Se considera que estos puntos son datos atípicos extremos (véase Fig. 1.25c).

La localización de la línea central de la caja es una indicación de la forma de la distribución. Si la línea está descentrada, sabremos que la distribución está sesgada en la dirección del extremo más largo de la caja.

Antes de ilustrar esta técnica, debe aclararse la idea de separadores. Puede demostrarse que, al muestrear a partir de una distribución normal, una distribución simétrica en forma de campana que se estudiará detalladamente en el Capítulo 5, sólo aproximadamente 7 valores de cada 1000 caerán fuera de los separadores interiores. Puesto que estos valores son muy inusuales, se consideran atípicos. Los datos atípicos deben tratarse con cuidado pues, como se habrá apreciado, su presencia puede tener un impacto crucial sobre *x, s^2, s* y el rango, es decir, sobre las medidas usuales de posición y variación. Cuando se encuentre un dato atípico. debería considerarse su origen. ¿Es legítimo un dato cuyo valor, inusualmente, es grande c pequeño? ¿Es un valor mal registrado? ¿Es el resultado de algún error o accidente en *te* experimentación? En los dos últimos casos puede borrarse el punto del conjunto de datos *y* completarse el análisis con los datos restantes. En el primer caso, sugerimos que se dé conocer la presencia del dato atípico y que los estadísticos se citen con y sin éste. De esta forma, el investigador, que es el experto en la materia, puede tomar la decisión de incluir o no el dato atípico en futuros análisis.

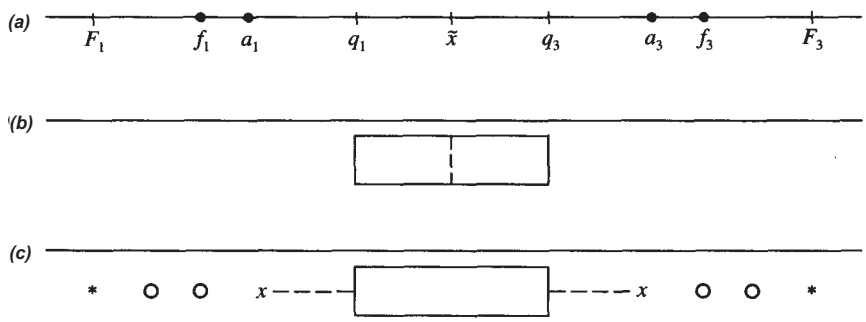


Figura 1.25. (a) Posiciones relativas de la mediana ($\tilde{x}$), cuartiles (q_1 y q_3) valores adyacentes (a_1 y a_3), separadores interiores (f_1 y f_3) y separadores exteriores (F_1 y F_3). (b) Se dibuja una caja que termina en (q_1 y q_3) y la línea interior en $\tilde{x}$. (c) Los valores adyacentes se indican mediante una *x,* los datos atípicos moderados se indican mediante círculos abiertos; los datos atípicos extremos se indican con asteriscos.

Ejemplo 1.6.1. En un estudio sobre la amnesia postraumática tras una lesión craneal, la variable estudiada fue el tiempo de hospitalización en días. En la Figura 1.26 se muestra el diagrama de tallo y hojas para los datos. (Basado en la información publicada en Jerry Mysia et al., «Prospective Assessment of Posttraumatic Amnesia: A Comparison of GOAT and the OGMS», *Journal of Head Trauma Rehabilitation,* marzo de 1990, págs. 65-77). Para estos datos, la posición de la mediana es (n + l)/2 = 11 y la mediana es 40 días. La posición cuartílica es q = (localización truncada de la mediana + l)/2 = 6. Los puntos $(q_1 \text{ y } q_3)$ son 32 y 47, respectivamente. El rango intercuartílico es $irq = q_3 - q_1 = 15$. Los separadores interiores son:

$$f_1 = q_1 - 1.5(iqr) \qquad f_3 = q_3 + 1.5(iqr)$$
$$= 32 - 22.5 \qquad = 47 + 22.5$$
$$= 9.5 \qquad = 69.5$$

Los valores adyacentes son $a_1 = 12$ y $a_3 = 61$. Los separadores exteriores son:

$$F_1 = q_1 - 2(1.5)(iqr) \qquad F_3 = q_3 + 2(1.5)(iqr)$$
$$= 32 - 45 \qquad = 47 + 45$$
$$= -13 \qquad = 92$$

El conjunto de datos contiene dos puntos, 8 y 89, que se califican como datos atípicos moderados. El punto 108 se califica como dato atípico extremo. Obsérvese que, dado que F_1 es negativo, es físicamente imposible observar un dato atípico extremo en el extremo inferior de la escala. En la Figura 1.27 se muestra el diagrama de cajas. Obsérvese que la línea central de la caja está cerca de su centro, indicando una distribución casi simétrica. Con respecto a los datos atípicos, ¿son observaciones reales que deben tenerse en cuenta, o son el resultado de errores en la recogida de datos? En este caso, sería fácil comprobar los registros de los pacientes para determinar la respuesta, y ello debería hacerse antes de proceder a cualquier otro análisis de los datos.

Un punto más a resaltar: el test de detección de datos atípicos está basado en el supuesto de que los datos provienen de una distribución normal. Si la distribución es asimétrica, es probable que los valores que definen la cola larga de la asimetría se identificarán como datos

0	8
1	2
2	7
3	0 2 5 6
4	0 0 0 0 1 2 5 7
5	0 2
6	1
7	
8	9
9	
10	8

Figura 1.26. Diagrama de tallo y hojas para los datos del Ejemplo 1.6.1. Los datos representan el tiempo de hospitalización de los pacientes con amnesia postraumática en días *(n* = 21).

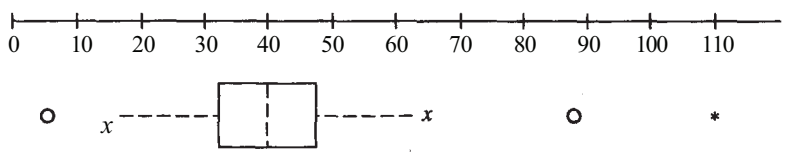


Figura 1.27. Diagrama de cajas para los datos del Ejemplo 1.6.1.

atípicos. ¿Lo son realmente? Posiblemente no. El test para datos atípicos debería usarse, pues, en conjunción con el diagrama de tallos y hojas o el histograma.

EJERCICIOS 1.6

1. En el estudio del Ejemplo 1.6.1 también se consideró la variable *X,* tiempo, en días, que estuvieron en coma los pacientes que padecían una lesión craneal. Se recogieron los datos siguientes:

2	8	9	14	16
6	10	8	7	
13	12	11	11	
11	13	15	10	
11	15	12	20	

   a) Construir un diagrama de tallo y hojas para estos datos. ¿Parecen estar los datos simétricamente distribuidos?
   b) Construir un diagrama de cajas para los datos ¿Da la misma impresión de simetría el diagrama de cajas, que la que se percibía en el diagrama de tallo y hojas?
   c) ¿Existen datos puntuales que puedan calificarse como datos atípicos?
2. Construir diagramas de cajas para cada uno de los conjuntos de datos del Ejemplo 1.5.6. Utilizar estas representaciones gráficas para comparar estos conjuntos de datos respecto a su localización y variabilidad. ¿Cuál es más simétrico? ¿Algún conjunto de datos contiene datos atípicos?
3. Se obtuvieron estos datos de la densidad media de los planetas que forman nuestro sistema solar y de la luna: (Basado en Nigel Henbert, «Rocky Dwarfs and Gassy Giants», *New Scientist,* 10 de febrero de 1990, págs. 1-4.)

Mercurio	5.42	Marte	3.94
Tierra	5.52	Júpiter	1.32
Saturno	0.69	Urano	1.26
Venus	5.25	Neptuno	1.64
Luna	3.34	Plutón	2.10

   ¿Alguno de estos valores puede calificarse como dato atípico?
4. Se midieron los niveles de ozono alrededor de Los Ángeles y ascendieron a 220 partes por billón (ppb). Las concentraciones de esta magnitud pueden ocasionar quemaduras en los ojos y son peligrosas tanto para las plantas como para la vida animal. También se obtuvieron datos del nivel de ozono en una zona boscosa cerca de Seattle, Washington, que fueron los siguientes: (Basado en la información publicada en «Twigs», *American Forests,* abril de 1990, pag. 71.)

160	176	160	180	167	164
165	163	162	168	173	179
170	196	185	163	162	163
172	162	167	161	169	178
161					

a) Construir un diagrama de tallo y hojas doble para estos datos. ¿Tiene aspecto de estar sesgados? Si es así, ¿en qué dirección?
b) Construir un diagrama de cajas para estos datos e identificar el dato atípico, en caso de que exista.
c) Supongamos que el dato atípico es una lectura legítima. En este caso, ¿qué medida de localización se ve menos afectada por el dato atípico? ¿Qué medida de variabilidad se ve menos afectada por el dato atípico?

5. Se están estudiando dos medicamentos, amantadina (A) y rimantadina (R), para combatir el virus de la gripe. Se han administrado por vía oral dosis únicas de 100 mg a adultos sanos. La variable estudiada es $T_{máx}$, tiempo requerido en minutos para alcanzar la concentración máxima de plasma. Se obtuvieron los datos siguientes: (Basado en la información publicada en Gordon Douglas Jr., «Drug Therapy», *New England Journal of Medicine,* febrero de 1990, págs. 443-449.)

$T_{máx}$ (A)			$T_{máx}$ (R)		
105	123	12.4	221	227	280
126	108	134	261	264	238
120	112	130	250	236	240
119	132	130	230	246	283
133	136	142	253	273	516
145	156	170	256	271	
200					

a) Construir un diagrama de cajas para cada conjunto de datos e identificar los datos atípicos.
b) Calcular x y s^2 para los datos del conjunto A.
c) Supongamos que el dato atípico del conjunto A es el resultado de un punto decimal mal colocado. Corregir el error borrando el decimal y observar qué cambios produce esto en el diagrama de cajas. Volver a calcular x y s^2, utilizando los datos correctos y comparar los resultados con los del apartado *b.*
d) ¿Hay algún dato atípico en el conjunto R? Si es así, ¿existe alguna razón legítima obvia para borrarlo del conjunto de datos?

6. Considerar los datos del Ejercicio 3 de la Sección 1.4.
a) Construir un histograma para cada conjunto de datos. Sea cuidadoso en el tratamiento del dato atípico en el conjunto de datos de las palomas de Carolina, según lo indicado en la Sección 1.3. Comentar la forma sugerida para cada distribución.
b) Construir un diagrama de cajas para cada conjunto de datos. ¿Parece el test para los datos atípicos apropiado en cada caso? Explíquelo.

7. La hidroponía es la ciencia que trata del crecimiento de las plantas solamente con una disolución de nutrientes. En un estudio realizado con *Echinacea* para comparar este método de crecimiento con el tradicional de cultivo en suelo, la variable medida fue la altura de cada planta al final. Los diagramas de cajas mostrados en la Figura 1.28 resumen los datos encontrados. En los diagramas, el grupo 1 es el grupo de hidroponía.
a) Calcular de forma aproximada la mediana para cada grupo.
b) ¿Qué grupo es más simétrico?
c) ¿Qué grupo tiene el rango intercuartílico interior más grande?
d) ¿Qué grupo tiene un dato que es posiblemente un dato atípico extremo? ¿Se puede decir con certeza qué punto es un dato atípico? Expliquese.

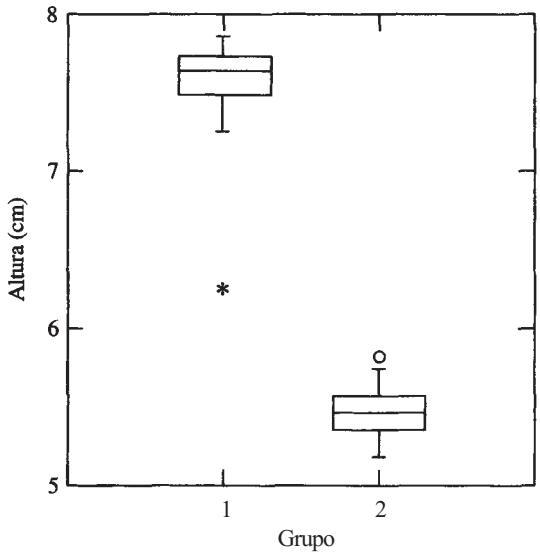


Figura 1.28. Diagrama de cajas de las alturas de la planta *Echinacea*

e) Basándose en los diagramas, ¿puede pensarse que la hidroponía tiende a producir plantas más altas que el cultivo en suelo? Expliquese.

(Basado en un estudio de Mary Cappuccio, Departamento de Biología, Universidad de Radford, 1996.)

1.7. MANEJO DE DATOS AGRUPADOS (OPCIONAL)

En muchos casos, en las publicaciones científicas se presentan los datos dispuestos en una tabla o en forma de clases. No se dan ni datos absolutos ni medidas resumen. Cuando esto ocurre, es conveniente extraer de la tabla al menos una primera aproximación de varias medidas resumen para el conjunto de datos subyacentes. Los métodos para hacerlo se presentan en los ejemplos siguientes.

Ejemplo 1.7.1. En un estudio del síndrome de Down, se examinaron 180 niños afectados. La Tabla 1.23. da la distribución de frecuencias para la variable *X,* cociente intelectual (CI) de los niños. Para calcular aproximadamente la media de CI, determinemos primero el punto medio de cada categoría o clase. El punto medio es la media aritmética de los límite» superior e inferior de la clase. Así, el punto medio para la primera clase, que designamos por m_1 es:

$$m_1 = \frac{10.5 + 20.5}{2} = 15.5$$

Los puntos medios sucesivos m_2, m_3,..., m_9 vienen dados en la Tabla 1.23.

Mientras que las clases no sean excesivamente amplias, el punto medio de cada clase, o marca de clase, sirve como una buena aproximación para cada uno de los valores de la clase Para aproximar la media muestral, cada una de las cuatro observaciones de la clase 1, cuyos; valores efectivos son desconocidos para nosotros, se reemplaza por el número 15.5; cada una de las 34 observaciones en la clase 2 se reemplaza por el número 25.5; el procedimiento

Tabla 1.23. Distribución de frecuencias de CI

Clase	Límites de clase	Marca de clase m_i	Frecuencia f_i
1	10.5 a 20.5	15.5	4
2	20.5 a 30.5	25.5	34
3	30.5 a 40.5	35.5	0
4	40.5 a 50.5	45.5	70
5	50.5 a 60.5	55.5	43
6	60.5 a 70.5	65.5	19
7	70.5 a 80.5	75.5	7
8	80.5 a 90.5	85.5	2
9	90.5 a 100.5	95.5	1

continúa para las otras clases hasta que, finalmente, la única observación de la clase 9 es reemplazada por el valor 95.5. Para aproximar *x,* sumamos estos valores y dividimos por 180, el número total de niños en el estudio. De este modo:

$$\bar{x} \cong \frac{4(15.5) + 34(25.5) + 0(35.5) + \cdots + 2(85.5) + 1(95.5)}{180}$$

$$= 47.4$$

donde el símbolo $\cong$ significa «aproximadamente igual a». El CI medio para este grupo de niños es aproximadamente 47.4.

Puede emplearse un procedimiento semejante para aproximar la varianza muestral para datos agrupados. Una vez más, el método está basado en la presunción de que el punto medio de la clase proporciona una buena aproximación para cada una de las observaciones de la clase. También se utiliza la fórmula del atajo para el cálculo de s^2, dada en el Ejercicio 1.5.13.

Definición 1.7.1. $\bar{x}$ y s^2, datos agrupados. Sea $x_1, x_2, x_3, \ldots x_n$ un conjunto de n observaciones correspondiente a una variable *X,* categorizadas en *k* clases. Sean m_i y f_i (i = 1,2,3,..., *k)* el punto medio de la clase y la frecuencia de la clase, respectivamente. Por consiguiente:

$$\bar{x} \cong \sum_{i=1}^{k} \frac{f_i m_i}{n}$$

y

$$s^2 \cong \frac{n \sum_{i=1}^{k} f_i m_i^2 - \left(\sum_{i=1}^{k} f_i m_i \right)^2}{n(n-1)}$$

Ejemplo 1.7.2. El valor aproximado de s^2 para los datos agrupados del Ejemplo 1.7.1 viene dado por:

$$s^2 \cong \frac{n \sum_{i=1}^{k} f_i m_i^2 - \left(\sum_{i=1}^{k} f_i m_i \right)^2}{n(n-1)}$$

$$= \frac{180 \sum_{i=1}^{9} f_i m_i^2 - \left(\sum_{i=1}^{9} f_i m_i \right)^2}{180(179)}$$

Para estos datos:

$$\sum_{i=1}^{9} f_i m_i^2 = 4(15.5)^2 + 34(25.5)^2 + \cdots + 2(85.5)^2 + 1(95.5)^2 = 445\,595$$

y

$$\sum_{i=1}^{9} f_i m_i = 4(15.5) + 34(25.5) + \cdots + 2(85.5) + 1(95.5) = 8540$$

Por tanto:

$$s^2 \cong \frac{180(445\,595) - 8540^2}{180(179)} = 225.81$$

y

$$s \cong \sqrt{225.81} \cong 15.0$$

Esbozaremos aquí un método que puede ser utilizado para aproximar cualquier observación específica en un conjunto de datos agrupados. Cuando el punto de interés es el de «en medio» del conjunto de datos, el método puede ser aplicado para obtener una aproximación para la mediana muestral.

Supongamos que $x_1, x_2, x_3, \ldots, x_n$ es un conjunto de observaciones, linealmente ordenadas, correspondientes a una variable X. Denominamos x_j. a cualquiera de esas observaciones. Para, aproximar x_j a partir del grupo de datos, se utilizan los siguientes pasos:

1. Localizar la clase en la que está *xy,* denominaremos *f* a la frecuencia de esta clase.
2. Encontrar los límites inferior y superior para esta clase; serán *l* y *u* respectivamente.
3. Encontrar la diferencia entre *j* y la frecuencia acumulada para la clase inmediatamente precedente a aquella en la que está *xy* denominamos *d* a esta diferencia.
4. El valor aproximado de x_j es:

$$x_j \cong l + \frac{d}{f}(u - l)$$

Resulta más sencillo aplicar este procedimiento que esbozarlo. Consideremos el Ejemplo 1.7.3.

Ejemplo 1.7.3. Se realiza un estudio para valorar el efecto del alcohol sobre los niveles de colesterol en suero. Una variable de interés es *X,* cantidad de alcohol consumido por semana y por sujeto. Los datos para los 923 sujetos que participan en el estudio vienen dados en la Tabla 1.24.

Tabla 1.24. Distribución de frecuencias en el consumo de alcohol (en onzas)

Clase	Límites de clase	Frecuencia *f*	Frecuencia acumulada	Frecuencia relativa acumulada
1	0 a 0.5	201	201	0.218
2	0.5 a 3.5	372	573*	0.621
3	3.5 a 9.5	260	833	0.903
4	9.5 a 19.5	80	913	0.989
5	≥19.5	10	923	1.000

El número total de observaciones es $n = 923$. Puesto que el número es impar, la posición de la mediana es $(n + 1)/2 = 462$ y la mediana muestral es $\tilde{x} = x_{462}$. De este modo, estamos tratando de aproximar $x_j = x_{462}$. De la distribución de frecuencias acumuladas puede deducirse que la observación 462 cae en la segunda clase (identificada por el símbolo *). La frecuencia de esta clase es $f = 372$. El límite inferior de la clase es $l = 0.5$ y el límite superior es $u = 3.5$. La diferencia entre j y la frecuencia acumulada para la clase 1 es $d = 462 - 201 = 261$. De este modo:

$$\tilde{x} = x_{462} \cong l + \frac{d}{f}(u - l)$$

$$= 0.5 + \frac{261}{372}(3.5 - 0.5)$$

$$= 2.6$$

(El símbolo $\cong$ significa «aproximadamente igual a».) La mediana aproximada del consumo de alcohol es 2.6 onzas por semana.

EJERCICIOS 1.7

1. Considérese la Tabla 1.25.
   a) Completar la Tabla 1.25, hallando el punto medio *(marca de clase)* y la frecuencia acumulada para cada clase.
   b) Obtener de forma aproximada la media, varianza, desviación típica y mediana muestrales.
2. Se realiza un estudio sobre la edad de las mujeres que utilizan anticonceptivos orales. Los datos agrupados están recogidos en la Tabla 1.26.
   a) Completar la Tabla 1.26, encontrando el punto medio y la frecuencia acumulada para cada clase.
   b) Obtener de forma aproximada la media, varianza, desviación típica y mediana muestrales.
3. Se realizó un estudio de la enfermedad de Hodgkin en pacientes menores de 40 años. Uno de los propósitos del estudio era comparar la distribución de casos por edad en hombres, con la de mujeres. Los datos agrupados se muestran en la Tabla 1.27.
   a) Completar la Tabla 1.27, hallando el punto medio y la frecuencia acumulada para cada clase.
   b) Para cada grupo, obtener de forma aproximada la media, varianza, desviación típica y mediana muestrales. Poner de manifiesto las semejanzas y las diferencias entre los dos grupos.

Tabla 1.25. Distribución de frecuencias

Clase	Límites de clase	Marca de clase	Frecuencia	Frecuencia acumulada
1	4.5 a 9.5		1	
2	9.5 a 14.5		2	
3	14.5 a 19.5		5	
4	19.5 a 24.5		3	

Tabla 1.26. Distribución de frecuencias por edades (en años)

Clase	Límites de clase	Marca de clase	Frecuencia	Frecuencia acumulada
1	14.5 **a** 19.5		171	
2	19.5 **a** 24.5		785	
3	24.5 **a** 29.5		837	
4	29.5 **a** 34.5		554	
5	34.5 **a** 39.5		382	
6	39.5 **a** 44.5		432	
7	44.5 **a** 49.5		562	
8	49.5 **a** 54.5		**610**	
9	54.5 **a** 59.5		**490**	
10	59.5 **a** 64.5		258	
11	**64.5 a** 69.5		153	
12	69.5 **a** 74.5		**60**	

Tabla 1.27. Distribución de casos por edades (en años)

Varones

Clase	Límites de clase	Marca de clase	Frecuencia	Frecuencia acumulada
1	4.5 **a** 9.5		**1**	
2	9.5 **a** 14.5		**4**	
3	14.5 **a** 19.5		**7**	
4	19.5 **a** 24.5		23	
5	24.5 **a** 29.5		**16**	
6	29.5 **a** 34.5		**7**	
7	34.5 **a** 39.5		**10**	

Mujeres

Clase	Límites de clase	Marca de clase	Frecuencia	Frecuencia acumulada
1	4.5 **a** 9.5		**0**	
2	9.5 **a** 14.5		**2**	
3	14.5 **a** 19.5		**10**	
4	**19.5 a** 24.5		**7**	
5	24.5 **a** 29.5		**3**	
6	29.5 **a** 34.5		**5**	
7	34.5 **a** 39.5		**2**	

HERRAMIENTAS COMPUTACIONALES

Los cálculos necesarios para resumir o presentar los datos de un conjunto de datos, incluso cuando éste sea pequeño, pueden ser tediosos y se consume mucho tiempo en ellos. Por un lado, hay muchas calculadoras de mano en el mercado que están programadas para hacer

cálculos estadísticos. Por otro, existen varios paquetes informáticos cuyo propósito principal es el análisis estadístico. En las secciones de Herramientas computacionales de este libro (al final de la mayoría de los capítulos) comentaremos un paquete informático y una calculadora científica. Si tiene acceso a estas herramientas, podrán seguirse las instrucciones dadas de forma exacta. Si se usan una calculadora o un programa diferentes, los pasos deberán modificarse. La calculadora elegida es la TI83; el paquete informático usado es el SAS.

La calculadora gráfica TI83

La calculadora gráfica T183 constituye una poderosa herramienta de ayuda al análisis de datos estadísticos. Es fácil de usar. Las instrucciones que aquí se dan, permitirán realizar muchas de las operaciones presentadas en el libro y, en concreto, se usarán los datos del texto para ilustrar el manejo de la calculadora. Si se tiene una, es recomendable estudiar el manual de usuario. Descubrirá así que la calculadora puede hacer más de lo que se explica en este libro.

I. Reiniciar/Limpiar

Empecemos por considerar cómo devolver la TI83 a su configuración de fábrica. En estas instrucciones se indican las teclas y el propósito de las mismas. Debería observar lo que aparece en pantalla. Muchos pasos son autoexplicativos y pronto será capaz de llevar a cabo estas tareas sin tener que memorizar las rutinas dadas en el libro.

Tecla/Comando de la TI83		**Propósito**
1.	2^{ND}	1. Reinicia la configuración de fábrica.
	+	
	5	
	1	
	2	
2.	2^{ND}	2. Oscurece la pantalla (si fuera necesario).
	△ (mantener)	
3.	CLEAR	3. Limpia la pantalla.

Nota: Al mantener pulsada la tecla 2^{ND} se ilumina la pantalla (si esto fuera necesario).

Este proceso de *reiniciar,* borra todos los datos almacenados, configuraciones y programas. Si no está guardando nada, puede usarse esta rutina para vaciar la memoria de la calculadora antes de empezar cada problema nuevo.

Si quiere borrar los datos contenidos en la memoria de la calculadora, sin cambiar nada más, puede usar las siguientes teclas.

Tecla/Comando de la TI83	**Propósito**
STAT	Limpia el editor de datos estadísticos;
4	Devuelve el color blanco a la pantalla.
2^{ND}	
1	
ENTER	
CLEAR	

Nota: Estas instrucciones borran los datos de la columna L1. Para borrar los de las columnas Ll y L2, añada:

, (coma)
2^{ND}
2
justo antes de ENTER

II. Histogramas

Puede utilizarse la calculadora TI83 para construir histogramas. Para hacerlo, debe conocerse el límite inferior de la primera clase, el límite superior de la última clase y la amplitud de la clase. Ilustraremos la construcción del histograma de la Figura 1.8. Los datos usados son los del Ejemplo 1.3.1. Se supone que se ha borrado la memoria de la calculadora.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 135 ENTER 137 ENTER ⋮ 155 ENTER	2. Introduce los datos (asegúrese de introducir todos los datos; los puntos indican que no se muestran aquí todos los datos).
3. WINDOW	3. Accede a la ventana para especificar los límites.
4. 129.5 ENTER 159.5 ENTER 6 ENTER (-)5 ENTER 20 ENTER ENTER	4. Establece el límite, $x_{mín}$, en 129.5. Establece el límite superior, $x_{máx}$, en 159.5. Establece el ancho de clase en 6. Establece un valor $y_{mín} = -5$ para proporcionar espacio debajo de histograma. Permite una frecuencia de clase máxima de 20.
5. 2^{ND} Y = ENTER cursor a ON ENTER ▽ ▷ ▷ ENTER GRAPH	5. Dibuja el histograma.
6. TRACE	6. Determina las frecuencias y límites moviendo el cursor a derecha e izquierda.

III. Hallar $\bar{x}$, s, s^2, q_1, q_3, y la mediana

La calculadora TI83 permite obtener la mayor parte de los estadísticos descritos en este libro. Mostraremos cómo se lleva a cabo esta tarea para un conjunto de datos, usando los del lugar I del Ejemplo 1.5.6. Obsérvese que la calculadora no estima q_1 y q_3 como se hizo en el texto. Si la posición de la mediana es un número entero, entonces la de los cuartiles se toma como la de la mediana dividida entre 2; si la posición de la mediana no es un número entero, la de los cuartíles coincidirá con lo dicho en este libro. Se supone que se ha borrado la memoria de la calculadora.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 10 ENTER 20 ENTER ⋮ 19 ENTER	2. Introduce los datos.
3. STAT ▷ 1 ENTER (use ▽ para ver otros)	3. Calcula los estadísticos básicos.
4. VARS 5 3 X^2 ENTER	4. Calcula s^2.

Observe que, puesto que la posición de la mediana es 12, que es un entero, la TI83 toma la posición del cuartil como el cociente entre la posición de la mediana y 2. En este caso, 6. Esto hace que $q_1 = 13$ y $q_3 = 31$. Estos valores difieren ligeramente de los datos proporcionados en el texto. La siguiente sección muestra cómo obtener los cuartíles de manera fácil.

IV. Ordenar

Si se desean hallar q_1 y q_3, tal como se hace en el texto, resultará útil la opción de ordenar. Ilustraremos el procedimiento con los datos del lugar I del Ejemplo 1.5.6. Se supone que se ha borrado la memoria de la calculadora.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 10 ENTER 20 ENTER ⋮ 19 ENTER	2. Introduce los datos (asegúrese de introducir todos los datos; los puntos indican que no se muestran todos ellos).

3.	STAT 2 2^{ND} 1 ENTER STAT 1	3. Ordena los datos y muestra la lista ordenada.

Observe que en el ejemplo del texto, la posición del cuartil es 6.5. Para hallar q_1 todo lo que teníamos que hacer era hallar los datos 6° y 7° de la nueva lista ordenada y hallar el punto medio. En este caso, el 6° valor era 13 y el 7° era 14, por lo que la media era 13.5. El cuartil q_3 se puede calcular encontrando los valores 6° y 7° contando desde el final.

V. Diagramas de cajas

La TI83 construye dos tipos de diagramas de cajas. Uno de ellos usa el valor máximo y el mínimo para calcular las *patillas.* El otro define éstas por los valores adyacentes, como se ha hecho en el texto. La última representación mostrará e identificará los datos atípicos. Mostraremos cómo se construye un diagrama de cajas usando los datos del Ejemplo 1.6.1.

Tecla/Comando de la TI83		**Propósito**
Max/Mín	**Atípicos**	
1. STAT 1	1. STAT 1	1. Accede al editor de datos estadísticos.
2. 8 ENTER 12 ENTER ⋮	2. 8 ENTER 12 ENTER ⋮	2. Introduce los datos (asegúrese de introducir cada dato; los puntos indican que no se muestran todos los datos).
3. WINDOW 0 ▽ 110 ▽	3. WINDOW 0 ▽ 110 ▽	Establece la escala *x* por debajo de x_{min} y por encima de $x_{máx}$; proporciona sitio debajo del diagrama.
4. 2^{ND} *Y* = ENTER cursor a ON ENTER ▽ ▷ ▷ ▷ ▷ ENTER GRAPH	4. 2^{ND} *Y* = ENTER cursor a ON ENTER ▽ ▷ ▷ ▷ ENTER GRAPH	4. Dibuja un diagrama de cajas.
5. TRACE	5. TRACE	5. Lee los valores de la mediana, q_1, q_2, *x* máxima, *x* mínima (o a_1, a_3 y datos atípicos)

El paquete estadístico informático SAS

Hay muchos paquetes estadísticos informáticos en el mercado. Entre los sistemas más extendidos están: SPSS *(Statistical Package for the Social Sciences,* McGraw-Hill), BMD *(Biomedical Computer Programs,* University of California Press), MINITAB *(Pennsylvania State University)* y SAS *(Statistical Analysis System,* SAS Institute, Inc.). Para usar cualquiera de ellos en análisis simples de datos, sólo se necesitan unos pocos conocimientos de informática.

Aquí presentamos una breve introducción a la programación en SAS, para darle alguna experiencia en paquetes estadísticos. Una vez conseguido esto, no será difícil proceder con otros paquetes, puesto que todos son similares. Introduciremos el SAS presentando algunos programas de ejemplo que pueden ser modificados para analizar los conjuntos de datos presentados en este capítulo. Debe consultar a un experto para instalar el SAS en su ordenador, así como para guardar y ejecutar programas.

I. Tablas de frecuencia de una entrada

Empecemos por presentar el programa usado para hacer la Tabla 1.4 de este libro. Los datos están en el Ejemplo 1.1.1.

Cada programa SAS consta de tres partes: datos, entrada de los datos y el análisis de los datos. La finalidad de la primera parte es nombrar cada conjunto de datos, nombrar las variables, configurar las opciones de impresión, y crear nuevas variables si es necesario. La primera línea de código que usaremos es:

```
OPTIONS LS = 80     PS = 60     NODATE;
```

La palabra OPTIONS es un comando del SAS. Debe escribirse correctamente. Esta sentencia indica que la salida debe imprimirse con 80 caracteres por línea y 60 líneas por página. Así se configura que la salida impresa se haga en un tamaño de papel similar al de un cuaderno o máquina de escribir, en lugar del tamaño de papel propio de los ordenadores, que es ligeramente superior. La opción NODATE suprime la impresión de la fecha en la salida de los datos. Obsérvese que los comandos SAS terminan con punto y coma. *Los comandos SAS pueden empezar en cualquier columna y siempre acaban con punto y coma.* La siguiente línea es:

```
DATA RESID;
```

El nombre del conjunto de datos, RESID, proviene de que los datos son de una residencia de adultos. Los nombres de los ficheros de datos no pueden superar los 8 caracteres y deben reflejar el ámbito del estudio. La palabra DATA es un comando SAS. A continuación nombraremos las variables por una sentencia INPUT. Los nombres de las variables tampoco pueden tener más de 8 caracteres y deben reflejar el tipo de información que estudian. En este caso, escribiremos:

```
INPUT     SEXO     $     DIAGNOST     $     EDAD     DESTINO;
```

Esta sentencia especifica que cada fila de datos contendrá el valor de cada una de las cuatro variables: sexo del paciente, diagnóstico, edad del paciente y destino después de dejar la residencia. El $ después de SEXO y DIAGNOST indica que los valores de esas variables serán letras en vez de números. La siguiente línea del programa es:

```
LINES;
```

Ésta indica a SAS que, a continuación, se introducirán los datos, con lo que concluye la primera parte del programa.

Comienza entonces la fase de introducción de los datos. Como se está considerando el Ejemplo 1.1.1, las líneas de datos son así:

```
M    EM    29   2 )
M    RM     3 5 7 } e s primeras líneas de datos
F    FE    34   7 )
⋮
F    RM    18   7 } última línea de datos
```

Obsérvese que los puntos indican que no se muestran todos los datos, aunque se deban introducir. Obsérvese también que no hay punto y coma al final de cada línea de datos. Para señalizar el final de las líneas de datos se pone punto y coma en una línea aparte:

Esto completa la fase de introducción de datos.

El SAS utiliza una serie de procedimientos, o PROCS, para analizar los datos. A lo largo de este libro se introducirán algunos de los más comunes. El procedimiento usado para generar la Tabla 1.4 fue PROC FREQ. El comando del SAS, PROC, significa *procedimiento;* el procedimiento deseado es uno que realiza cálculos de frecuencia, de frecuencias acumuladas y porcentajes. Su nombre es PROC FREQ. Para indicar a SAS lo que se quiere, escribimos:

```
PROC    FREQ;    TABLES    DIAGNOST;
```

TABLES también es un comando SAS. Con él indicamos a SAS que obtenga una tabla de frecuencias de la variable DIAGNOST, una de las cuatro de nuestro conjunto de datos. Esto completa la parte final, la del análisis de los datos, de nuestro primer programa.

Si se desea, se pueden añadir títulos tras el paso PROC. En este caso titulamos nuestra tabla:

```
Frecuencias y Porcentajes:
Variable Diagnóstico
```

Para conseguir esto escribimos:

```
TITLE     'Frecuencias y Porcentajes:';
TITLE2    'Variable Diagnóstico';
```

Obsérvese que la primera sentencia TITLE, que produce la primera línea del título, no tiene que ir numerada. La segunda línea, que produce la segunda línea del título, sí debe numerarse. Obsérvese también que los títulos están encerrados en comillas simples y terminan con punto y coma. En esta exposición, y en las siguientes, escribiremos el código SAS en mayúsculas para facilitar la lectura. No obstante, *el código no tiene por qué introducirse en el ordenador en mayúsculas.*

Nuestro primer programa quedaría así:

```
OPTIONS LS = 80 PS = 60 NODATE;
DATA RESID;
INPUT SEXO $ DIAGNOST $ EDAD DESTINO;
LINES;
M    EM    29    2
M    RM    35    7
F    FE    34    7
⋮
```

```
F     RM    18    7

PROC FREQ; TABLES DIAGNOST;
TITLE 'Frecuencias y Porcentajes: ';
TITLE2 'Variable Diagnóstico';
```

Puede utilizarse este programa para manipular otros datos, cambiando el nombre del conjunto de datos, así como los nombres de las variables y el título.

II. Diagrama de barras horizontal: datos discretos

Ahora presentamos el código requerido para hacerun gráfico de barras horizontal. Los datos utilizados son los del Ejemplo 1.1.1. El código siguiente producirá el gráfico mostrado en la Figura 1.3. Se da el código SAS y una breve explicación. Las líneas en blanco sólo están para dejar espacio a la explicación, cuando se introduzca en el ordenador el código no debe haber líneas en blanco.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Configura la salida impresa.
DATA RESID;	Nombra el conjunto de datos.
INPUT SEXO $ DIAGNOST $ EDAD DESTINO;	Nombra las variables.
LINES;	Indica que, a continuación, vienen los datos.
M EM 29 2 M RM 35 7 F FE 34 7 ⋮ F RM 18 7	Líneas de datos.
	Señala el final de los datos.
PROC CHART;	Pide el procedimiento para crear gráficos de barras.
HBAR DIAGNOST/DISCRETE;	Indica que se dibuje un gráfico de barras horizontales para la variable DIAGNOST; comunica al SAS que la variable es discreta.
TITLE 'Gráfico de barras horizontales:';	Título (**1.ª** línea).
TITLE2 'Variable Diagnóstico';	Título (**2.ª** línea).

De la misma forma se pueden hacer gráficos de barras verticales. Para ello, basta con cambiar el código **HBAR** por VBAR.

III. Tablas de doble entrada

Pueden crearse las tablas de doble entrada mediante el procedimiento PROC FREQ. A continuación se presenta el código usado para crear la Tabla 1.8, basándose en los datos sin procesar dados en el Ejemplo 1.1.1.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Configura la salida impresa.
DATA RESID;	Nombra el conjunto de datos.
INPUT SEXO $ DIAGNOST $ EDAD DESTINO;	Nombra las variables.
LINES;	Indica que, a continuación, vienen los datos.
M EM 29 2 M RM 35 7 F FE 34 7 ⋮	Líneas de datos.
F RM 18 7	Señala el final de los datos.
PROC FREQ;	Pide el procedimiento para crear gráficos de barras.
TABLES SEX*DIAGNOST;	Indica que se dibuje un gráfico de barras horizontal para la variable DIAGNOST; comunica a SAS que la variable es discreta.
TITLE 'Tabla de doble entrada';	Título (1.ª línea).
TITLE2 'Relación entre Sexo y Diagnóstico';	Título (2.ª línea).

IV. Histogramas: vertical y horizontal

Ilustraremos ahora el procedimiento SAS para construir un histograma vertical, partiendo de los datos del Ejemplo 1.3.1. El código usado para hacer los histogramas de las Figuras 1.10 y 1.12, es el siguiente.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Configura la salida impresa.
DATA JERBOS;	Nombra el conjunto de datos.
INPUT GRUPO $ @; DO I = 1 to 30; INPUT LLAMADAS @ @; OUTPUT;	Introduce los datos; permite hacerlo con sólo introducir el identificador de grupo una vez.
LINES;	Indica que los datos vienen a continuación.
e 135 137 148 152 ... 155	Introduce los datos del grupo experimental.
c 123 109 118 116 ... 90	Introduce los datos del grupo de control.
	Señala el fin de los datos.
DATA EXP; SET JERBOS; IF GRUPO = 'e';	Forma un nuevo conjunto de datos que contiene sólo los animales experimentales.
PROC CHART; VBAR LLAMADAS/ MIDPOINTS = 132.5 138.5 144.5 150.5 156.5;	Especifica un histograma vertical para la variable LLAMADAS; indica a SAS que utilice los mismos puntos medios que los del texto; reproduce la Figura 1.10.

TITLE 'HISTOGRAMA DE FRECUENCIAS: EXPERIMENTALES';	Titula la salida impresa.
PROC CHART; HBAR LLAMADAS/ MIDPOINTS = 132.5 138.5 144.5 150.5 156.5;	Selecciona un histograma horizontal para la variable LLAMADAS; especifica puntos medios; reproduce la Figura 1.12.
TITLE 'TABLA RESUMEN E HISTOGRAMA HORIZONTAL';	Título (1.ª línea).
TITLE2 'EXPERIMENTALES';	Título (2.ª línea)

V. Resumen de estadísticos/tallo y hojas/diagramas de cajas

Los datos del Ejemplo 1.6.1 se usan para ilustrar el procedimiento PROC UNIVARLATE. Este procedimiento calculará todos los estadísticos citados en el capítulo y también construirá un diagrama de tallo y hojas y un diagrama de cajas para los datos.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Configura la salida impresa.
DATA AMNESIA;	Nombra el conjunto de datos.
INPUT DÍAS @ @;	Nombra las variables; @ @ permite poner más de un dato puntual por línea; debe haber al menos un espacio entre cada valor.
LINES;	Señala que los datos vienen a continuación.
8 12 20 27 30 32 35 36 40 40 40 40 41 42 45 47 50 52 61 89 108	Líneas de datos.
	Señala el fin de los datos.
PROC UNIVARLATE PLOT;	Pide el procedimiento que calcula los estadísticos básicos y dibuja los gráficos.
TITLE 'ESTADÍSTICOS BÁSICOS Y GRÁFICOS TITLE2 'VÍA PROC UNIVARIATE';	

En la salida impresa aparecen estos estadísticos (los números rodeados por un círculo, corresponden a los estadísticos señalados en la salida impresa):

① Tamaño muestral
② Media muestral, $\bar{x}$
③ Desviación típica muestral, s
④ Varianza muestral, s^2
⑤ Mediana, $\tilde{x}$
⑥ Tercer cuartil, q_3
⑦ Primer cuartil, q_1
⑧ Rango
⑨ Rango intercuartílico, *iqr*
⑩ Tallo y hojas
⑪ Diagrama de cajas

ESTADÍSTICOS BÁSICOS Y GRÁFICOS
VÍA PROC UNIVARIATE

Procedimiento para una sola variable

Variable = DÍAS

Momentos

①	N	21	Sum Wgts	21
②	Mean	42.61905	Sum	895
③	StdDev	22.613	Variance	511.3476 ④
	Skewness	1.42 0384	Kurtosis	3.131485
	USS	48371	CSS	10226.95
	CV	53.05843	Std Mean	4.93456
	T:Mean=0	8.636848	Pr>|T|	0.0001
	Num ^= 0	21	Num > 0	21
	M(Sign)	10.5	Pr>=|M|	0.0001
	Sgn Rank	115.5	Pr>=|S|	0.0001

Quantiles(Def=5)

	100%	Max	108	99%	108
⑥	75%	Q3	47	95%	89
⑤	50%	Med	40	90%	61
⑦	25%	Q1	32	10%	20
	0%	Min	8	5%	12
				1%	8
⑧	Range		100		
⑨	Q3-Q1		15		
	Mode		40		

Extremos

Lowest	Obs	Highest	Obs
8(	1)	50 (	17)
12 (	2)	52 (	18)
20(	3)	61 (	19)
27 (	4)	89(	20)
30(	5)	108 (	21)

	Stem	Leaf	#	Boxplot
	10	8	1	*
	8	9	1	0
⑩	6	1	1	| ⑪
	4	0000125702	10	+-+-+
	2	070256	6	+---+
	0	82	2	0

--+--+--+--+--+

Multiply Stem.Leaf by 10**+1

Introducción al cálculo de probabilidades y al cálculo combinatorio

En el Capítulo 1 hemos presentado algunos de los métodos utilizados para describir un conjunto de datos. Si el único propósito del investigador es describir los resultados de un experimento concreto, tales métodos pueden considerarse suficientes. No obstante, si lo que se pretende es utilizar la información obtenida para extraer conclusiones generales sobre todos aquellos objetos del tipo de los que han sido estudiados, entonces los métodos del Capítulo 1 constituyen solamente el principio de los análisis. Para obtener conclusiones válidas y hacer predicciones correctas acerca de una población a través de la observación de una parte de ella, debe recurrirse a métodos de inferencia estadística. Estos métodos implican el uso inteligente de la teoría de probabilidades.

La teoría de probabilidades es una rama interesante de las matemáticas y la base de la inferencia estadística. En las Secciones 2.2 a 2.6, se describen las técnicas de cálculo combinatorio y su utilización en el cálculo clásico de probabilidades. El Capítulo 3 presenta las definiciones básicas, axiomas y teoremas que dirigen el comportamiento de las probabilidades. Si se dispone de tiempo suficiente, deberán estudiarse estas secciones. No obstante, las Secciones 2.1 y 2.2 proporcionan una base suficiente para la comprensión del uso de las probabilidades en el análisis de datos. Por tanto, tras leer tales secciones, es posible pasar al Capítulo 4 sin pérdida de continuidad.

2.1. INTERPRETACIÓN DE LAS PROBABILIDADES

Cuando se formula la pregunta: «¿conoce usted algo sobre probabilidad?», la mayor parte de la gente responde con rapidez: «¡no!». Aunque, generalmente, no es éste el caso. Nuestra cultura asume la habilidad para interpretar adecuadamente la idea de probabilidad, por lo menos intuitivamente. Se oyen frases tales como «la probabilidad de que llueva hoy es del 95 %» o «hay un 10 % de probabilidad de que llueva». Se da por supuesto que el público en general sabe interpretar adecuadamente estos valores. Resumiendo, la interpretación de probabilidades puede sintetizarse de la siguiente forma:

1. Las probabilidades son números comprendidos entre 0 y 1, ambos inclusive, que reflejan las expectativas con respecto a que un suceso físico determinado ocurra.

2. Probabilidades próximas a 1 indican que cabe esperar que ocurran los sucesos de que se trate. No indican que el suceso vaya a producirse, sólo que es un tipo de suceso que generalmente se produce.
3. Probabilidades próximas a 0 indican que no cabe esperar que ocurran los sucesos de que se trate. No indican que el suceso no vaya a producirse, sólo que este tipo de sucesos se considera raro.
4. Probabilidades próximas a $\frac{1}{2}$ indican que es tan verosímil que el suceso se produzca como que no.

El intervalo [0, 1] puede entenderse como una escala que se utiliza para determinar la probabilidad de que se produzca un suceso. Cuanto más cerca de 1 se encuentre la probabilidad, más confianza tenemos en que éste se produzca. A un suceso que deba producirse con absoluta certeza se le asigna una probabilidad de 1. Cuanto más cerca de 0 se encuentre la probabilidad, más seguros estamos de que no se producirá el suceso. A un suceso que es físicamente imposible se le asigna una probabilidad de 0.

¿Qué podemos considerar como una probabilidad grande o pequeña? Sin duda, una probabilidad de 1 es grande y una probabilidad de 0 es pequeña. ¿Cuán cercana a estos extremos debe encontrarse una probabilidad para ser considerada grande o pequeña? No existe una respuesta bien definida para esta pregunta. La interpretación de las probabilidades es un tanto subjetiva, y la interpretación dada puede depender de las consecuencias de estar equivocado. Una probabilidad que se considera grande en un entorno puede parecer pequeña en otro. Por ejemplo, supongamos que tengo la oportunidad de hacer un paseo al aire libre y oigo que la probabilidad de que llueva es 0.1. Podría considerar que esta probabilidad es pequeña y concluir que es más probable que no llueva. Si estoy equivocado, sólo estaré incómodo. Sin embargo, supongamos que me han dicho que he sido seleccionado para ser el primer civil en descender al fondo del océano en un nuevo minisubmarino y que la probabilidad de que el vehículo falle es 0.1. ¡Yo rechazaría esta oferta! Las consecuencias del fallo son demasiado serias para aceptarlas. En el primer caso una probabilidad de 0.1 se considera pequeña, en el segundo caso, parece grande.

Las probabilidades que acabamos de presentar sirven de guía para la interpretación de las probabilidades una vez puestos a nuestra disposición una serie de valores dados, pero no indican el modo de asignar un valor de probabilidad a uno u otro suceso. Generalmente, se utiliza uno de los tres métodos siguientes: la estimación personal, la estimación por medio de la frecuencia relativa y la estimación clásica. El uso de cada uno de ellos tiene sus ventajas y sus inconvenientes.

> **Ejemplo 2.1.1.** Un paciente sufre de cálculos renales, y no se ha conseguido mejora alguna a partir de los métodos ordinarios. Su médico está planteándose llevar a cabo una intervención quirúrgica y debe responder a la siguiente pregunta. ¿Cuál es la probabilidad de que la operación sea un éxito? Varios factores, como son la edad del paciente, su estado general de salud y su actitud frente a la operación, intervienen en este caso. Esta particular combinación de factores es una peculiaridad de este paciente. El médico no se ha enfrentado antes con un caso *exactamente igual* a éste, ni espera enfrentarse a otro igual en el futuro. Es una situación peculiar y es preciso establecer un juicio de valores para resolverla. En este caso, cualquier probabilidad que se asigne al suceso «la operación será un éxito» es una *apreciación personal.*

Este ejemplo ilustra las ventajas e inconvenientes de una estimación personal. Su mayor ventaja consiste en que siempre es aplicable. Cualquiera puede establecer una apreciación personal sobre lo que sea. Su mayor inconveniente es obvio: su acierto depende de lo correcta que sea la información de que dispone y de la capacidad del científico para evaluarla adecuadamente.

Ejemplo 2.1.2. Un investigador trabaja en un nuevo fármaco para insensibilizar a los pacientes frente a picaduras de abejas. De 200 sujetos sometidos a prueba, 180 presentaron una disminución en la gravedad de los síntomas tras sufrir una picadura, después de ser sometidos al tratamiento. Es natural suponer, entonces, que la probabilidad de que ocurra lo mismo en otro paciente que reciba el mismo tratamiento es por lo menos de *aproximadamente*

$$\frac{180}{200} = 0.90$$

Basándose en este estudio, se informa de que el fármaco es eñcaz en un 90 % de los casos para disminuir la reacción de pacientes sensibles a las picaduras de abejas. Tal probabilidad *no* es simplemente una opinión personal. Es una asignación numérica basada en la repetición de una experiencia y en la observación de los resultados. Se trata, de hecho, de una *frecuencia relativa.*

El Ejemplo 2.1.2 ilustra las características de la frecuencia relativa como forma de estimación de la probabilidad. Es aplicable a cualquier situación en la que el experimento pueda ser repetido varias veces y sus resultados observados. Por lo tanto, la probabilidad aproximada de que se produzca un suceso determinado *A,* simbolizada por *P[A],* viene dada por

$$P[A] \cong \frac{f}{n} = \frac{\text{número de veces que ocurre } A}{\text{número de veces que se realiza el experimento}}$$

donde el símbolo $\cong$ significa «aproximadamente igual a». El inconveniente de este método de estimación es que puede que el experimento no se lleve a cabo siempre en las mismas condiciones; la experiencia debe ser repetible. La ventaja de este método sobre la aproximación personal es que, generalmente, es más precisa, porque se basa en la observación real más que en la apreciación personal. Debe tenerse en cuenta que cualquier probabilidad obtenida a partir de la frecuencia relativa es una aproximación. Se trata de un valor calculado sobre *n* pruebas. Si se llevasen a cabo experimentos adicionales podríamos obtener un valor aproximado diferente. Sin embargo, cuando el número de pruebas aumenta, las diferencias entre los valores aproximados obtenidos tienden a desaparecer. Por tanto, para un número grande de experimentos, la probabilidad aproximada que resulta a partir de la frecuencia relativa es, generalmente, muy acertada.

En la Sección 1.3 tratamos las probabilidades de frecuencia relativa. La distribución de frecuencias relativas y la distribución de frecuencias relativas acumuladas que encontramos allí pueden ser interpretadas como si fueran probabilidades. Por ejemplo, la Tabla 1.19 muestra que la aparición de la reacción a una picadura de insecto se produjo entre 8.2 y 10.3 minutos en 10 de los 40 pacientes estudiados. Podemos utilizar esta información para decir que la probabilidad de que un futuro paciente experimente una reacción en este mismo período de tiempo es de aproximadamente $\frac{10}{40}$. Es decir, esperamos que, aproximadamente, 1 de cada 4 pacientes experimente una reacción en el intervalo de 8.2 a 10.3 minutos. Obsérvese que la frecuencia relativa acumulada para la clase 5 es de $\frac{38}{40}$, A partir de esto, podemos decir que la probabilidad aproximada de que una reacción tarde como máximo 14.7 minutos en aparecer es de $\frac{38}{40}$. Las distribuciones de frecuencias relativas y relativas acumuladas pueden utilizarse para determinar aproximadamente el comportamiento futuro de una variable aleatoria. No podemos asegurar qué valores aparecerán en alguna prueba futura del experimento, pero podemos anticipar qué valores es probable que surjan y cuáles se considerarán raros.

Ejemplo 2.1.3. ¿Cuál es la probabilidad de que un niño nacido de una pareja, cada uno de cuyos miembros posee genes para ojos castaños y para ojos azules, tenga los ojos castaños? Para resolver esta cuestión, observemos que, dado que el niño recibe un gen de cada uno de sus padres, las posibilidades para él son (castaño, azul) (azul, castaño) (azul, azul) y (castaño, castaño), donde el gen que aparece representado en primer lugar en cada uno de los pares es el gen que procede del padre. Puesto que cada uno de los padres tiene exactamente la misma probabilidad de aportar un gen para ojos azules que uno para ojos castaños, las cuatro alternativas son equiprobables. Al ser dominante el gen para ojos castaños, tres de los cuatro pares dan como resultado un niño de ojos castaños. En consecuencia, la probabilidad de que el niño tenga los ojos castaños es $\frac{3}{4} = 0.75$.

Esta probabilidad no es una apreciación personal, ni tampoco se basa en la repetición de un experimento. De hecho, la hemos calculado por el *método clásico*. Tal método puede usarse siempre que los resultados posibles de un experimento sean *equiprobables*. En este caso, la probabilidad de que ocurra el suceso A viene dada por:

$$P[A] = \frac{n(A)}{n(S)} = \frac{\text{número de veces que puede producirse } A}{\text{número de resultados que puede dar el experimento}}$$

Este método tiene también ventajas e inconvenientes. Su principal inconveniente es que no siempre es aplicable; se necesita que los resultados posibles sean equiprobables. Su mayor ventaja es que, si es aplicable, la probabilidad obtenida es exacta. Por otra parte, no exige la realización de experiencias ni la recogida de datos y es de fácil uso.

Los tres métodos desempeñarán su papel en algún momento y serán utilizados con frecuencia posteriormente.

EJERCICIOS 2.1

En cada uno de los ejercicios del 1 al 10 se pide calcular probabilidades. ¿Qué método (personal, frecuencia relativa o clásico) considera el más apropiado para resolver el problema? Justifique su elección. En los casos en que sea posible, halle la probabilidad, exacta o aproximada, pedida.

1. Una mujer contrae la rubéola durante el embarazo. ¿Cuál es la probabilidad de que su hijo nazca con algún defecto congénito?
2. Se somete a prueba un fármaco que se pretende utilizar en el tratamiento de la afección cutánea producida por una hiedra venenosa. De 190 personas a las que les fue aplicado, 150 obtuvieron algún beneficio del mismo. ¿Cuál es la probabilidad de que este fármaco sea eficaz con el siguiente paciente al que se aplique?
3. Un etólogo estudia un numeroso grupo de babuinos en libertad. Observa que de los 150 animales del grupo, 5 tienen el pelo de un color extremadamente claro. ¿Cuál es la probabilidad de que la siguiente cría de babuino que nazca en el grupo porte esta coloración clara?
4. Un bioquímico planea aislar y purificar una enzima que ha sido recientemente descubierta en las hojas de las espinacas. Consulta la literatura especializada para obtener alguna información que le sirva de guía en el diseño de un procedimiento de purificación, pero, puesto que se trata de una enzima muy particular que no ha sido aislada con anterioridad, no hay ninguna metodología específica disponible. ¿Cuál es la probabilidad de que un procedimiento de nuevo diseño tenga éxito?

5. Un hombre es zurdo y su mujer diestra. La pareja tiene dos niños. Cada uno de ellos tiene exactamente la misma probabilidad de ser zurdo que de ser diestro. ¿Cuál es la probabilidad de que los dos sean zurdos?
6. En un laboratorio se cometen errores en los análisis cruzados de sangre, en uno de cada 2000 análisis. ¿Cuál es la probabilidad de que un determinado análisis esté equivocado?
7. Un químico sabe por experiencia que, aproximadamente, 8 de cada 100 de las muestras que recibe para localizar fosfatos contienen demasiado poco para que éstos puedan ser detectados en un análisis rutinario. ¿Cuál es la probabilidad de que tenga que usar un método alternativo, más sensible, en la siguiente muestra que reciba para su análisis?
8. De 140 trabajadores de la construcción sometidos a examen, 98 presentaban un número excesivo de astillas de amianto. ¿Cuál es la probabilidad de que uno de estos trabajadores seleccionado al azar tenga en los pulmones una cantidad indebida de astillas de amianto?
9. En un banco de sangre hay cinco unidades disponibles de sangre del grupo A+. Una de ellas está erróneamente etiquetada y es, de hecho, del grupo 0. Se selecciona aleatoriamente una de estas cinco unidades para llevar a cabo una transfusión. ¿Cuál es la probabilidad de que la unidad elegida sea la que ha sido erróneamente etiquetada?
10. Una plaga que ataca a los pinos es tal que el 50 % de los árboles afectados muere en un año. Hay tres árboles afectados. ¿Cuál es la probabilidad de que mueran todos los árboles? ¿Cuál es la probabilidad de que muera al menos uno de ellos?
11. Basándose en la información de la Tabla 1.19, determinar, aproximadamente, estas probabilidades:
    a) La probabilidad de que la aparición de la reacción se produzca en 12.5 minutos.
    b) La probabilidad de que no se produzca ninguna reacción en los primeros 5.9 minutos.
    c) La probabilidad de que se produzca una reacción en el intervalo de 12.6 a 14.7 minutos.
12. Utilizar los datos del Ejercicio 7 de la Sección 1.3 para calcular aproximadamente la probabilidad de que el porcentaje de cenizas en una futura muestra de turba de pantano sea como máximo del 1.8 %.
13. Se sabe que la infección por el VIH puede ser contraída por enfermeros, técnicos de urgencias, doctores y otros profesionales por un pinchazo accidental con una aguja. Supóngase que, para cada pinchazo, la probabilidad de que ocurra es 0.05. ¿Cree que es alta o, por el contrario, le parece baja esa probabilidad? Explíquelo.
14. Un método de extinción de incendios forestales consiste en preparar deliberadamente un petardeo. Supóngase que en un caso determinado se ha estimado que la probabilidad de que este método tenga éxito es 0.05. ¿Cree que es alta o baja? Explíquelo.

2.2. DIAGRAMA DE ÁRBOL Y GENÉTICA ELEMENTAL

Los problemas que hemos comentado en la Sección 2.1 son elementales porque el número de posibilidades en cada caso es muy pequeño. Cuando los experimentos son más complejos, es útil tener un método sistemático para obtener todos los resultados posibles. Un método para hacerlo es el *diagrama de árbol.* Esta técnica resulta útil cuando el experimento puede visualizarse como si se produjera en unos pocos pasos o etapas diferentes. Cada paso del experimento se representa como una ramificación del árbol. El árbol se forma determinando primero cuántas etapas están implicadas. En cada etapa las ramas del árbol representan las posibilidades en ese punto determinado. Una vez completado el árbol, pueden leerse las

secuencias de sucesos siguiendo lo que se denominan «trayectorias» a lo largo del árbol. El método se muestra en el Ejemplo 2.2.1.

Ejemplo 2.2.1. Una mujer es portadora de hemofilia clásica. Esto significa que, aunque la mujer no tenga hemofilia, puede transmitir la enfermedad a sus hijos. Da a luz a tres hijos. ¿Cuáles son las posibilidades de este experimento?

Puesto que lo que fundamentalmente nos interesa es si los hijos presentan o no la enfermedad, debemos generar un árbol que nos proporcione esta información. En este caso, tenemos tres etapas naturales, cada una de las cuales representa el nacimiento de un hijo. El primero tendrá (sí) o no tendrá (no) la enfermedad. Esto aparece en el diagrama de árbol de la Figura 2.1a, donde sí = *s* y no = *n.* Del mismo modo, el segundo hijo tendrá o no tendrá la enfermedad. Esto aparece en la Figura *2.1b.* Finalmente, el tercer hijo tendrá o no tendrá la enfermedad. Como consecuencia, el árbol completo será el que aparece en la Figura 2.1c. Se determina una trayectoria a lo largo del árbol empezando por el extremo izquierdo del mismo hasta el extremo derecho, sin desviar o levantar el lápiz. Así, la primera trayectoria a través del árbol es *sss,* lo que corresponde al hecho de que los tres hijos tienen hemofilia clásica.

El conjunto de posibilidades del experimento puede leerse en el árbol siguiendo cada una de las ocho trayectorias distintas a través del árbol. Este conjunto es

$$\{sss, ssn, sns, snn, nss, nsn, nns, nnn\}$$

Cuando el contexto físico del problema garantiza que cada trayectoria a lo largo del árbol es tan probable que se produzca como cualquier otra, entonces el árbol puede utilizarse para calcular las probabilidades utilizando el método clásico. Por ejemplo, en el Ejemplo 2.2.1 se sabe que un portador de hemofilia clásica es tan probable que transmita la enfermedad a un

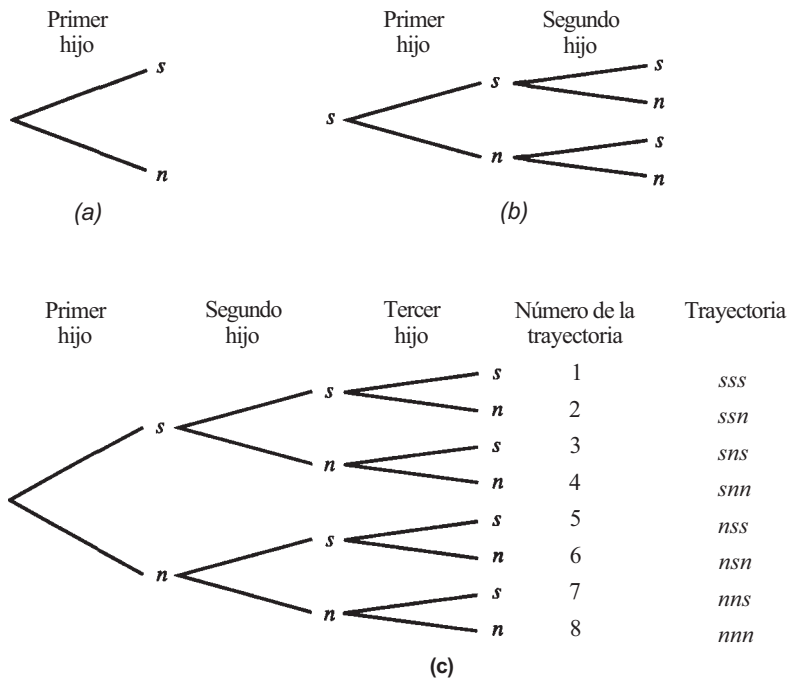


Figura 2.1. Construcción de un diagrama de árbol.

hijo como que no lo haga. Por esta razón, tenemos tantas posibilidades de elegir una ramificación como otra cualquiera en cada etapa del diagrama de la Figura 2.1. Esto implica que cada una de las ocho trayectorias a lo largo del árbol es equiprobable. Puede utilizarse este hecho para calcular la probabilidad de diversos sucesos relacionados con el experimento. El Ejemplo 2.2.2 muestra cómo se hace.

Ejemplo 2.2.2. ¿Cuál es la probabilidad de que una mujer con tres hijos y que es portadora de hemofilia clásica no transmita su enfermedad a ninguno de sus hijos? Sabemos que hay ocho trayectorias a lo largo del árbol en la Figura 2.1. De ahí que este experimento tenga ocho resultados equiprobables. En el árbol podemos ver que una trayectoria, *nnn,* representa el suceso en cuestión. Por lo tanto

$$P[\text{ninguno de los tres tiene la enfermedad}] = \frac{1}{8}$$

Del mismo modo, podemos decir **que** la **probabilidad** de que exactamente dos de los tres hijos tengan hemofilia es de $\frac{3}{8}$. Las tres trayectorias correspondientes a la aparición de este suceso son *ssn, sns, nss.*

En el Capítulo 4 se abordará el cálculo de probabilidades utilizando diagramas de árbol en los que las trayectorias no son equiprobables.

Genética elemental (opcional)

Los rasgos hereditarios de un organismo vienen determinados por unidades denominadas *genes.* Los genes se producen por parejas de individuos y se dan en formas que se pueden contrastar. Estas formas se denominan *alelos.* Por ejemplo, consideremos el gen que determina la altura de la planta del guisante. Este gen tiene dos alelos, T para la altura elevada y *t* para la altura baja. Así, existen tres posibles composiciones genéticas o *genotipos* respecto a esta característica. Estos son *TT, Tt y tt.* Cuando dos genes son de la misma forma, decimos que el organismo es *homocigoto* para esta característica; de lo contrario, es *heterocigoto.* Una característica que se manifieste cuando esté presente el alelo que controla esta característica se denomina una característica *dominante,* y el alelo es un alelo dominante. En caso contrario, su característica de contraste o alelo se dice que es *recesivo.* En el guisante, el alelo para la altura elevada es dominante. Así, los genotipos *TT* y *Tt* darán como resultado una planta alta, mientras que el genotipo *tt* dará como resultado una planta de altura baja. Obviamente, los alelos dominantes se representarán con letras mayúsculas y los alelos recesivos con minúsculas. Para cada característica, la descendencia hereda un gen aleatorio de cada uno de sus progenitores.

Pueden utilizarse los diagramas de árbol para resolver problemas genéticos simples. Para ilustrarlo, reconsideremos el Ejemplo 2.2.3.

Ejemplo 2.2.3. Cada uno de los miembros de una pareja tiene alelos tanto para ojos castaños como azules. En términos genéticos son heterocigotos para el color de los ojos. En el caso del color de los ojos, el alelo para los ojos castaños, que representamos con *B,* es dominante sobre el de los ojos azules, *b.* Es decir, cualquiera que tenga el alelo *B* tendrá ojos castaños. En el momento de la concepción, cada progenitor contribuye con un alelo para el color de los ojos. Por lo tanto, podemos considerar el experimento de la determinación del color de los ojos de un niño como un proceso de dos etapas. La etapa 1 representa la herencia de un alelo de la madre; la etapa 2 representa la herencia del padre. En la Figura 2.2a se muestra el árbol para el proceso de dos etapas. Obsérvese que, dado que los alelos se heredan aleatoriamente,

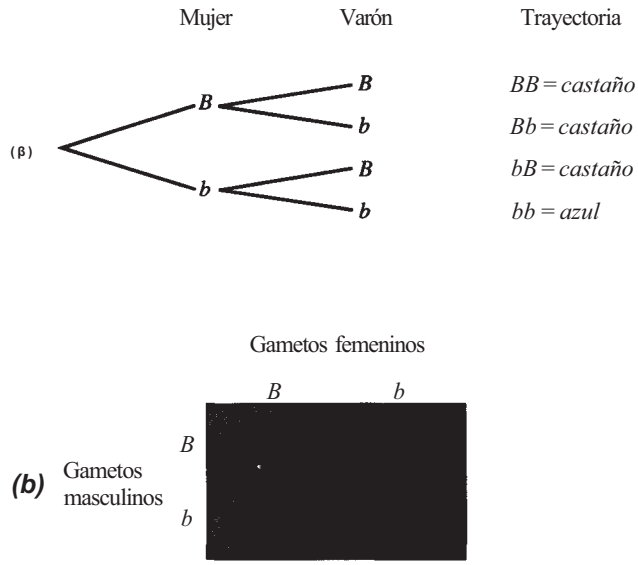


Figura 2.2. (a) Diagrama de árbol para la herencia del color de los ojos de una pareja, cada uno de los cuales es heterocigoto para el color de los ojos. *(b)* Representación biológica del problema en forma de cuadro de Punnett.

en cada paso tenemos tantas posibilidades de heredar un alelo *B* como uno *b.* Cada una de las cuatro trayectorias a lo largo del árbol es equiprobable. Puesto que tres de las cuatro trayectorias dan como resultado un niño de ojos castaños, podemos utilizar el cálculo de probabilidades clásico para concluir que

$$P[\text{niño de ojos castaños}] = \frac{3}{4}$$

Quizá haya visto este problema resuelto en un texto de biología utilizando lo que se denomina un cuadro de Punnett o tablero de ajedrez. En la Figura *2.2b* se muestra el tablero de ajedrez para este problema. Obsérvese que da la misma información que el árbol.

En el último ejemplo, un alelo, el de los ojos castaños, domina sobre el otro. De ahí que la presencia del alelo *B* dé como resultado un individuo de ojos castaños. A veces, existen dos alelos pero ninguno de ellos domina sobre el otro. Nuestro siguiente ejemplo ilustra una situación de este tipo.

Ejemplo 2.2.4. La planta conocida como dondiego de noche puede tener ñores rojas, blancas o rosas. El alelo para el color rojo se representa con *R* y el del blanco con *r.* Una flor roja tiene dos alelos *R* y se dice que es homocigota para el color; una flor blanca es homocigota con el genotipo *rr.* Cuando se cruzan plantas blancas puras con rojas puras, la flor resultante tiene el genotipo *Rr.* Dado que no existe un alelo dominante, la flor resultante es rosa. Cuando se cultivan dos de estas plantas heterocigotas, se obtiene el resultado del árbol de la Figura 2.3. Cada una de las cuatro trayectorias a lo largo del árbol el equiprobable. Utilizando el cálculo clásico de probabilidades podemos concluir que la probabilidad de obtener una flor blanca del cruce es $\frac{1}{4}$.

En una determinación genética pueden utilizarse los árboles para estudiar más de una característica simultáneamente. Para ello, simplemente ampliamos la idea desarrollada en los últimos dos ejemplos.

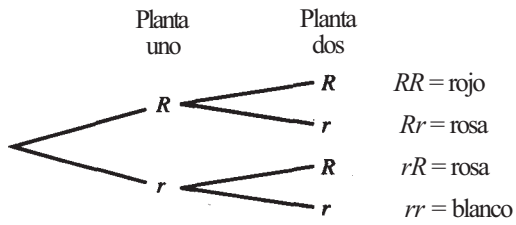


Figura 2.3. Resultados que se obtienen cuando se cruzan dos flores dondiego de noche heterocigotas.

Ejemplo 2.2.5. En los seres humanos, el alelo para la pigmentación normal de la piel *S* es dominante sobre el del albinismo *s.* El alelo para los lóbulos de las orejas separados *F* es dominante sobre el de los lóbulos pegados/. Una mujer tiene el genotipo *SsFF,* y su marido tiene el genotipo *ssFf.* De ahí que la mujer tenga una pigmentación normal de la piel y los lóbulos de las orejas separados; su marido es albino con los lóbulos de las orejas separados. ¿Cuáles son los resultados posibles para su descendencia? Esto lo podemos ver como un experimento de cuatro etapas, siendo éstas las siguientes:

1. Hereda un alelo de la madre para la pigmentación de la piel.
2. Hereda un alelo del padre para la pigmentación de la piel.
3. Hereda un alelo de la madre para la forma de la oreja.
4. Hereda un alelo del padre para la forma de la oreja.

En la Figura 2.4 se muestra el árbol resultante. En el árbol vemos que hay cuatro resultados equiprobables, Puede utilizarse el árbol para ver que la probabilidad de que el niño sea albino es $\frac{1}{2}$; la probabilidad de que tenga los lóbulos de las orejas separados es 1; y la probabilidad de que tenga piel normal y los lóbulos de las orejas separados es $\frac{1}{2}$.

Obsérvese que puede utilizarse el cálculo clásico de probabilidades para resolver estos problemas dado que en cada paso de la descendencia de la pareja es tan probable heredar un gen como el otro de cada uno de los padres. Esto garantiza que las trayectorias a lo largo del árbol son equiprobables. En el Capítulo 3 consideraremos árboles en los cuales esto no es así.

EJERCICIOS 2.2

1. Una familia tiene cuatro hijos. Identificando a cada uno de ellos solamente por su sexo, utilizar un diagrama de árbol para determinar los 16 tipos de nacimientos posibles de los

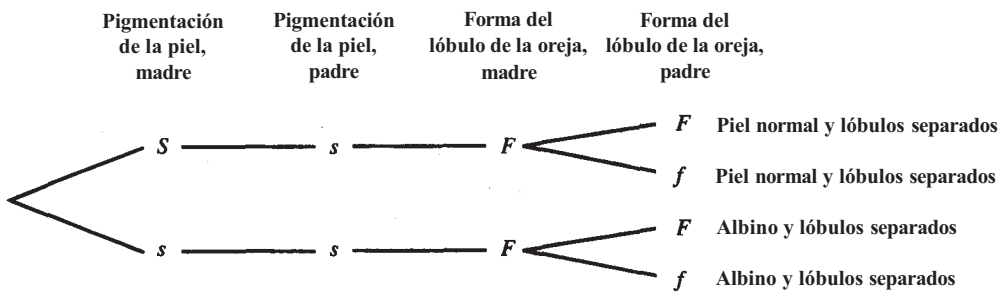


Figura 2.4. Un árbol de cuatro etapas utilizado para estudiar dos rasgos simultáneamente.

hijos. Supongamos que cada hijo tenga exactamente la misma posibilidad de ser niño que de ser niña. Determinar la probabilidad de cada uno de los sucesos siguientes:

A: el primer hijo es un niño.
B: exactamente dos de los cuatro son niños.
C: el mayor y el más pequeño son niños.
D: dos son niñas y dos son niños.

2. Un tetrapéptido bioactivo (un compuesto formado por cuatro aminoácidos ligados en cadena) tiene la siguiente dotación de aminoácidos: alanina (A), ácido glutámico (G), lisina (L) e histidina (H). Por ejemplo, ALGH y LGHA son cadenas típicas de cuatro aminoácidos.
   a) Diseñar un diagrama de árbol para representar las veinticuatro formas posibles en las que estos cuatro aminoácidos pueden ligarse para formar una cadena de cuatro aminoácidos.
   b) Si cada cadena es equiprobable, determinar la probabilidad del suceso *A:* se encuentea ácido glutámico en uno u otro extremo de la cadena.
   c) Determinar la probabilidad del suceso *B:* no se encuentra lisina en ningún extremo de la cadena.
3. Se planifica un experimento para estudiar el efecto de tres tipos de fertilizantes en el crecimiento del trigo. Se prepara una parcela de tierra y se divide en tres franjas de igual tamaño. Se aplica un fertilizante en cada una de las franjas. Denominamos a los fertilizantes *A, B y C.*
   a) Diseñar un diagrama de árbol para representar las seis formas en las que pueden asignarse los fertilizantes a las franjas.
   b) Si la asignación se ha realizado aleatoriamente, de manera que cada trayectoria del diagrama de árbol es equiprobable, ¿cuál es la probabilidad de que la primera franja reciba el fertilizante A?
4. El gato montés que vive en tierras de pastoreo públicas puede ser una amenaza para el ganado vacuno y para las ovejas. Es interesante estimar el número de gatos monteses que viven en una zona concreta. Diez son capturados, marcados y liberados. Más tarde, se capturan cuatro gatos monteses y se clasifica a cada uno de ellos según haya sido marcado (*t*) o no haya sido marcado *(u).* Así, un resultado típico del experimento es *tuut* en el cual el primero y el último animal están marcados mientras que el segundo y el tercero no.
   a) Dibujar un diagrama de árbol para representar los 16 resultados posibles de este experimento.
   b) Indicar las trayectorias que correspondan al acontecimiento del suceso *A:* el primer y el último animal capturados están marcados.
   c) Dar las trayectorias que correspondan al suceso *B:* están marcados exactamente tres animales.
   d) Indicar las trayectorias que correspondan al acontecimiento simultáneo de los sucesos *A y B.*
   e) Si, de hecho, hay 100 gatos monteses viviendo en la región, ¿son equiprobables las 16 trayectorias a lo largo del árbol? Si no es así, ¿qué trayectoria es más probable que se produzca? ¿Cuál es la menos probable que se produzca?
5. Al examinar a un paciente, un médico observa la presencia *(p)* o ausencia *(a)* de cuatro síntomas: dolor de cabeza, fiebre, erupción cutánea o dolor muscular.
   a) Construir un diagrama de árbol para representar las 16 combinaciones de síntomas posibles.
   b) Para diagnosticar alergia sistémica a los alimentos, deberá existir erupción cutánea. Enumerar las trayectorias para las cuales es posible este diagnóstico.

c) Para diagnosticar gripe, el paciente debe tener fiebre y dolor muscular. Decir las trayectorias para las que es posible este diagnóstico.

d) ¿Algunas trayectorias posibilitan el diagnóstico tanto de gripe como de alergia a los alimentos? Si es así, ¿cuáles?

e) Enumerar las trayectorias en las que no es posible el diagnóstico de gripe ni el de alergia a los alimentos.

6. En una zona concreta de Smoky Mountains, muchos árboles de hoja perenne han sido atacados por el escarabajo del pino. Cuatro árboles fueron seleccionados aleatoriamente para su estudio. Se supone que cada uno tiene una probabilidad de ser atacado del 50%.

   a) Dibujar un diagrama de árbol que represente los dieciséis resultados que pueden producirse. (Simbolice con *s* el hecho de que el árbol está infectado y con *n* que no lo está.)

   b) Utilizar el diagrama para calcular la probabilidad de que ningún árbol esté infectado; y la de que al menos uno lo esté.

   c) ¿Consideraría raro que tres de los cuatro árboles estuvieran infectados? Explíquelo.

7. Una forma de determinar el progreso del tratamiento de un enfermo de SIDA es realizando un recuento de linfocitos T4. Un enfermo de SIDA es analizado tres veces. Cada vez el recuento celular es marcado como *n* (normal) o *b* (bajo).

   a) Dibujar un diagrama de árbol que represente las ocho formas en las que esta serie de tres tests puede producirse.

   b) ¿Bajo qué condiciones las ocho trayectorias podrían ser igualmente probables? Hasta donde se conoce sobre el SIDA, ¿cree probable que esto pueda ocurrir?

8. La anemia drepanocítica es una enfermedad mortal en la que los hematíes de la sangre tienden a adoptar forma de hoz en el interior de los vasos sanguíneos. Esto da como resultado el taponamiento de los vasos capilares lo que, eventualmente, conduce a la muerte. Representemos el alelo responsable de la formación de hematíes normales en la sangre como 5 y el que conduce a la formación de hematíes falciformes como *s*. Una persona tendrá anemia drepanocítica si, y sólo si, su genotipo es *ss*. Para cada uno de los casos siguientes construir un diagrama de árbol para representar el (los) resultado(s) posible(s) para un niño hijo de los padres descritos. Utilizar el diagrama de árbol para hallar la probabilidad de que el niño tenga anemia drepanocítica.

   a) La madre es *Ss y* el padre es *Ss*.

   b) La madre es *SS y* el padre es *Ss*.

9. Consideremos el experimento del Ejemplo 2.2.5. Supongamos que cada uno de los padres son del genotipo *SsFf*.

   a) Dibujar un diagrama de árbol con el propósito de representar los resultados posibles de su descendencia.

   b) Utilizar el diagrama de árbol para hallar la probabilidad de que

      i. El hijo sea albino con los lóbulos de las orejas pegados.

      ii. El hijo tenga la piel normal y los lóbulos de las orejas separados.

      iii. El hijo sea albino.

      iv. El hijo tenga los lóbulos de las orejas separados.

10. En los conejillos de indias, el pelo corto (L) es dominante sobre el pelo largo (*l*) y el pelo negro (B) es dominante sobre el pelo albino *(b)*. Una hembra negra de pelo corto es apareada con un macho albino de pelo largo.

    (a) ¿Cuáles son los genotipos posibles para la hembra? ¿Cuáles son los genotipos posibles para el macho?

    b) Para cada genotipo posible de la hembra, construir un diagrama de árbol para representar los resultados posibles de la descendencia.

c) Hallar la probabilidad de obtener un albino de pelo corto en cada caso.

11. En la planta del guisante, la altura elevada (7) es dominante sobre la altura baja *(t).* Las semillas amarillas *(Y)* son dominantes sobre las verdes (y), y la forma redonda (W) es dominante sobre la arrugada (w). Supongamos que se cruzan dos plantas. Una tiene el genotipo *TTYYWw* y la otra *TtYyWw.*
    α) Describir cada una de las plantas madre con relación a las tres características antes mencionadas.
    b) Construir un diagrama de árbol para representar las maneras posibles en las que puede producirse el cruce. *(Indicación:* ampliar la idea del Ejemplo 2.2.5 a un proceso de seis etapas.)
    c) Describir la planta relacionada con cada trayectoria a lo largo del árbol.
    d) Utilizar el diagrama de árbol para hallar la probabilidad de que el cruce dé como resultado una planta alta.
    e) ¿Cuál es la probabilidad de que el cruce dé como resultado una planta alta, de semilla amarilla y arrugada?
    f) ¿Cuál es la probabilidad de que el cruce produzca un guisante verde?
12. Los melocotoneros dan frutos con pelusa y las nectarinas los dan suaves. El alelo para la pelusa es dominante. Cada tipo de fruto puede ser amarillo o blanco, siendo dominante el amarillo. Se cruza un melocotonero blanco con una nectarina amarilla.
    a) ¿Cuáles son los genotipos posibles para el melocotonero?
    b) ¿Cuáles son los genotipos posibles para la nectarina?
    c) Hay cuatro formas posibles de emparejar los genotipos de ambos árboles. Dibujar diagramas de árbol para cada uno de ellos.
    d) Utilizar los árboles del apartado c para hallar la probabilidad de obtener un melocotonero blanco en cada caso.
13. El cuerpo de la mosca de la fruta puede ser de dos colores: gris (*E*) y ébano (e), siendo dominante el gris. También se observan dos tipos de alas, normal (V) y corta o vestigial (v), siendo dominantes las normales. Las moscas homocigotas (*EEVV y eew)* se aparean para obtener descendientes doblemente heterocigotos *(EeVv).* Estos a su vez también los apareamos.
    a) Diseñar un diagrama de árbol para demostrar los resultados posibles de los descendientes.
    b) ¿Cuántas trayectorias darán como resultado una mosca de la fruta color ébano?
    c) ¿Cuántas trayectorias darán como resultado una mosca de la fruta con alas normales ?
    d) Si se selecciona aleatoriamente una mosca entre un gran número de moscas producidas por medio del experimento anterior, ¿cuál es la probabilidad de que tenga alas normales? ¿Cuál es la probabilidad de que sea de color ébano y tenga las alas cortas? ¿Cuál es la probabilidad de que *no* sea de color ébano y tenga las alas cortas?
14. Los alelos que determinan el albinismo se indican como *A y a,* siendo *A* dominante y el responsable de que se produzca un individuo normal con respecto a esta característica. Para ser albino, un individuo ha de recibir el gen recesivo *a* de los dos progenitores. Los individuos *Aa* son normales, pero pueden transmitir el carácter a sus descendientes; se les llama *portadores.*
    a) De una pareja de individuos normales nace un hijo albino. ¿Cuál es el genotipo de cada uno de los progenitores?
    b) ¿Cuál es la probabilidad de que el siguiente hijo nacido de la pareja sea también albino? ¿Cuál es la probabilidad de que sea portador de albinismo?
    c) Una mujer tiene madre normal y padre albino. Su abuela materna también es albina. ¿Cuál es la probabilidad de que la mujer sea portadora de albinismo?

2.3. PERMUTACIONES Y COMBINACIONES (OPCIONAL)

Como se ha indicado en la Sección 2.1, hay varias formas de determinar la probabilidad de un suceso físico. La aproximación clasica, cuando es aplicable, tiene la ventaja de ser exacta. Recordemos que para poder aplicar el método clásico se requieren experimentos en los que los posibles resultados físicos son equiprobables. En este caso, la probabilidad de que se dé un determinado suceso A viene dada por:

$$P[A] = \frac{n(A)}{n(S)}$$

Para calcular una probabilidad utilizando la aproximación clásica tiene que ser posible contar dos cosas: $n(A)$, número de veces que el suceso A puede darse, y $n(S)$, número de resultados que puede dar el experimento. Cuando el experimento es más bien simple, puede optarse por enumerar los resultados o por construir un diagrama de árbol. No obstante, si el experimento es más complejo, estos métodos son incómodos y requieren demasiado tiempo. Habrá que desarrollar métodos alternativos para contar. En el resto de este capítulo se presentan brevemente el cálculo y el concepto clásico de probabilidad. Estos métodos son aplicables a numerosos problemas de las ciencias biológicas y en ellos subyacen muchas de las teorías elementales de la genética.

Por medio de la aproximación clásica, pueden resolverse dos tipos de problemas notablemente distintos, a saber, los que implican permutaciones y los que implican combinaciones. Antes de considerar cómo manejarlos matemáticamente, es necesario poder distinguir el uno del otro.

Definición 2.3.1. Permutación. *Una permutación* es una distribución de objetos en un orden determinado.

Definición 2.3.2. Combinación. Una *combinación* es una selección de objetos con independencia de su ordenamiento.

De las Definiciones 2.3.1 y 2.3.2 resulta evidente que la característica que distingue a una permutación de una combinación es el *orden.* Si el orden en que se realiza una cierta acción es importante, entonces es un problema de permutaciones, y puede resolverse por medio del principio de multiplicación expuesto en la Sección 2.4. Si el orden es irrelevante entonces es un problema de «combinación», lo que implica el uso de la fórmula para las combinaciones desarrollada en la Sección 2.6.

Ejemplo 2.3.1.

a) Escriba su número de afiliado a la Seguridad Social. ¿Es una permutación o una combinación? Es, obviamente, una permutación. 239-62-5558 *no* es el mismo número que 329-62-5558. El orden en que los dígitos están escritos es importante.

b) Hay veinte aminoácidos diferentes de aparición frecuente en polipéptidos y proteínas. Un polipéptido compuesto por cinco aminoácidos

alanina-valina-glicina-cisteína-triptófano

tiene propiedades diferentes y es, de hecho, un compuesto distinto que el polipéptido

alanina-glicina-valina-cisteína-triptófano

que contiene los mismos aminoácidos. Los polipéptidos, ¿son permutaciones o combinaciones de unidades de aminoácidos? Son permutaciones porque es importante 1a secuencialidad, u orden, de los aminoácidos en la cadena.

c) Un biólogo dispone de 10 plantas para un experimento. Sólo ocho son necesarias para realizarlo. Las ocho plantas necesarias son seleccionadas aleatoriamente. ¿Representa esta colección de plantas una permutación o una combinación? Es una combinación dado que el punto de interés radica solamente en las ocho plantas seleccionadas y no en el orden en que han sido escogidas.

El Ejemplo 2.3.1 tiene como meta proporcionar una noción intuitiva de la diferencia entre permutaciones y combinaciones. Su importancia en el análisis de datos científicos no ha sido más que esquematizada.

EJERCICIOS 2.3

1. Escriba su nombre. Indique si esa cadena de letras constituye una permutación de letras o combinación de letras.
2. Considérense todas las familias que tienen cinco hijos. Atendiendo al sexo de cada uno de ellos, hay 32 posibles ordenaciones. Por ejemplo, *VVVHH* representa una situación en la que los tres primeros hijos nacidos son varones y los dos últimos hembras. Dar un ejemplo de alguna otra ordenación con tres varones y dos hembras. Las cadenas del cinco letras que representan el orden de los nacimientos, ¿son permutaciones o combinaciones de letras?
3. Un científico tiene seis jaulas diferentes de ratas blancas en el animalario y desea seleccionar tres de las seis para utilizarlas en un experimento. El grupo de animales elegidos, ¿es una permutación o una combinación de animales?
4. Ocho pacientes, alérgicos a las picaduras de insectos, están siendo insensibilizados. Se seleccionan cuatro aleatoriamente y se les trata con un compuesto de veneno activo de insectos. Con fines comparativos, los otros cuatro son tratados con un compuesto formado por insectos muertos. El conjunto de pacientes elegidos para recibir el veneno de insectos, ¿es una permutación o una combinación de pacientes?
5. En el curso de unas pruebas de tolerancia frente a la glucosa, se mide y registra el nivel de azúcar en sangre de un paciente cada media hora durante dos horas y media. En el registro se utilizan las anotaciones «normal» (n), «alto» *(a)* o «bajo» *(b)*. Considérese la cadena *nααnb*. Explicar la forma en que se ha desarrollado el test para este paciente. Dar un ejemplo de alguna otra cadena posible con dos anotaciones normales, dos altas y una baja. ¿Estas cadenas de letras son permutaciones o combinaciones de letras?
6. Un químico toma diez muestras de agua del depósito de una factoría de papel. Se seleccionan tres aleatoriamente para verificar el grado de acidez. ¿Representa este conjunto de tres muestras una permutación de muestras o una combinación de muestras?
7. Dos de las trayectorias del diagrama de árbol del Ejercicio 4 de la Sección 2.2 son *tutu* y *tuut*. Cada una de ellas es una permutación con dos *t* y dos *u*. Explicar la diferencia entre ambas, dentro del contexto del experimento descrito en el Ejercicio 4.
8. Al abrir una caja fuerte, marcamos la «combinación» de la misma. ¿Es la secuencia de números o letras marcadas una combinación en el sentido matemático?
9. Un guardabosques selecciona cuatro cornejos rosas y cuatro blancos entre los árboles disponibles en la tienda de un proveedor mayorista. ¿Constituyen los árboles una combinación o una permutación de árboles rosas y blancos? Los árboles deben plantarse en hilera a lo largo de la carretera de entrada de un parque nacional. Imagine dos formas de plantar los árboles de manera que el primero y el último de cada hilera sean

rosas. ¿Forma esta hilera de ocho árboles una combinación o una permutación de colores?

10. Un guardabosques desea estimar el número de especies de árboles dentro de un gran bosque forestal. Obtiene un mapa de la zona donde están marcados y numerados cuadrados de 10 x 10 metros. Se escogen aleatoriamente 15 de estos cuadrados, y se determina el número de especies por cuadrado mediante una inspección. ¿Representan los 15 cuadrados una combinación de cuadrados o una permutación de cuadrados?

2.4. PRINCIPIO DE MULTIPLICACIÓN (OPCIONAL)

Una vez identificada la importancia del orden en un problema dado, la cuestión siguiente a dilucidar es: ¿cuántas permutaciones de los objetos dados son posibles? La respuesta viene generalmente dada por el *principio de multiplicación.*

Para aplicar el principio en primer lugar debemos preguntarnos: ¿de cuántas etapas o pasos consta el experimento considerándolo como un todo? O ¿cuántas decisiones debo tomar en el curso del experimento? Se dibuja una raya para representar cada etapa o decisión. En cada etapa preguntamos: ¿de cuántas formas puede realizarse este paso en el experimento? Estos números se colocan en las rayas. El número total de maneras para realizar el experimento completo es el producto de los valores de las rayas.

Este principio es más difícil de describir que de utilizar, y mucha gente lo aplica intuitivamente aun cuando haya tenido una cierta educación matemática. Ilustraremos su empleo en el Ejemplo 2.4.1.

Ejemplo 2.4.1. Los biólogos están interesados en el orden en que los cuatro ribonucleótidos adenina (A), uracilo (U), guanina (G) y citosina (C) se combinan para formar cadenas pequeñas. Estos nucleótidos constituyen las subunidades principales de RNA, molécula intermediaria portadora de la información que actúa en la traducción del código genético del DNA. ¿Cuántas cadenas formadas por dos nucleótidos *diferentes* pueden formarse? La cuestión puede resolverse muy fácilmente por medio del diagrama de árbol de la Figura 2.5.

La solución es evidentemente 12. Obsérvese que estamos considerando que la cadena AC es distinta de la CA. Es decir, que el orden en que se disponen los nucleótidos es importante. Hemos demostrado, por tanto, que hay 12 permutaciones de cuatro elementos distintos tomados de dos en dos. Este resultado puede predecirse sin necesidad de recurrir al diagrama, simplemente contestando a tres preguntas.

1. ¿De cuántas etapas o pasos consta el experimento como un todo? Respuesta: dos. La primera etapa corresponde a la colocación del primer nucleótido, la segunda al segundo, que ha de ser *diferente* al primero. Representamos el hecho de que consta de dos etapas dibujando dos rayas:

_____	_____
primer nucleótido	segundo nucleótido

2. ¿De cuántas formas puede realizarse la primera fase del experimento? Respuesta: cuatro. Hay cuatro nucleótidos disponibles, cualquiera de los cuales puede ser colocado en primera posición. Lo indicaremos poniendo un 4 sobre la primera raya:

4	
_____	_____
primer nucleótido	segundo nucleótido

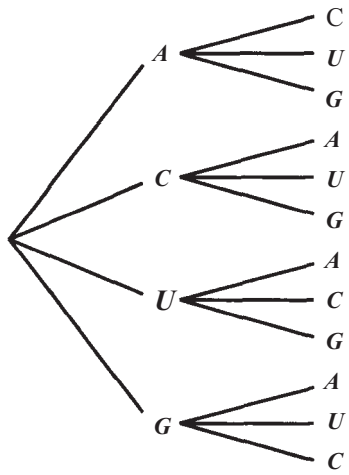


Figura 2.5. Cadenas de dos nucleótidos.

3. Tras haber completado la primera fase, ¿de cuántas maneras puede llevarse a cabo la segunda fase? La respuesta es tres. Puesto que cada cadena está formada por dos nucleótidos diferentes, no se permite la repetición. El nucleótido que aparece en el primer lugar de la cadena queda fuera de juego. El segundo elemento de la cadena puede ser cualquiera de los restantes nucleótidos. Lo indicaremos poniendo un 3 sobre la segunda raya.

$$\underset{\text{primer nucleótido}}{\underline{\overset{4}{}}} \qquad \underset{\text{segundo nucleótido}}{\underline{\overset{3}{}}}$$

El principio de multiplicación dice que, para determinar el número total de permutaciones posibles, solamente es necesario multiplicar estos dos números; se obtiene pues, nuevamente, 12 como solución. Obsérvese que el 4 de la primera raya corresponde directamente al primer punto de ramificación del diagrama en árbol y el 3 al segundo.

El principio de multiplicación puede utilizarse para resolver problemas que surgen en la vida diaria. El Ejemplo 2.4.2 lo demuestra.

Ejemplo 2.4.2

a) ¿Cuántas palabras clave de acceso al ordenador pueden formarse con cinco letras diferentes? Este es un proceso de cinco pasos. Existen 26 elecciones para la primera letra de la palabra clave. Dado que las letras sólo pueden utilizarse una vez, el número de elecciones desciende de 1 en 1 en las etapas sucesivas. Por lo tanto, el número total de palabras clave es

$$\underline{26} \cdot \underline{25} \cdot \underline{24} \cdot \underline{23} \cdot \underline{22} = 7\,893\,600$$

b) ¿Cuántas marcas de licencia pueden formarse con tres letras diferentes seguidas de tres dígitos diferentes? Este es un proceso de seis etapas. En primer lugar, elegimos tres letras diferentes en sucesión entre las 26 letras del alfabeto.

$$\underset{\text{Letras}}{\underline{26} \cdot \underline{25} \cdot \underline{24}} \qquad \underset{\text{Dígitos}}{\underline{}\ \underline{}\ \underline{}}$$

A continuación, elegimos tres dígitos diferentes en sucesión entre los dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Todo junto sería

$$\underset{\text{Letras}}{26 \cdot 25 \cdot 24} \qquad \underset{\text{Dígitos}}{10 \cdot 9 \cdot 8} = 11232\,000$$

posibles marcas que cubren las condiciones establecidas,

c) ¿De cuántas formas puede responderse a una prueba de cinco preguntas de tipo verdadero-falso? Este es un proceso de cinco pasos. Puesto que cada pregunta puede responderse como verdadera o falsa, existen dos posibilidades en cada etapa de la prueba. Existen

$$2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 32$$

formas de responder a la prueba.

d) ¿Cuántos números de teléfono de 10 dígitos pueden formarse si el código de zona de cada uno de ellos es 703 y el código local no puede contener ni 0 ni 1? Este es un proceso de 10 pasos. Puesto que el código de zona debe ser el 703, sólo tenemos una elección para cada una de las tres primeras etapas. La primera debe ser 7, la segunda 0 y la tercera 3.

$$\underset{\text{Código de zona}}{1 \quad 1 \quad 1} \qquad \underset{\text{Código local}}{\_ \quad \_ \quad \_} \qquad \underset{\text{Número}}{\_ \quad \_ \quad \_ \quad \_}$$

Puesto que el código local no puede ser ni 0 ni 1, el cuarto paso tiene ocho posibilidades.

$$\underset{\text{Código de zona}}{1 \quad 1 \quad 1} \qquad \underset{\text{Código local}}{8 \quad \_ \quad \_} \qquad \underset{\text{Número}}{\_ \quad \_ \quad \_ \quad \_}$$

No existen restricciones especiales en las otras posiciones, por lo que hay 10 elecciones en cada caso. El número total de posibilidades es

$$\underset{\text{Código de zona}}{1 \quad 1 \quad 1} \qquad \underset{\text{Código local}}{8 \cdot 10 \quad 10} \qquad \underset{\text{Número}}{10 \quad 10 \quad 10 \quad 10} = 8\,000\,000$$

Al aplicar el principio de multiplicación, debemos tener en cuenta algunos detalles que se exponen a continuación.

Directrices para la aplicación del principio de multiplicación

1. Posibilidad de repetición o no repetición. Unas veces los objetos pueden repetirse (tal como los dígitos en un número de teléfono); y otras veces no (como sucede en los Ejemplos *2.4.2a y 2A.2V).* Generalmente, el contexto físico del problema permite o excluye la posibilidad de repetición.
2. Posibilidad de recurrir a la sustracción. Consideremos un suceso *A.* En ocasiones, será difícil, si no imposible, calcular *n(A)* directamente. No obstante, podríamos ser capaces de hallar el total *n(S)* y el número de veces en que *A* no se produce de manera sencilla. Representemos esto último con *n(A')* [la prima (') se lee «no»] Puesto que el suceso *A* puede ocurrir o no, el total es la suma de los números *n(A)* y *n(A').* Es decir, $n(S) = n(A) + n(A')$. Ello implica que $n(A) = n(S) - n(A')$.

3. Si hay alguna etapa del experimento que se encuentre sometida a alguna restricción especial, conviene ocuparse primero de la restricción (como en el Ejemplo *2.4d).*

Estos puntos aparecen ilustrados en el Ejemplo 2.4.3.

Ejemplo 2.4.3. El código DNA-RNA es un código molecular en el que el orden secuencial de las moléculas proporciona una información genética importante. Cada segmento del RNA está compuesto por «palabras». Cada palabra es específica para un determinado aminoácido, y está compuesta por una cadena de tres ribonucleótidos que no son necesariamente distintos. Por ejemplo, la palabra UUU corresponde al aminoácido fenilalanina, mientras que AUG identifica la metionina.

a) Consideremos el experimento consistente en formar una palabra de RNA. ¿Cuántas palabras pueden formarse? Esto es, ¿cuánto vale n(S)? Cada uno de los tres ribonucleótidos de la cadena es uno de los cuatro mencionados en el Ejemplo 2.4.1, es decir, adenina (A), uracilo (U), guanina (G) y citosina (C). Obsérvese que los podemos repetir. Se trata, pues, de un experimento en tres etapas con cuatro posibilidades por etapa. Por el principio de multiplicación, $n(S) = 4 \cdot 4 \cdot 4 = 64$.

b) ¿Cuántas de las palabras del apartado α tienen al menos dos nucleótidos idénticos? La cuestión es fácilmente abordable por sustracción. Sea R el suceso «la palabra contiene nucleótidos repetidos». El suceso R' es el suceso donde no hay repetición de nucleótidos. Consideremos el diagrama de la Figura 2.6. Sabemos ya que n(S) = 64. Por el principio de multiplicación, $n(R') = 4 \cdot 3 \cdot 2 = 24$. Así que, por sustracción, ¿el número de palabras con alguna repetición es:

$$n(R) = n(S) - n(R') = 64 - 24 = 40$$

c) Si se forma una palabra aleatoriamente, ¿cuál es la probabilidad de que contenga alguna repetición de nucleótidos? Usando la probabilidad clásica, tenemos:

$$P[R] = \frac{n(R)}{n(S)} = \frac{40}{64} = 0.625$$

d) Consideremos el suceso B compuesto por palabras formadas aleatoriamente que terminen en U (uracilo) y que no contengan repeticiones. Hallar $P[B]$. Puesto que la última posición en la palabra ha de estar ocupada por el uracilo, solamente hay una opción para dicho lugar, tal como se indica:

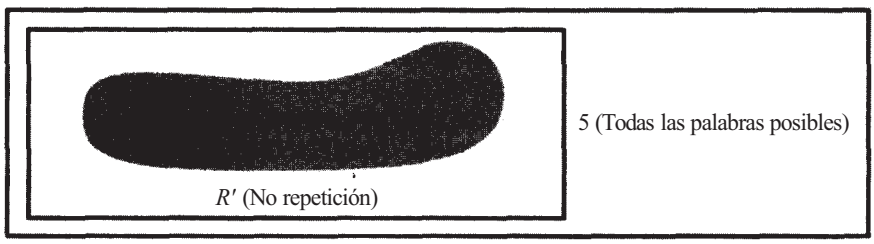


Figura 2.6. Puesto que cada palabra tiene (R) o no tiene (R') repetición, $n(S) = n(R) + n(R')$. Por lo que n(R) = $n(S)$ - n(R').

Una vez que nos hemos ocupado de la restricción, nos fijaremos en que las repeticiones no están permitidas. Así que la primera posición puede ser ocupada por cualquiera de los tres nucleótidos restantes, y la segunda por cualquiera de los otros dos.

$$\underline{\quad 3 \quad} \quad \underline{\quad 2 \quad} \quad \underline{\quad 1 \quad}$$

Por tanto, $n(B) = 6$ y

$$P[B] = \frac{n(B)}{n(S)} = \frac{6}{64}$$

No se ha dado una fórmula para resolver problemas de permutaciones porque muchos de ellos en los que importa el orden son demasiado complejos para que una simple fórmula sea apropiada. Más bien se ha propuesto que, cuando el orden importe, debe pensarse en el «principio de multiplicación». En algunos casos, no obstante, se puede aplicar una fórmula para las permutaciones. Esta fórmula se estudia en el Ejercicio 15.

EJERCICIOS 2.4

1. Una antena parabólica para satélites puede recibir señales de 20 satélites de comunicación diferentes. Cada satélite emite a 15 estaciones. ¿Cuántas estaciones están disponibles para el propietario de esta antena?
2. Un médico tiene cuatro pacientes en la sala de espera. Uno de los casos es una urgencia, pero el paciente no se lo dice al médico.
   a) ¿De cuántas maneras distintas pueden atenderse estos pacientes?
   b) ¿De cuántas de estas maneras se atenderá primero al de la urgencia?
   c) Si el orden se selecciona aleatoriamente, ¿cuál es la probabilidad de que el caso de urgencia sea atendido en primer lugar?
3. ¿Cuántas palabras clave de acceso al ordenador pueden formarse con cinco letras diferentes si cada palabra debe empezar con una vocal (A, E, I, O, U) y terminar con una consonante?
4. Cuatro personas se registran en el mostrador de una línea aérea. El encargado de los pasajes asigna los asientos y a continuación coloca cada pasaje aleatoriamente dentro de una carpeta que tiene escrito el nombre del pasajero. ¿De cuántas formas puede realizarlo? Dibujar un diagrama de árbol para ilustrar el problema. ¿Cuál es la probabilidad de que ninguno de ellos reciba su propio pasaje?
5. a) ¿Cuántas palabras de RNA que comiencen con U (uracilo) y terminen con A (adenina) o G (guanina) pueden formarse? (Recuérdese que pueden utilizarse los cuatro ribonucleótidos —A, U, G y C— y que una palabra está constituida por tres de ellos, no necesariamente diferentes.)
   b) ¿Cuántas de las palabras del apartado *a* no contienen ninguna repetición?
   c) ¿Cuál es la probabilidad de que una palabra formada aleatoriamente empiece con U, termine con A o con G y contenga alguna repetición?
   d) Comprobar las soluciones construyendo el diagrama de árbol correspondiente a *a y b.*
6. De las 64 palabras de RNA posibles, 61 son códigos para los 20 aminoácidos existentes. Las otras tres son códigos de «parada», que causan la terminación del polipéptido. Si formamos aleatoriamente una palabra, ¿cuál es la probabilidad de que se trate del código de algún aminoácido?

7. Una palabra es código para la treonina si, y sólo si, empieza por AC. ¿Cuántos sinónimos (palabras con idéntico significado) existen para la treonina?
8. Considérese el segmento UUUAUUUUA de RNA. Tal segmento contiene tres palabras UUU (fenilalanina), AUU (isoleucina) y UUA (leucina) no solapadas. Obsérvese que si se presenta una mutación y la cadena de nucleótidos cambia a UUAAUUUUA, entonces el segmento no codifica, en forma alguna, la misma secuencia de aminoácidos. Existen dos sinónimos para la fenilalanina, tres para la isoleucina y seis para la leucina. ¿De cuántas formas diferentes, pero equivalentes, puede expresarse la secuencia de los tres aminoácidos citados?
9. Considérese cualquier segmento de tres palabras no solapadas. ¿De cuántas maneras podría expresarse tal segmento? ¿Cuántos de estos segmentos no contienen palabras repetidas? Seleccionando aleatoriamente un segmento, ¿cuál es la probabilidad de que tenga alguna repetición de palabras? ¿Cuál es la probabilidad de que cada palabra sea diferente y de que cada una codifique un aminoácido? (Véase Ejercicio 6.)
10. Un médico puede elegir entre cinco fármacos diferentes, dos de los cuales son experimentales, para tratar a un paciente afectado de hipertensión arterial. También puede optar por uno de entre cuatro programas, de los cuales dos implican el desarrollo de actividades domiciliarias y los otros dos son relativos a actividades externas. Hay tres dietas posibles, una de ellas totalmente desprovista de sal.
    a) ¿Cuántos tratamientos, formados por un fármaco, un programa y una dieta son posibles?
    b) ¿Cuántos tratamientos de entre los del apartado *a* implican el uso de un fármaco experimental?
    c) Dado que todos los tratamientos del apartado *a* tienen la misma posibilidad de ser elegidos, si seleccionamos uno aleatoriamente, ¿cuál es la probabilidad de que implique el uso de un fármaco experimental y un programa de actividades externas?
    d) Si se da el hecho de que un determinado fármaco de entre los de carácter experimental es peligroso cuando se combina con una dieta sin sal, ¿cuál es la probabilidad de que, a pesar de ello, tal tratamiento fuera casualmente prescrito?
11. Se ponen a prueba siete fármacos para descubrir su grado de eficacia contra el acné. Se pide a un investigador que ordene los fármacos de 1 a 7; el fármaco más eficaz recibirá el valor 1.
    a) ¿De cuántas maneras pueden asignarse los siete valores a los siete fármacos?
    b) Los fármacos A y B han sido fabricados por la compañía interesada en la prueba, lo cual es desconocido por el investigador. Si se da el hecho de que éste no pueda establecer distinción alguna entre los productos y realmente asigna a cada fármaco un valor aleatoriamente, ¿cuál es la probabilidad de que esos dos fármacos reciban la valoración más alta?
    c) Si se realiza el experimento y A y B reciben los dos valores más altos, ¿piensa usted que la compañía puede estar segura de que sus productos son más eficaces que los de sus competidores? Razónese sobre la base de la respuesta que usted diera al apartado *b.*
12. Se está elaborando un estudio para investigar el efecto del tipo de polímero, la temperatura, la dosis de radiación, la tasa de la dosis de radiación y el pH en la capacidad para extraer vestigios de benceno del agua. Existen dos tipos de polímeros (A y B), tres temperaturas (alta, media, baja), tres dosis de radiación, tres tasas de las dosis de radiación y tres niveles de pH (ácido, básico, neutro).
    a) ¿Cuántas condiciones experimentales deberán estudiarse?
    b) Si cada condición experimental debe replicarse (repetirse) cinco veces, ¿cuántos ensayos experimentales deberán realizarse?

c) ¿Cuántos ensayos se han de realizar con el polímero A a baja temperatura?
d) ¿Cuántos ensayos se han de realizar con el polímero B a alta o media temperatura y bajo pH?

13. La unidad de almacenamiento básico de un ordenador digital es un *bit.* Un bit es una posición de almacenamiento que puede designarse bien como activada (1), bien como desactivada (0) en cualquier momento dado. En la digitalización de imágenes de manera que puedan transmitirse electrónicamente, se utiliza un elemento de imagen denominado *píxel.* Cada píxel es cuantificado en niveles de gris y codificado utilizando un código binario. Por ejemplo, un píxel con cuatro niveles de gris puede codificarse utilizando 2 bits designando los niveles de gris 00, 01, 10, 11.
   a) ¿Cuántos niveles de gris pueden cuantificarse utilizando un código de 4 bits?
   b) ¿Cuántos bits son necesarios para codificar un píxel cuantificado en 32 niveles de gris?
14. Se elabora un experimento para investigar el efecto de la temperatura del suelo sobre la tasa de germinación de una nueva variedad de semillas de hierba. Sólo tenemos disponible una habitación a temperatura ambiente para realizar el experimento, por lo que sólo puede considerarse la temperatura en un momento dado. Las temperaturas se codifican como A = alta, M = moderada, B = baja ¿En cuántos órdenes pueden ensayarse las temperaturas? Dibujar un diagrama de árbol para verificar la respuesta y enumerar todas las condiciones experimentales posibles.
15. *Permutaciones de* n *objetos distintos tomados de* r *en* r: ${}_nP_r$. Consideremos n objetos distintos y tomemos $r \leqslant n$ de esos objetos para formar un grupo. En el grupo, los objetos no se repiten y pueden ocupar cualquier posición. Representamos por ${}_nP_r$ el número de grupos que se pueden formar. Para obtener una fórmula para ${}_nP_r$ se utiliza el principio de multiplicación. Comenzaremos dibujando r posiciones en la forma siguiente

Posición 1	Posición 2	Posición 3	⋯	Posición r

   a) ¿De cuántas formas puede rellenarse la posición 1 ? ¿Y la posición 2? ¿Y la 3? ¿Y la r?
   b) Por medio del principio de multiplicación, justificar que

$$ {}_nP_r = n(n - 1)(n - 2) \cdots (n - r + 1) $$

   c) Sean $n = 10$ y $r = 4$. Utilizar estos valores para verificar la fórmula del apartado *b.*
   d) Hallar la fórmula para ${}_nP_n$.
   e) Un farmacéutico tiene remedios contra la jaqueca de 10 marcas diferentes. Dispone de sitio en la parte alta de una estantería solamente para tres. ¿De cuántas formas pueden ser colocados esos remedios en la estantería?

2.5. PERMUTACIONES DE OBJETOS INDISTINGUIBLES (OPCIONAL)

Hasta el momento, hemos descrito problemas sencillos en los que puede o no producirse repetición. Consideraremos ahora situaciones en las que la repetición es inevitable. La pregunta a responder es: ¿cuántas ordenaciones distintas de n objetos son posibles, si algunos de ellos son idénticos y además indistinguibles? Para contestar esta pregunta, consideraremos primero una notación simbólica llamada *notación factorial,* que será ampliamente utilizada a lo largo del texto.

Definición 2.5.1. *n* factorial. Sea n un entero positivo. El producto $n(n - 1)\ (n - 2)... 3 .2 . 1$ se llama *n factorial* y se representa por $n!$

Ejemplo 2.5.1. $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$

Definición 2.5.2. Cero factorial

$$0! = 1$$

Volvamos ahora al problema de determinar el número de permutaciones posibles cuando algunos de los objetos son indistinguibles unos de otros

Ejemplo 2.5.2. Consideremos los nucleótidos adenina (A) y uracilo (U) ¿Cuántas palabras —cadenas de tres ribonucleótidos no necesariamente diferentes— pueden formarse usando solamente esos dos símbolos? Puesto que una palabra está constituida por tres símbolos y sólo pueden utilizarse dos símbolos distintos, la repetición es inevitable. Por el principio de multiplicación existen $2 \cdot 2 \cdot 2 = 8$ palabras posibles. Esto no es nada nuevo. No obstante, si preguntamos ahora «¿cuántas de las ocho palabras contienen dos veces uracilo y una vez adenina?»; la pregunta es nueva. Estamos disponiendo un total de $n = 3$ objetos, pero dos son idénticos (dos U) y, por tanto, indistinguibles el uno del otro. Para solucionar la cuestión, escribamos las distintas posibilidades:

AUU UAU UUA

La respuesta es 3. ¿Cómo podríamos haber previsto esto sin tener que hacer una lista? Observamos que

$$3 = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1)(1)} = \frac{3!}{2!\ 1!}$$

El factorial de 3 del numerador representa el hecho de que están siendo permutados un total de tres objetos. Hay dos factoriales en el denominador porque existen dos tipos de objetos implicados, dos Ues y una A.

El caso anterior no es el resultado de una simple coincidencia. Puede demostrarse que se produce en cualquier situación en la que haya que ordenar n objetos siendo algunos de ellos indistinguibles de otros. El Teorema 2.5.1 formaliza esta idea.

Teorema 2.5.1. Permutaciones de objetos indistinguibles. Consideremos n objetos de los cuales n_1 son del tipo 1, n_2 del tipo 2, ..., n_k son del tipo k. El número de formas en que pueden disponerse los n objetos viene dado por

$$\frac{n!}{n_1!\, n_2! \cdots n_k!} \qquad n_1 + n_2 + \ldots + n_k = n$$

Ejemplo 2.5.3. Se utilizan 15 pacientes en un experimento para comparar un fármaco estándar, un fármaco experimental y un placebo. Se asigna aleatoriamente a cada paciente uh tratamiento. ¿De cuántas formas distintas pueden asignarse los tres tratamientos a los 15 pacientes? ¿Cuál es la probabilidad de que, asignando aleatoriamente los tratamientos a los pacientes, salga la alternativa de que 10 pacientes reciben el placebo, 3 el fármaco experimental y 2 el fármaco estándar?

La primera pregunta no es nueva. Hay $3 \cdot 3 \cdot 3 \cdots 3 = 3^{15} = 14\,348\,907$ formas alternativas de asignar el tratamiento a los pacientes. La segunda pregunta sí. Para hallar la probabilidad

que se pide aquí, debemos determinar cuántas de las alternativas posibles incluyen 10 veces el placebo, tres veces el fármaco experimental y dos veces el fármaco estándar. Por medio del Teorema 2.5.1 obtenemos inmediatamente

$$\frac{15!}{10!\ 3!\ 2!} = \frac{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 \cdot \cancel{10!}}{\cancel{10!}(3 \cdot 2 \cdot 1 \cdot 2 \cdot 1)} = 30\,030$$

La probabilidad buscada es, por lo tanto,

$$\frac{30\,030}{14\,348\,907} \cong 0.0021$$

EJERCICIOS 2.5

1. ¿De cuántas formas pueden disponerse las letras de la palabra BOO para formar palabras distintas?
2. Un investigador tiene ocho plantas para experimentar con ellas. Se están investigando dos regímenes de riego diferentes: agua corriente (7) y agua ligeramente ácida *(A),* para simular la lluvia ácida. Cuatro plantas reciben agua corriente y el resto la solución ácida. Una asignación típica de los tratamientos a las plantas es *ATTTAATA* ¿Cuántas asignaciones posibles hay?
3. Un laboratorio médico posee una máquina que se utiliza para contar los leucocitos en la sangre de un paciente. Cada mañana se hace un conteo de prueba, con una muestra de la que se conoce el recuento correcto. Si la máquina da un recuento aceptablemente cercano al recuento real, entonces se supone que la máquina está «bajo control» o funcionando correctamente; de lo contrario, está «fuera de control» y debe ajustarse. Un técnico informa de que en cuatro de los últimos 14 días, la máquina ha estado fuera de control. ¿De cuántas maneras puede haber sucedido?
4. ¿De cuántas formas se pueden recolocar las letras de su apellido para formar diferentes palabras distinguibles?
5. Durante la observación de los biorritmos de los ratones, se aisla un ratón en una caja que contiene una rueda de ejercicios. El ratón se mantiene con luz durante las 24 horas y su actividad se controla durante un período de 24 horas a través de un monitor computadorizado conectado a la rueda. Se informa de que existen 16 secuencias de tiempo reconocibles, con las frecuencias siguientes: R = reposo (8), SA = corto período de actividad (4), MA = período de actividad moderada (2), LA = largo período de actividad (2). A los períodos de actividad siempre les sigue uno de reposo. ¿De cuántas formas puede desarrollarse este experimento? Si la actividad del ratón es aleatoria, de forma que todos los resultados posibles son equiprobables, ¿cuál es la probabilidad de que se produzcan consecutivamente dos períodos de larga actividad?
6. Consideremos el experimento del Ejercicio 9 de la Sección 2.3. Si un árbol se distingue sólo por el color, ¿de cuántas maneras pueden plantarse los ocho cornejos? ¿Cuántas de estas disposiciones tienen al menos dos árboles del mismo color en cada uno de los lados?
7. Se pretende formar una secuencia de RNA de 10 palabras. La secuencia incluye la palabra ACU (treonina) tres veces, la GGU (glicina) dos veces, cuatro a GAA (ácido glutámico) y una la UAA (código de parada). ¿Cuántas secuencias de 10 palabras que presenten esta composición son posibles? ¿Cuál es la probabilidad de que una secuencia de este

grupo aleatoriamente seleccionada, presente el código de parada en algún lugar que no sea al final de la secuencia?

8. Se dispone de quince animales de experimentación que se utilizarán para comparar tres dietas diferentes. Cada dieta será puesta a prueba sobre cinco animales seleccionados aleatoriamente. ¿De cuántas formas diferentes pueden distribuirse las dietas entre los sujetos de experimentación?
9. Probar que ${}_nP_r$ puede expresarse como:

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

2.6. COMBINACIONES (OPCIONAL)

Hasta aquí hemos considerado problemas de recuento en los que el orden, natural o impuesto, era importante. Las palabras *orden* o *disposición* aparecen, generalmente, en el enunciado de estos problemas. Indican el uso del principio de multiplicación o la fórmula para permutación nes de objetos indistinguibles para su solución. Volveremos ahora nuestra atención a situaciones en las que el orden es irrelevante. Es decir, nos ocuparemos de problemas más bien relativos a combinaciones que a permutaciones.

Las palabras clave que identifican una combinación son las palabras *seleccionar, elegir o escoger*. Las combinaciones se resolverán mediante una fórmula adecuada. El modelo utilizado en la fórmula es fácil de deducir. El Ejemplo 2.6.1 ilustra esta idea.

Ejemplo 2.6.1. Cinco personas se ofrecen voluntarias para participar en un programa experimental. Se necesitan solamente dos para llevar a cabo el estudio. ¿De cuántas formas pueden seleccionarse dos personas de entre las cinco?

En este caso el orden no importa. Lo que interesa únicamente es el hecho de que sean dos los seleccionados, no el orden en que lo son. Estamos, pues, preguntando, ¿cuántas combinaciones de cinco elementos tomados dos a dos existen? La cuestión puede resolverse adjudicando una de las letras *A, B, C, D, E* a cada uno de los voluntarios y formando una lista con todos los subconjuntos posibles de tamaño dos, del siguiente modo:

{A, B}	*{A,E}*	*{B, E}*	*{D, E}*
{A, C}	*{B,C}*	*{C,D}*	
{A, D}	*{B,D}*	*{C,E}*	

Obviamente, existen 10 combinaciones. Escribiremos ${}_5C_2 = 10$, donde el 5 indica el número disponible de objetos, el 2 el número de objetos que hay que seleccionar, y *C* las combinadones en cuestión. Existe la alternativa de escribir $\binom{5}{2} = 10$. ¿Cómo podía haberse obtenido el valor 10 sin hacer uso de la lista? Observamos simplemente que

$$ 10 = \frac{5!}{2!\,3!} $$

El numerador de esta expresión es 5! porque disponemos de cinco objetos. En el denominador tenemos dos números, 2! y 3! Ello se debe al hecho de que los cinco objetos han de ser divididos en dos grupos, uno de los cuales constituye el de los que se seleccionan para el estudio (2) y el otro el de los que se omiten (5—2 = 3).

El modelo descrito no es resultado de una coincidencia y se generaliza por medio del Teorema 2.6.1.

Teorema 2.6.1. El número de combinaciones de n objetos diferentes, tomados en grupos de r objetos, simbolizado por ${}_nC_r$ o bien $\binom{n}{r}$, viene dado por:

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} $$

El Ejemplo 2.6.2 ilustra el uso de la fórmula de las combinaciones en la solución de problemas de probabilidades.

Ejemplo 2.6.2. Un banco de sangre dispone de 10 unidades de sangre tipo A^*. De ellas, cuatro están contaminadas con suero de hepatitis. Se seleccionan aleatoriamente tres unidades de entre las 10 para utilizarlas con tres pacientes diferentes. ¿Cuál es la probabilidad de que un solo paciente esté expuesto a contraer la hepatitis por esta causa?

Esta pregunta se refiere a las combinaciones, ya que sólo estamos interesados en las unidades seleccionadas, no en el orden en que se seleccionan. Consideremos el diagrama de la Figura 2.7. El número total de formas de seleccionar tres unidades de entre las 10 disponibles es

$$ {}_{10}C_3 = \frac{10!}{3!\,7!} = 120 $$

Para que un solo paciente esté expuesto a contraer la hepatitis por esta causa, la unidad seleccionada lo ha de ser de entre las cuatro contaminadas. La unidad contaminada puede seleccionarse de

$$ {}_4C_1 = \binom{4}{1} = \frac{4!}{1!\,3!} = 4 \text{ formas} $$

Las unidades no contaminadas pueden ser seleccionadas de

$$ {}_6C_2 = \binom{6}{2} = \frac{6!}{2!\,4!} = 15 \text{ formas} $$

En total hay (4)(15) = 60 formas de seleccionar, en las que un solo paciente está expuesto a la hepatitis por esta causa. Dando por supuesto que las 120 formas posibles de seleccionar tres unidades de entre 10 son equiprobables, podemos hacer uso del método clásico para concluir que

$$ P[\text{un solo paciente esté expuesto al riesgo}] = \frac{60}{120} = 0.5 $$

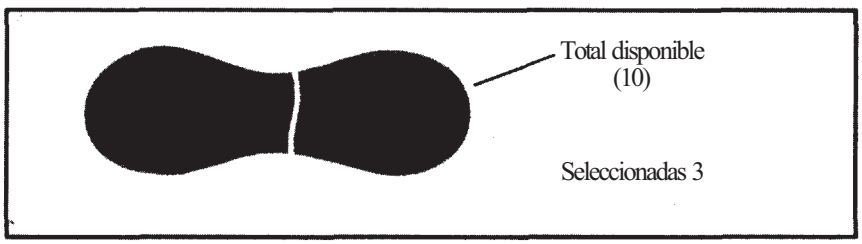


Figura 2.7. Partición del conjunto de unidades de sangre.

EJERCICIOS 2.6

1. Calcular:
   a) ${}_6C_2$ *)* $\binom{4}{0}$ *d)* $\binom{3}{3}$
   b) ${}_8C_5$
   e) Si ${}_nC_2 = 21$, ¿cuánto es *n?*
   f) Si ${}_nC_3 = 20$, ¿cuánto es *n?*
   g) Demostrar que ${}_5C_3 = {}_5C_2$.
2. Se dispone de un grupo de 12 pacientes para un determinado estudio. Se seleccionará a cinco de ellos con el fin de someterlos a un tratamiento experimental; los otros siete recibirán el tratamiento estándar y constituirán el grupo control. ¿De cuántas formas se puede seleccionar el grupo control? ¿Cuál es la probabilidad de que un individuo concreto A sea seleccionado para el grupo control? *Ayuda:* Imagine el grupo de pacientes dividido como aparece en la Figura 2.8.
3. Se está preparando un nuevo compuesto para ayudar a reducir la sequedad de la piel. Se pide a quince mujeres, de aproximadamente la misma edad y con el mismo tipo de piel, que tomen parte en un experimento comparativo. Se seleccionan siete de ellas aleatoriamente y se les pide que se apliquen el producto en las manos durante dos semanas. El resto de las mujeres constituye el grupo control. Al final del experimento, se pide a un juez imparcial que escoja a las siete mujeres cuyas manos estén en mejores condiciones. Si, de hecho, el tratamiento no ha tenido ningún efecto, ¿cuál es la probabilidad de que el juez seleccione por casualidad a las siete mujeres sobre las que se ha aplicado el compuesto experimental? ¿Cuál es la probabilidad de que se seleccionen cinco de las siete mujeres sobre las que se ha aplicado el compuesto experimental?
4. Un científico tiene seis jaulas diferentes de ratas blancas en el animalario. De ellas, dos contienen algunas ratas enfermas. ¿Cuál es la probabilidad de que, en una selección aleatoria de tres jaulas, ninguna con animales enfermos sea seleccionada? ¿Cuál es la probabilidad de que exactamente una de las jaulas con animales enfermos sea seleccionada?
5. Un químico tiene 10 muestras de agua tomadas de las aguas residuales de una fábrica de papel. Sin saberlo el químico, cuatro de las muestras son excesivamente ácidas. En una selección aleatoria de tres muestras, ¿cuál es la probabilidad de que exactamente (los sean en exceso ácidas?
6. Se han capturado, marcado y liberado diez osos salvajes. Más tarde, se captura una muestra de ocho osos y se cuenta cuántos están marcados. Se supone que no es más probable que se capture un oso que otro, por lo que cualquier conjunto de tamaño 8 tiene igual probabilidad. Supongamos que la población de osos en la región asciende a 100.
   a) ¿Cuántos subconjuntos de ocho pueden seleccionarse?

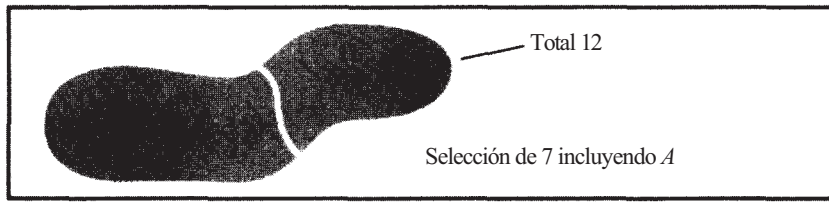


Figura 2.8. Partición del conjunto de pacientes.

b) ¿Se sorprendería si no se hubiera capturado ningún oso marcado? Arguméntelo basándose en la probabilidad de que ello ocurra.

c) ¿Cuál es la probabilidad de que los ocho osos estén marcados?

7. A un farmacéutico se le suministran 100 comprimidos de penicilina. Sin que él lo sepa, cinco han perdido su eficacia. Se cumplimenta una prescripción seleccionando aleatoriamente 15 comprimidos entre todos los existentes en *stock.* ¿Cuál es la probabilidad de que ninguno de los comprimidos elegidos haya perdido su eficacia? ¿Cuál es la probabilidad de que exactamente uno haya perdido su eficacia? ¿Cuál es la probabilidad de que al menos uno haya perdido su eficacia?
8. El proyecto Delta es un proyecto para determinar si en Alaska tendría éxito una producción agrícola a gran escala. En este proyecto se seleccionan 22 personas de un grupo de 103 candidatos especializados y se les concede el derecho de adquirir parcelas de terreno para trabajar con fines agrícolas. (Basado en la información de «Expanding Subartic Agriculture», *Interdisciplinary Science Reviews,* vol. 7, núm. 3, 1982, págs. 178-187.)
   a) ¿De cuántas maneras distintas pueden seleccionarse las 22 personas? (Sólo el planteamiento.)
   b) Supongamos que usted es una de las personas del grupo de candidatos. ¿En cuántos de los subgrupos del apartado *a* estaría incluido? (Sólo el planteamiento.)
   c) Si el proceso de selección se realiza aleatoriamente, cada uno de los subgrupos del apartado *a* es equiprobable. Si su nombre está en el grupo de candidatos, ¿cuál es la probabilidad de que se le otorgue el derecho de adquirir tierras?
9. Consideremos el Ejercicio 10 de la Sección 2.3. ¿De cuántas formas pueden seleccionarse los 15 cuadrados a inspeccionar si el mapa contiene un total de 50 cuadrados?
10. Se realiza un estudio para comparar la efectividad de 10 compuestos diferentes contra el dolor de cabeza. Si los compuestos se prueban a pares, ¿cuántas comparaciones diferentes son posibles?

HERRAMIENTAS COMPUTACIONALES

VI. $n!$, ${}_nC_r$, ${}_nP_r$

La calculadora TI83 incorpora en memoria la fórmula para $n!$, ${}_nC_r$ y ${}_nP_r$. Mostraremos su utilización calculando $10!$, $10!\,{}_{10}C_{10}P_3$.

Tecla/Comando de la TI83	**Propósito**
1. 10	1. Introduce el número 10.
2. MATH	2. Calcula $10!$ y borra la pantalla.
◁	
4	
ENTER	
CLEAR	
3. 10	3. Calcula ${}_{10}C_3$ y borra la pantalla.
MATH	
◁	
3	
3	
ENTER	
CLEAR	

4. 10
MATH
◁
2
3
ENTER
CLEAR

4. Calcula ${}_{10}P_3$ y borra la pantalla.

Teoría de probabilidades y resolución de problemas

En el Capítulo 2, hemos examinado la interpretación de las probabilidades y algunos métodos elementales para determinarlas. En este capítulo, continuamos nuestro estudio con la explicación de alguno de los teoremas útiles en la resolución de problemas para casos más complejos que los expuestos en el Capítulo 2.

3.1. DIAGRAMAS DE VENN Y LOS AXIOMAS DE PROBABILIDAD (OPCIONAL)

Diagramas de Venn

Antes de comenzar a desarrollar las reglas básicas que rigen el comportamiento de las probabilidades, presentaremos un diagrama que resulta útil para organizar las probabilidades. El diagrama, llamado *diagrama de Venn,* se denomina así en honor de John Venn (1834-1923). En este diagrama representamos el conjunto de posibilidades para un experimento mediante un rectángulo. A este conjunto le llamamos *espacio muestral* y lo representamos con la letra mayúscula *S* (Fig. 3.1a). Un suceso de interés se representa mediante una curva cerrada dentro del rectángulo y se indica mediante una letra mayúscula distinta de 5. En la Figura 3.1b se ha representado el suceso A. El suceso «que no se produzca A», se indica mediante A' y se representa en la región del rectángulo que queda fuera de *A* (Fig. 3.1c). El suceso *A′* se denomina suceso *complementario* de A. Cuando dos sucesos A_1 y A_2 están relacionados en el mismo experimento, dividen el rectángulo en cuatro áreas separadas. Cada área representa una forma exclusiva de combinar los dos sucesos. Éstas se muestran en la Figura *3. 1d* a *g.* En el ejemplo, ilustraremos esta idea.

> **Ejemplo 3.1.1.** Se diseña un estudio para investigar el peso y el hábito de fumar de los pacientes con hipertensión. Aquí *S* representa a todos los pacientes con hipertensión. Establezcamos que A_1 representa a los pacientes con sobrepeso y A_2 a los fumadores. La Figura 3. 1d representa a los pacientes con sobrepeso que no fuman; la Figura 3.le representa a los

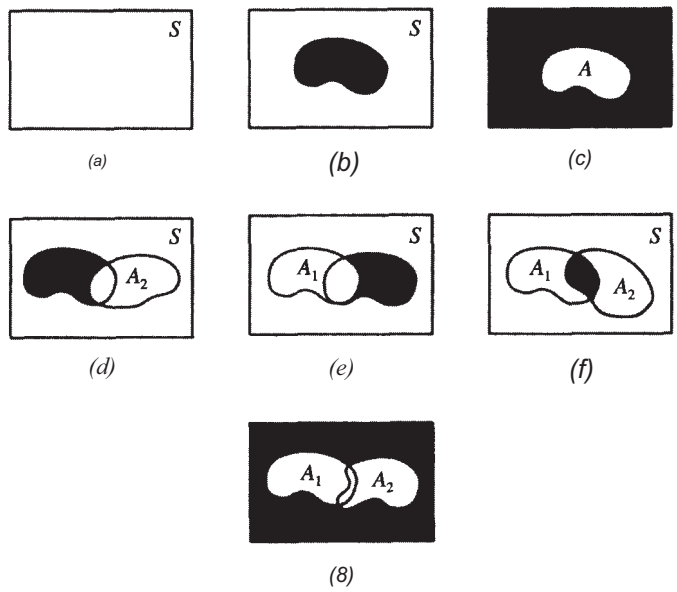


Figura 3.1. (a) El espacio muestral se representa mediante un rectángulo; *(b)* el suceso *A* se representa mediante una curva cerrada dentro del rectángulo; (c) el suceso *A'* es el suceso de que *A* no ocurra; (d) ocurre A_1 pero no A_2; *(e)* ocurre A_2 pero no A_1; *(f)* ocurren tanto A_1 como A_2 *(g)* no ocurre ni A_1 ni A_2.

que fuman pero que no tienen sobrepeso. La Figura 3.1/ muestra a los pacientes con sobrepeso y fumadores, mientras que los que ni fuman ni tienen sobrepeso están representados en la Figura 3.1g.

La palabra castellana «o» tiene dos significados diferentes. Cuando se utiliza con sentido de inclusión, significa «lo uno o lo otro» o «quizá ambos»; en el sentido de exclusión significa «o uno u otro» pero no «ambos». En este texto, la palabra «o» se utiliza en sentido de inclusión, salvo que se especifique lo contrario. Por ejemplo, si decimos que un paciente que sufre hipertensión tiene sobrepeso o fuma, queremos decir que el paciente presenta, al menos, una de estas características. El o ella tiene (1) sobrepeso pero no fuma o (2) fuma pero no tiene sobrepeso o (3) fuma y tiene sobrepeso. En la Figura 3.2 se muestra el diagrama de Venn para los sucesos A_1 *o* A_2.

Axiomas de probabilidad

Comenzamos considerando tres axiomas de probabilidad. Estos axiomas, que se admiten como ciertos y que no requieren demostración, son de origen intuitivo. Mucha gente los aplica de forma bastante natural sin tener la menor idea de lo que está haciendo.

Antes de establecer los axiomas, desarrollaremos una definición. Considérense los dos sucesos, A_1: el paciente *A* se recupera de una operación de corazón, y A_2: el paciente *A* fallece en la mesa de operaciones. Es evidente que estos sucesos no pueden producirse simultáneamente. El hecho de que se produzca uno excluye que sea posible el otro. Cuando esto ocurre, decimos que los sucesos A_1 y A_2 son *mutuamente excluyentes.* En la Figura 3.3a se muestra la representación del diagrama de Venn de dos sucesos mutuamente excluyentes. Obsérvese que, en este caso especial, las curvas que representan los dos sucesos no se superponen. La idea se extiende a un conjunto de sucesos mutuamente excluyentes en la Figura *3.3b.*

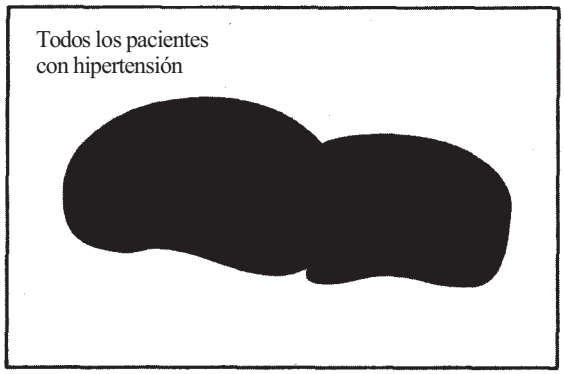


Figura 3.2. Pacientes que fuman o con sobrepeso.

Para sentar las bases de los teoremas básicos de resolución de los problemas de este capítulo se utilizan tres axiomas. Son los siguientes:

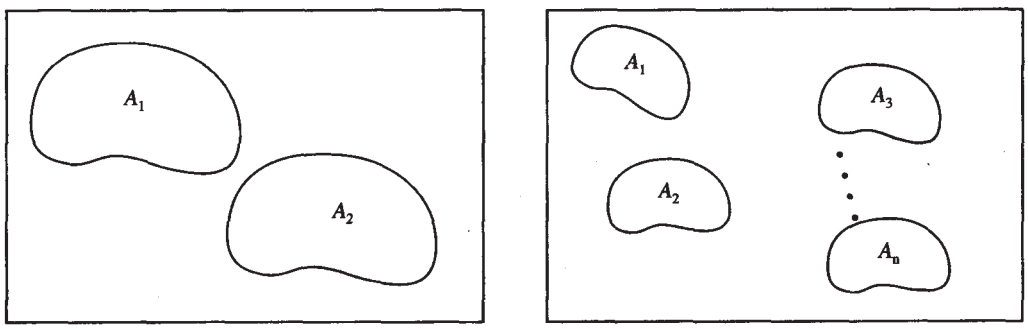
Axiomas de probabilidad

1. Sea *S* el espacio muestral de un experimento. Entonces $P[S] = 1$.
2. $P[A] > 0$ para todo suceso A.
3. Sea A_1, A_2, A_3, ... un conjunto finito o infinito de sucesos mutuamente excluyentes. Entonces $P[A_1 \text{ o } A_2 \text{ o } A_3 \text{ o } ...] = P[A_1] + P[A_2] + P[A_3] + ...$

El Axioma 1 alude a un hecho que a la mayoría de las personas le parecerá obvio, es decir, la probabilidad asignada al suceso seguro, al suceso cierto, es 1. El Axioma 2 afirma que la probabilidad nunca puede ser negativa. El Axioma 3 garantiza que cuando se tiene una serie de sucesos mutuamente excluyentes, la probabilidad de que ocurra uno *u* otro de los sucesos puede calcularse sumando las probabilidades individuales. Estos axiomas conducen fácilmente al Teorema 3.1.1.

Teorema 3.1.1. $P[\varnothing] = 0$.

El teorema establece que la probabilidad asociada al suceso «imposible», $\varnothing$, es 0. Puesto que el suceso imposible corresponde al suceso físico que no puede ocurrir, recurriremos a nuestros axiomas para asignar a tales sucesos la probabilidad 0. Por ejemplo, consideremos el experimento consistente en tirar un único dado corriente de seis caras. Las caras del dado contienen los números del 1 al 6. Si preguntamos cuál es la probabilidad de obtener 8 en una



(a) *(b)*

Figura 3.3. (a) Los sucesos A_1 y A_2 son mutuamente excluyentes. Si ocurre uno, el otro es imposible. *(b)* Un conjunto de *n* sucesos mutuamente excluyentes.

sola tirada, la respuesta es 0. El suceso descrito es físicamente imposible. La demostración del teorema se indica en el Ejercicio 13 de esta sección.

El Axioma 3 es especialmente importante porque nos proporciona la capacidad de calcular la probabilidad de un suceso cuando los puntos del espacio muestral del experimento no son equiprobables. Para comprender esta idea, consideremos el Ejemplo 3.1.2.

Ejemplo 3.1.2. La distribución de tipos de sangre en Estados Unidos entre los individuos de raza blanca es aproximadamente la siguiente:

A: 40% AB: 4%
B: 11% 0:45%

Tras un accidente de automóvil, un individuo de raza blanca es conducido a una clínica de urgencia. Se le hace un análisis de sangre para establecer el grupo al que pertenece. ¿Cuál es la probabilidad de que sea del tipo A, o del B, o del AB? Para hallar la probabilidad deseada se puede utilizar el Axioma 3. Vamos a denominar A_1, A_2 *y* A_3 a los sucesos relativos a que el paciente sea del grupo sanguíneo A, B y AB, respectivamente. Vamos a calcular $P[A_1 \text{ o } A_2 \text{ o } A_3]$. Dado que es imposible que un individuo tenga dos grupos sanguíneos diferentes, estos sucesos son mutuamente excluyentes. Por el Axioma 3,

$$\begin{aligned} P[A_1 \text{ o } A_2 \text{ o } A_3] &= P[A_1] + P[A_2] + P[A_3] \\ &= 0.40 + 0.11 + 0.04 \\ &= 0.55 \end{aligned}$$

Hay un 55% de posibilidades de que el paciente tenga uno de los tres grupos sanguíneos mencionados. (Basado en la información del *Technical Manual,* American Association of Blood Banks, 1985.)

Supongamos que conocemos la probabilidad de que se produzca el suceso A, y deseamos hallar la probabilidad de que *A* no se produzca. Podemos hacerlo fácilmente restando de 1. Por ejemplo, basándonos en una investigación realizada recientemente, estimaremos que la probabilidad de «curar» la leucemia infantil es de $\frac{1}{3}$. («Curar» significa que el niño se libra de la enfermedad durante al menos 4 años una vez finalizado el tratamiento.) Por lo tanto, la probabilidad de que la enfermedad no esté curada es $1 - \frac{1}{3} = 2/3$.

Esta idea, que parece evidente, se justifica con el Teorema 3.1.2, cuya demostración se presenta en el Ejercicio 14 de esta sección. Recuerde que *A'* indica el suceso de que *A no* ocurra.

Teorema 3.1.2. $P[A'] = 1 - P[A]$.

Obsérvese que este teorema proporciona una forma de hallar la probabilidad del suceso corriplementario del suceso *A*.

EJERCICIOS 3.1

1. Sea *L* el suceso que un paciente tiene leucemia y *W* el suceso que el recuento de leucocitos es alto. Considérense los diagramas de Venn de la Figura 3.4. Describir, en cada caso, los pacientes representados por la región sombreada.
2. Sea *H* el suceso que un árbol está situado en un lugar muy alto y *G* el suceso que el crecimiento de los árboles es deficiente. Considérense los diagramas de Venn de la Figura 3.5. Describir, en cada caso, los árboles representados por la región sombreada
3. En un estudio de vacunación realizado con niños en edad preescolar, el interés se centro en las vacunas contra la parotiditis y contra el sarampión. *P* representa el suceso de

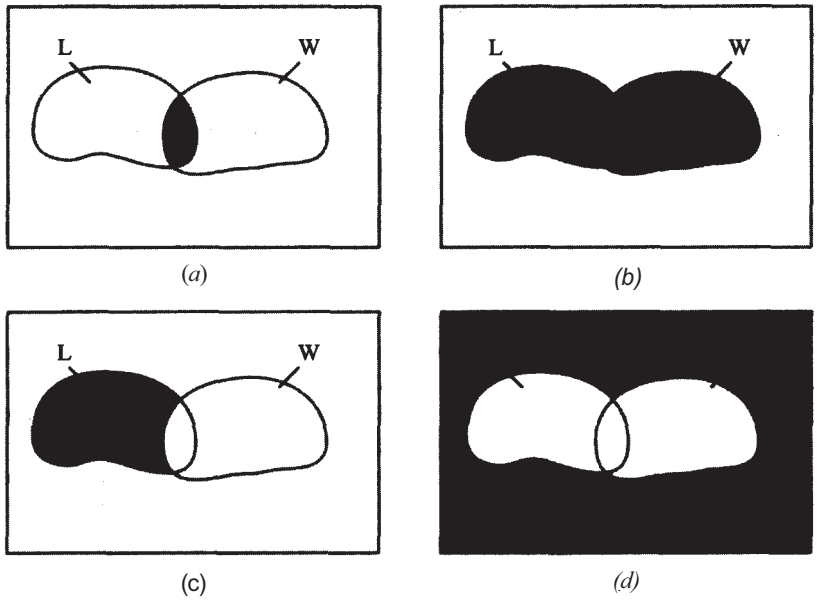


Figura 3.4

que un niño haya recibido la vacuna contra la parotiditis y *M* el suceso de que un niño haya recibido la vacuna contra el sarampión.

a) Describir a los niños del suceso *P* y *M*.

b) Dibujar un diagrama de Venn para representar al conjunto de niños que han recibido la vacuna contra el sarampión, pero no la vacuna contra la parotiditis.

c) Dibujar un diagrama de Venn para representar al conjunto de niños que no han recibido ninguna vacuna.

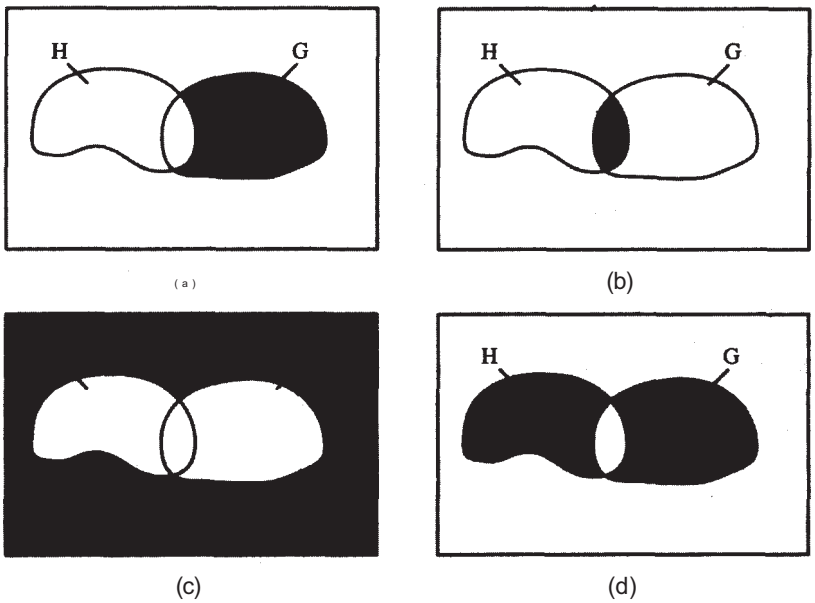


Figura 3.5

d) Dibujar un diagrama de Venn para representar al conjunto de niños que han recibido la vacuna contra la parotiditis o contra el sarampión.
e) Dibujar un diagrama de Venn para representar al conjunto de niños que han recibido la vacuna contra la parotiditis o el sarampión, pero que no han recibido ambas.

4. En un estudio sobre el efecto que produce el dióxido de azufre en los árboles a lo largo de las autopistas principales de los Smokies, se han identificado dos sucesos. Estos son: *L,* el árbol tiene las hojas dañadas, y *T,* el árbol ha crecido poco.
   a) Dibujar un diagrama de Venn para representar el conjunto de árboles que no han crecido poco.
   b) Dibujar un diagrama de Venn para representar el conjunto de árboles que han crecido poco, pero que no tienen las hojas dañadas.
   c) Dibujar un diagrama de Venn para representar el conjunto de árboles que no presentan ninguna de estas características.
   d) Dibujar un diagrama de Venn para representar el conjunto de árboles que han crecido poco o que tienen las hojas dañadas.
   e) Dibujar un diagrama de Venn para representar el conjunto de árboles que han crecido poco o que tienen las hojas dañadas, pero que no tienen ambos problemas.
5. ¿Cuáles de los siguientes pares de sucesos son mutuamente excluyentes?
   a) A: El hijo de Jane tiene hemofilia.
   B: La hija de Jane es portadora de hemofilia.
   b) A: El 65 % de las semillas de guisante que han sido plantadas germinará.
   B: El 50 % de las semillas de guisante que han sido plantadas no llegará a germinar.
   c) A: José sufre hipotermia.
   B: La temperatura de José es de 39 °C.
   d) *A:* El pH de una muestra de superficie de terreno es igual a 7:0.
   B: La muestra de superficie de terreno es alcalina.
   e) A: Un paciente tiene SIDA.
   B: El paciente ha recibido una transfusión de sangre.
   f) A: El animal es un mamífero.
   B: El animal es un delfín.
   C: El animal está cubierto de pelo.
   g) A: El árbol es de hoja perenne.
   B: El árbol es un encino.
   C: El árbol es un cornejo.
   h) A: El bosque es una extensión virgen.
   B: El bosque fue talado hace 10 años.
6. Tratando a bebés prematuros, la cantidad de oxígeno recibido puede afectar a su visión. Se puede categorizar a cada niño tratado como de visión normal, de lesión media, de lesión moderada, de lesión grave o ciego. Un estudio muestra que la probabilidad de que ocurra cada uno de estos sucesos es de 0.80, 0.10, 0.06, 0.02 y 0.02, respectivamente.
   a) Determinar la probabilidad de que un niño nazca con visión defectuosa.
   b) Determinar la probabilidad de que un niño nazca con visión normal.
7. Un determinado análisis químico tiene un alcance más bien limitado. Generalmente, el 15 % de las muestras están demasiado concentradas para que puedan contrastarse sin llevar a cabo una dilución previa, el 20 % están contaminadas con algún material obstaculizante que deberá ser eliminado antes de llevar a cabo el análisis. El resto puede ser analizado sin pretratamiento. Supongamos que las muestras no están en ningún caso concentradas y contaminadas a la vez. ¿Cuál es la probabilidad de que una muestra seleccionada aleatoriamente pueda ser contrastada sin pretratamiento?

8. La diabetes constituye un problema delicado durante el embarazo, tanto para la salud de la madre como para la del hijo. Entre las embarazadas diabéticas se presentan toxemias en un 25 % de los casos, hidroamnios en un 21 % y deterioro fetal en un 15 %. En un 6 % de los casos se dan otras complicaciones. Supongamos que no fuera posible que dos de estas complicaciones pudiesen presentarse simultáneamente en un mismo embarazo. ¿Cuál es la probabilidad de que, seleccionando aleatoriamente a una embarazada diabética, demos con un embarazo normal? ¿Cuál es la probabilidad de que exista algún tipo de complicación?
9. El índice de contaminación atmosférica elaborado por una central meteorológica clasifica los días como: extremadamente buenos, buenos, tolerables, malos o extremadamente malos. La experiencia anterior indica que el 50 % de los días se clasifican como extremadamente buenos, el 22 % como buenos, el 18 % como tolerables, el 8 % como malos y el 2 % como extremadamente malos. Se emite un pronóstico de los días clasificados como malos o extremadamente malos. ¿Cuál es la probabilidad de que un determinado día, elegido aleatoriamente, esté incluido en ese pronóstico?
10. Estudios sobre la depresión muestran que la aplicación de un determinado tratamiento mejora el estado del 72 % de aquellas personas sobre las que se aplica, no produce efecto alguno en un 10 %, y empeora el estado del resto. Se trata a un paciente que sufre de depresión, por estos medios, ¿cuál es la probabilidad de que empeore? ¿Cuál es la probabilidad de que el tratamiento no vaya en detrimento de su estado?
11. Los árboles de Mount Mitchell y otras zonas del sur de los Apalaches se han visto afectados por la polución. Supongamos que en una zona concreta el 40 % de los árboles de hoja perenne presentan daños leves, el 15 % daños moderados, el 10 % están muy afectados, el 8 % están muertos y el resto no están afectados. Si se selecciona aleatoriamente un árbol para un estudio, cuál es la probabilidad de que esté:
    a) No afectado.
    b) Muy poco afectado.
    c) Gravemente afectado o muerto.
    d) Ni gravemente afectado ni muerto.
12. La distribución del grupo sanguíneo de los individuos de raza negra de Estados Unidos es

    0: 49% B:20%
    A: 27% AB: 4%

    Si se lleva a una mujer de raza negra a una clínica de urgencias, ¿cuál es la probabilidad de que sea del tipo A, B o AB? (Basado en la información de *Technical Manual,* American Association of Blood Banks, 1985.)
13. Demostrar el Teorema 3.1.1. *Sugerencia:* Obsérvese que $S = S$ o $\varnothing$ y que S y $\varnothing$ son mutuamente excluyentes. Aplicar los Axiomas 3 y 1.
14. Demostrar el Teorema 3.1.2. *Sugerencia:* Obsérvese que $S = A$ o $\boldsymbol{A'}$ y que A y $\boldsymbol{A'}$ son mutuamente excluyentes. Aplicar los Axiomas 1 y 3.
15. Sean A y *B* dos sucesos tales que A está contenido en *B* (véase la Fig. 3.6). Obsérvese que

$$B = Ao(B \text{ pero NO } A)$$

    y que los sucesos de la parte derecha de la igualdad son mutuamente excluyentes.
    a) Utilizar la información y los teoremas y axiomas desarrollados en esta sección para probar que

$$\boldsymbol{P[A] \leqslant P[B]}$$

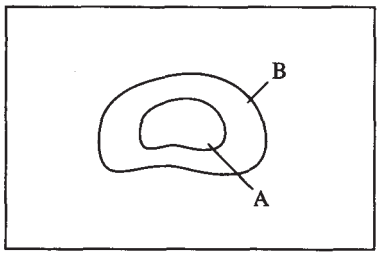


Figura 3.6. Diagrama de Venn mostrando el suceso *A* contenido en el suceso *B*.

b) Sea *C* un suceso cualquiera. Probar que $P[C] < 1$ utilizando el apartado a y los teoremas y axiomas estudiados. Se puede comprobar así la afirmación hecha en el Capítulo 2 acerca de que las probabilidades no pueden exceder de 1.

3.2. REGLA GENERAL DE LA ADICIÓN

En la Sección 3.1, vimos cómo tratar cuestiones relativas a la probabilidad de que se produzca uno u otro de dos sucesos mutuamente excluyentes. En esta sección, vamos a examinar la regla general de la adición. Su propósito es permitir el manejo del caso más general, calcular la probabilidad de que ocurra al menos uno de dos sucesos que no es necesario que sean mutuamente excluyentes.

Comenzaremos por observar el diagrama de Venn de la Figura 3.7. Obsérvese que A_1 y A_2 *no* son mutuamente excluyentes. Por tanto, la región sombreada no es una región vacía. Si calculamos $P[A_1$ o $A_2]$ como en la Sección 3.1, concluiremos que

No obstante, dado que la región sombreada está contenida en A_1 y A_2, incluimos $P[A_1$ y $A_2]$ dos veces en el cálculo anterior. Para corregirlo, debemos restar $P[A_1 \, y \, A_2]$ del miembro de la derecha de la ecuación. La expresión resultante es la regla general de la adición.

Teorema 3.2.1. Regla general de la **adición.** Sean los sucesos A_1 y A_2. Entonces

$$P[A_1 \, o \, A_2] = P[A_1] + P[A_2] - P[A_1 \, y \, A_2]$$

La palabra clave para, dado un determinado problema, saber si puede aplicarse en él la regla general de la adición, es la palabra «o». Por el tercer axioma de la probabilidad y la regla general de la adición, se puede decir con seguridad que si en un problema de probabili-

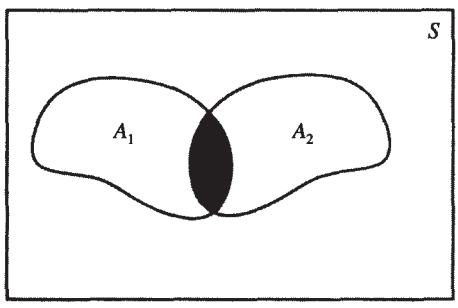


Figura 3.7. A_1 y A_2 no son mutuamente excluyentes. La zona sombreada no es una zona vacía.

dad aparece la palabra *o,* la adición está casi siempre involucrada. En las Secciones 3.5 y 3.6 se verá que la palabra «y» es un indicativo de que se multiplicarán los números para resolver el problema de que se trate. El empleo de esta regla se explica en el Ejemplo 3.2.1.

Ejemplo 3.2.1. Se estima que el 30 % de los habitantes de Estados Unidos son obesos (A_1) y que el 3 % sufre diabetes (A_2). El 2 % es obeso y padece de diabetes. ¿Cuál es la probabilidad de que una persona aleatoriamente elegida sea obesa o sufra diabetes? Se nos da $P[A_1] = 0.3$, $P[A_2] = 0.03$ y $P[A_1$ y $A_2] = 0.02$. Hemos de hallar $P[A_1$ *o* $A_2]$. Aplicando la regla general de la adición, obtenemos:

$$\begin{aligned} P[A_1 \text{ o } A_2] &= P[A_1] + P[A_2] - P[A_1 \text{ y } A_2] \\ &= 0.30 + 0.03 - 0.02 \\ &= 0.31 \end{aligned}$$

La regla general de la adición no sólo es útil para calcular $P[A_1$ o AJ, sino que, disponiendo de la información adecuada, podemos utilizarla para hallar $P[A_1$ y $A_2]$. El Ejemplo 3.2.2 ilustra cómo se hace.

Ejemplo 3.2.2. Se sabe por informes recientes que el 18 % de los estudiantes de segunda enseñanza sufre depresión en algún período de su escolarización (A,), que el 2 % piensa en el suicidio (A_2) y que el 19 % padece depresión o piensa en el suicidio. ¿Cuál es la probabilidad de que un estudiante de secundaria elegido aleatoriamente sufra depresión y piense en el suicidio? ¿Cuál es la probabilidad de que un estudiante de secundaria elegido aleatoriamente sufra depresión *pero* no piense en el suicidio?

Sabemos que $P[A_1] = 0.18$, $P[A_2] = 0.02$ y $P[A_1$ o $A_2] = 0.19$. Hemos de hallar, primero, $P[A_1$ y $A_2]$. Aplicando la regla general de la adición, obtenemos

$$P[A_1 \text{ o } A_2] = P[A_1] + P[A_2] - P[A_1 \text{ y } A_2]$$

o bien

$$\begin{aligned} P[A_1 \text{ y } A_2] &= P[A_1] + P[A_2] - P[A_1 \text{ o } A_2] \\ &= 0.18 + 0.02 - 0.19 \\ &= 0.01 \end{aligned}$$

Para resolver la segunda cuestión propuesta, utilizamos la información dada por el diagrama de Venn. Puesto que $P[A_1$ *y* $A_2] = 0.01$, sabemos que el 1 % del área total del diagrama corresponde a la región representada por A, y A_2, como muestra la Figura 3.8a. Puesto que $P[A_1] = 0.18$, del área total, el 18 % corresponde a la región marcada por A_1; dado que $(A_1$ *y* $A_2)$ está contenido en A_1 el 17% del área corresponde a la región sombreada de la Figura 3.8b. Análogamente, puesto que $P[A_2] = 0.02$ y (A, y A_2) está contenido en A_2, el 1 % del área corresponde a la región sombreada de la Figura 3.8c. Ya que P[S] = 1 y que tenemos ya contabilizado el 17 + 1 + 1 = 19% del área, el 81 % restante corresponde a la región sin sombrear de la Figura *3.8d.* Ahora podemos resolver la segunda cuestión buscando la región apropiada en el diagrama de Venn, es decir, A_1 y A'_2. Puede verse que la probabilidad asociada a esta región es 0.17. Por lo tanto, la probabilidad de que un estudiante de segunda enseñanza sufra depresión pero no haya pensado en el suicidio es 0.17.

Obsérvese que, si los porcentajes registrados en problemas como éstos están basados en los datos de población, las probabilidades calculadas utilizando la regla general de la adición son exactas. Sin embargo, si los porcentajes están basados en muestras extraídas de una población mayor, las probabilidades calculadas son frecuencias relativas. Son *aproximaciones* a la probabilidad real de que se produzca el suceso en cuestión. Dado que muchos porcen-

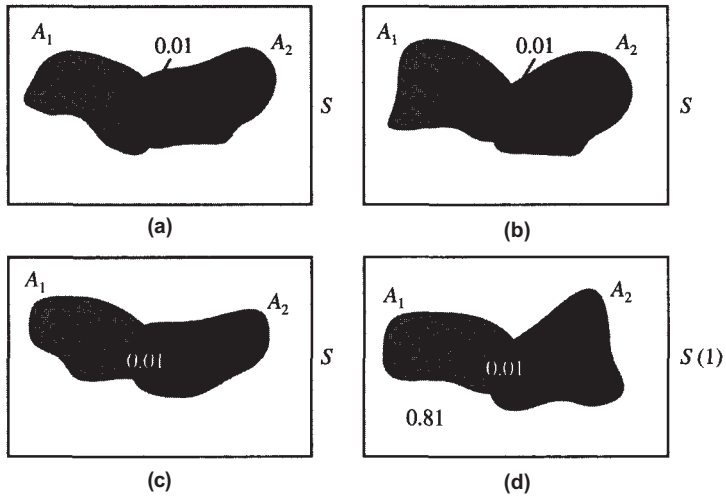


Figura 3.8. Cálculo de probabilidades mediante los diagramas de Venn. (a) $P[A_1$ y $A_2] = 0.01$; *(b)* $P[A_1] = 0.18$; (c) $P[A_2] = 0.02$; (d) $P[A_1$ o $A_2] = 0.19$, lo cual implica que $P[(A_1$ *o* $A_2)'] = 0.81$.

tajes de los registrados en la bibliografía se basan en muestras, la mayoría se interpreta correctamente como frecuencias relativas. Utilizamos la palabra *probabilidad,* pero entendiendo que las probabilidades dadas y calculadas utilizando los teoremas de este capítulo son, en muchos de los casos, sólo aproximaciones.

EJERCICIOS 3.2

1. Supongamos que $P[A_1$ y $A_2] = 0.04$, $P[A_1] = 0.06$, $P[A_2] = 0.10$. Hallar
   a) $P[A_1$ o $A_2]$
   b) $P[A_1$ y $A_2']$
   c) $P[A_1'$ y $A_2]$
   d) $P[A_1'$ y $A_2']$
   e) Diseñar un diagrama de Venn para visualizar la descomposición de las probabilidades dentro del espacio muestral, según se ha descrito en la Figura 3.8.
2. Supongamos que $P[A_1$ o $A_2] = 0.30$, $P[A_1] = 0.15$, $P[A_2] = 0.20$. Hallar
   a) $P[A_1$ y $A_2]$
   b) $P[A_1$ y $A_2']$
   c) $P[A_1'$ y $A_2]$
   d) $P[A_1'$ o $A_2']$
   e) $P[(A_1$ o $A_2)']$
   f) Diseñar un diagrama de Venn para visualizar la descomposición de las probabilidades dentro del espacio muestral, según se ha descrito en la Figura 3.8.
3. Para satisfacer la demanda de los granjeros de utilizar pinos blancos jóvenes como protección contra el viento, los empleados del servicio forestal tomaron muestras de los granjeros del estado. Hallaron que el 30 % había adquirido árboles del servicio forestal en años anteriores, el 40 % había anticipado el pedido de árboles para el año siguiente, el 10% había adquirido árboles en el pasado y anticipado el pedido de árboles para el año siguiente. ¿Cuál es la probabilidad de que un granjero seleccionado aleatoriamente haya adquirido árboles en el pasado o haya anticipado el pedido para el año siguiente? ¿Cuál

es la probabilidad de que un granjero seleccionado aleatoriamente haya adquirido árboles en el pasado pero no haya realizado ningún pedido por adelantado para el año siguiente? Si a cada granjero que solicita árboles se le conceden como máximo 100 y hay 5000 granjeros en el estado, hallar una aproximación del número máximo de árboles necesarios para completar todas las peticiones para el año siguiente.

4. Los datos recogidos en un banco de sangre concreto indican que el 0.1 % de todos los donantes da positivo en el test para el virus de inmunodeficiencia humana (VIH) y el 1 % da positivo para el test del herpes. Si el 1.05 % da positivo para uno u otro de estos problemas, ¿cuál es la probabilidad de que un donante seleccionado aleatoriamente no tenga ninguno de estos problemas? ¿Le sorprendería hallar un donante con ambos problemas? Explíquelo basándose en la probabilidad estimada de que suceda.
5. Se ha determinado que el 62 % de todos los servicios sanitarios está financiado por fundaciones privadas, que el 70 % se financia por medio de cooperativas de empresarios y trabajadores, y que el 50 % se financia tanto por fundaciones privadas como por medio de cooperativas de empresarios y trabajadores. ¿Cuál es la probabilidad de que un paciente elegido al azar sea atendido por unos servicios sanitarios que dependan financieramente de una fundación privada o de una cooperativa de empresarios y trabajadores? ¿Cuál es la probabilidad de que un paciente elegido al azar sea atendido por unos servicios sanitarios financieramente dependientes de una cooperativa de empresarios y trabajadores, pero no de una fundación privada?
6. Ciertos estudios muestran que un 12 % de las personas tratadas por médicos es atendido en el hospital. De ellas el 1 % sufre alguna alergia a medicamentos, y el 12.4 % recibe atención en un hospital o es alérgico a los medicamentos. ¿Cuál es la probabilidad de que un paciente elegido al azar reciba atención en un hospital y sea alérgico a los medicamentos? ¿Cuál es la probabilidad de que un paciente elegido al azar sea ingresado en un hospital pero no sufra alergia a medicamentos? ¿Cuál es la probabilidad de que un paciente elegido al azar sea alérgico a los medicamentos pero no reciba atención en un hospital?
7. Un químico analiza muestras de agua de mar para detectar la presencia de dos metales pesados: plomo y mercurio. Encuentra que el 38 % de las muestras tomadas en las proximidades de la desembocadura de un río en cuyas orillas se localizan numerosas plantas industriales tiene niveles tóxicos de plomo o de mercurio, y que el 32 % tiene nivel tóxico de plomo. De estas muestras, el 10 % contiene un nivel alto de ambos metales. ¿Cuál es la probabilidad de que una muestra dada contenga un alto nivel de mercurio? ¿Cuál es la probabilidad de que una muestra dada contenga solamente plomo?
8. Si a ratones de una cierta raza suiza se les suministra 1 mg de compuesto A por kg de peso, muere el 50 % de los animales (una dosis que mata a un 50 % de los animales puestos a prueba se conoce como la LD50 del fármaco o veneno) y el 40 % de los animales tratados, supervivientes o no, presenta cianosis (es decir, su piel tiene un tono azulado que indica una inadecuada oxigenación en la sangre). Una cuarta parte de los animales muere y muestra una evidente cianosis. ¿Cuál es la probabilidad de que un animal al que se le ha administrado el compuesto A (la dosis LD_{50}) muera o esté cianótico? ¿Cuál es la probabilidad de que un animal al que se le ha administrado el compuesto A viva y esté cianótico?

3.3. PROBABILIDAD CONDICIONADA

En esta sección introducimos la noción de probabilidad condicionada. El nombre es, en sí mismo, significativo de lo que vamos a hacer. Pretendemos determinar la probabilidad de que ocurra un suceso A_2 «condicionado por» el hecho de que algún otro suceso A_1 haya ocurrido ya. Las palabras clave a las que debe prestarse atención para identificar una probabilidad

condicionada son *si y dado que.* Utilizaremos la notación $P[A_2 \mid A_1]$ para designar la probabilidad del suceso A_2 condicionada por el hecho de que haya sucedido previamente A_1. Obsérvese que, a pesar de que en esta expresión intervienen dos sucesos, se alude únicamente a una probabilidad. El primero de los sucesos reseñados es aquel que no sabemos si ocurrirá o no; la barra se lee «dado que»; el segundo suceso es el que se supone que ha ocurrido ya.

Ejemplo 3.3.1. Una mujer tiene tres hijos. ¿Cuál es la probabilidad de que los dos primeros sean chicos (A_1)? ¿Cuál es la probabilidad de que exactamente dos sean chicos (A_2)? ¿Cuál es la probabilidad de que se satisfagan ambas condiciones?

Estas son preguntas no condicionadas y fáciles de contestar utilizando un diagrama en árbol (véase Fig. 3.9). Si suponemos que cada hijo tiene la misma posibilidad de ser chico que chica, entonces los ocho puntos muestrales representados en el diagrama son igualmente probables. Por esta razón, puede usarse la aproximación clásica para calcular las probabilidades deseadas. En particular

$$P[A_1] = \tfrac{2}{8}$$
$$P[A_2] = \tfrac{3}{8}$$
$$P[A_1 \text{ y } A_2] = \tfrac{1}{8}$$

Supongamos que ya sabemos que los dos primeros hijos son chicos. Ahora, ¿cuál es la probabilidad de que haya exactamente dos chicos en la familia? Esto es, ¿cuál es $P[A_2|A_1]$? Puesto que sabemos que los dos primeros hijos son chicos, el espacio muestral para el experimento lógicamente no estará constituido por los ocho puntos, sino que, de hecho, ahora contendrá solamente los dos puntos MMM y MMF. El resto de los puntos no son consistentes con la información que tenemos. La pregunta condicionada planteada se resuelve mediante este nuevo espacio muestral formado por dos puntos. Ya que estos dos puntos son igualmente probables, y sólo uno de ellos corresponde a tener exactamente dos chicos en la familia,

En este caso observamos que $\frac{1}{2} = P[A_2 \mid A_1] \neq P[A_2] = \frac{3}{8}$. La nueva información afecta a la probabilidad asignada al suceso de que exactamente dos de los niños sean varones.

El Ejemplo 3.3.1 es una simplificación del problema general. La mayor parte de las preguntas que se plantean sobre probabilidad condicionada se refiere a situaciones en las que no es conveniente trabajar directamente con un espacio muestral restringido explícitamente. Así que es necesario desarrollar una fórmula para la probabilidad condicionada que, en esencia,

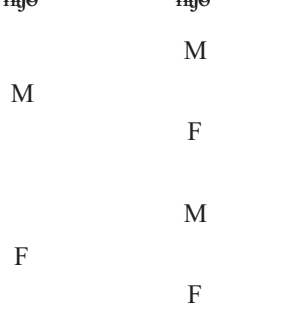


Figura 3.9. Orden de nacimiento en el árbol filial de la familia.

reduzca automáticamente el espacio muestral hasta hacerlo coherente con la información dada, y que permita calcular la probabilidad pedida relativa a este espacio muestral reducido. Para encontrar esta fórmula sólo necesitamos mirar el modelo del Ejemplo 3.3.1. Con ello es más que suficiente. Obsérvese que

$$P[A_2 \mid A_1] = \frac{1}{2} = \frac{\frac{1}{8}}{\frac{2}{8}} = \frac{P[A_1 \text{ y } A_2]}{P[A_1]}$$

Esta relación no es exclusiva de este problema. Se trata en realidad de la definición general de la probabilidad condicionada del suceso A_2, dado A_1.

Definición 3.3.1. Probabilidad condicionada. Sean A_1 y A_2 dos sucesos tales que $P[A_1] \neq 0$. La probabilidad condicionada de A_2 dado A_1, denotada $P[A_2|A_1]$ se define por

$$P[A_2 \mid A_1] = \frac{P[A_1 \text{ y } A_2]}{P[A_1]}$$

En la práctica, la condición $P[A_1] \neq 0$ no es restrictiva. Si A_1 ya ha ocurrido, ha de tener originalmente una probabilidad no nula. La Definición 3.3.1 se recuerda fácilmente del modo siguiente:

$$\text{Probabilidad condicionada} = \frac{P[\text{ambos sucesos}]}{P[\text{suceso dado}]}$$

Ejemplo 3.3.2. Se estima que el 15 % de la población adulta padece hipertensión, pero que el 75 % de todos los adultos cree no tener este problema. Se estima también que el 6 % de la población tiene hipertensión pero no es consciente de padecer dicha enfermedad. Si un paciente adulto opina que no es hipertenso, ¿cuál es la probabilidad de que la enfermedad, de hecho, exista?

Siendo A_1 el suceso «el paciente no cree tener la enfermedad» y A_2 el suceso «la enfermedad existe», se nos ha dado que $P[A_1] = 0.75$, $P[A_2] = 0.15$ y $P[A_1 \text{ y } A_2] = 0.06$. Pretendemos hallar $P[A_2|A_1]$.

Por la Definición 3.3.1,

$$P[A_2 \mid A_1] = \frac{P[\text{ambos}]}{P[\text{dado}]} = \frac{P[A_1 \text{ y } A_2]}{P[A_1]}$$

$$= \frac{0.06}{0.75} = 0.08$$

Hay un 8 % de posibilidades de que un paciente que opine que no tiene problemas de hipertensión padezca, de hecho, la enfermedad. Del mismo modo podemos preguntar: si la enfermedad existe, ¿cuál es la probabilidad de que el paciente lo sospeche? Es decir, ¿cuál es $P[A'_1| A_2]$? Antes de aplicar la Definición 3.3.1, organicemos los datos por medio de un diagrama de Venn, como se muestra en la Figura 3.10. Por la Definición 3.3.1,

$$P[A'_1 \mid A_2] = \frac{P[\text{ambos}]}{P[\text{dado}]} = \frac{P[A'_1 \text{ y } A_2]}{P[A_2]}$$

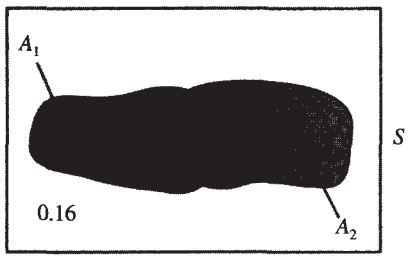


Figura 3.10. A_1 = No creen que exista la enfermedad. A_2 = La enfermedad existe.

Observando el diagrama de Venn, tenemos

$$P[A_2] \quad 0.15$$

Es decir, si el paciente opina que tiene hipertensión, existe un 60 % de probabilidad de que esté en lo cierto.

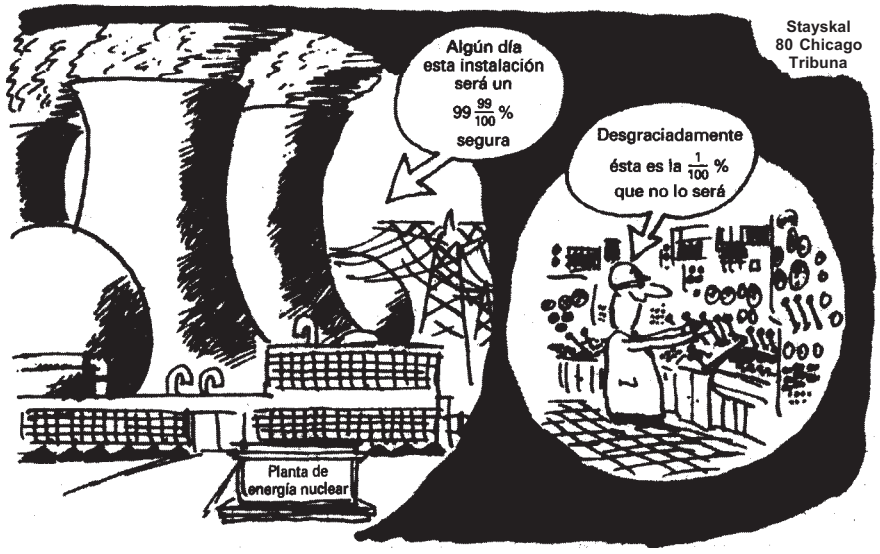
EJERCICIOS 3.3

1. Supongamos que una familia tiene cuatro hijos.
   a) Hallar la probabilidad de que exactamente dos sean varones.
   b) ¿Cuál es la probabilidad de que exactamente dos sean varones si el nacido en primer lugar es varón?
   c) Hallar la probabilidad de que el último hijo nazca varón.
   d) ¿Cuál es la probabilidad de que el último hijo nazca varón si los tres primeros son mujeres?
2. Supongamos que una plaga afecta al 50 % de todos los cornejos de un área dada. Se toma una muestra de tres árboles y cada uno se clasifica como afectado por la plaga *(s)* o no afectado (n). Dado que *P*[afectado] = *P*[no afectado], cada uno de los ocho resultados posibles del experimento tiene la misma probabilidad.
   a) Dibujar un árbol para representar los ocho elementos muestrales.
   b) Hallar P[al menos dos están afectados].
   c) Hallar P[al menos dos están afectados | el primero está afectado].
   d) Hallar inexactamente dos están afectados | el primero está afectado].
3. Un estudio indica que el 10 % de la población de Estados Unidos tiene 65 o más años, *y* que el 1 % de la población total padece insuficiencia cardíaca moderada. Además, el 10.4 % de la población tiene 65 o más años o padece insuficiencia cardíaca moderada. Eligiendo a un individuo al azar:
   a) Hallar la probabilidad de que el individuo tenga 65 o más años y padezca de insuficiencia cardíaca moderada.
   b) Utilizar la solución del apartado *α* para organizar los datos en un diagrama de Venn.
   c) Si un individuo tiene 65 o más, ¿cuál es la probabilidad de que padezca de insuficiencia cardíaca moderada?
   d) Si un individuo es menor de 65 años, ¿cuál es la probabilidad de que padezca die insuficiencia cardíaca moderada?
4. En un estudio sobre alcohólicos se informa de que el 40 % de los mismos tiene padre alcohólico y el 6 *%,* madre alcohólica. El 42 % tiene al menos uno de los padres alcohólicos. ¿Cuál es la posibilidad de que elegido uno al azar
   a) Tenga ambos padres alcohólicos?

b) ¿Tenga una madre alcohólica si lo es el padre?
c) ¿Tenga una madre alcohólica pero no un padre alcohólico?
d) ¿Tenga una madre alcohólica si el padre no lo es?

En un estudio sobre sensibilidad, se practican necropsias en encéfalos de pacientes afectados de demencia senil o degeneración arteriosclerótica cerebral. Se informa de que el 35 % tiene alteraciones asociadas principalmente con la demencia senil, el 45 % tiene alteraciones asociadas con la degeneración arteriosclerótica cerebral, y el 20 % muestra signos de ambas. Basándose en esta información ¿cuál es la probabilidad de que un paciente con el cerebro dañado a consecuencia de una degeneración arteriosclerótica tenga también alteraciones cerebrales características de la demencia senil? ¿Cuál es la probabilidad de que un paciente que no tiene alteraciones debidas a la demencia senil padezca degeneración arteriosclerótica cerebral?

6. En un estudio de aguas localizadas en las proximidades de centrales eléctricas y de otras plantas industriales que vierten sus desagües en el hidrosistema, se ha llegado a la conclusión de que el 5 % muestra signos de contaminación química y térmica, el 40 % de contaminación química y el 35 % de contaminación térmica. Suponiendo que los resultados del estudio reflejen correctamente la situación general, ¿cuál es la probabilidad de que un arroyo que muestra cierta contaminación térmica presente también signos de contaminación química? ¿Cuál es la probabilidad de que un arroyo que muestra cierta contaminación química no presente signos de contaminación térmica?



(© Copyright 1980. Chicago Tribune Company. Todos los derechos reservados. Reproducido con autorización.)

7. Unos estudios muestran que los ejemplares de una cierta raza de liebres de alta montaña (liebre esquiadora) mueren antes de lo normal, aun en ausencia de depredadores o de enfermedad conocida alguna. Dos de las causas de muerte identificadas son: baja cantidad de azúcar en sangre y convulsiones. Se estima **que** el 7 % **de** los **animales** presenta ambos síntomas, el 40 % tiene bajo nivel de azúcar en sangre, y el 25 % sufre convulsiones, ¿cuál es el porcentaje de muertes producidas por causas que no sean las que hemos mencionado? ¿Cuál es la probabilidad de que un animal elegido aleatoriamente que tiene bajo nivel de azúcar en sangre sufra también convulsiones?

8. Utilizar los datos del Ejercicio 3 de la Sección 3.2 para hallar la probabilidad de que un granjero pida árboles para el año que viene dado que ya ha pedido árboles en el pasado. Hallar la probabilidad de que un granjero no efectúe ningún pedido de árboles para el año que viene dado que ya ha adquirido árboles en el pasado. ¿Cuál es la relación matemática entre las dos respuestas? Explicar por qué ha sucedido esto.
9. Utilizar los datos del Ejercicio 4 de la Sección 3.2 para hallar la probabilidad de que un donante seleccionado aleatoriamente sea negativo para el test del VIH. Hallar la probabilidad de que un donante seleccionado aleatoriamente sea negativo para el test del VIH dado que da negativo en el test del herpes.

3.4. TESTS DE DIAGNÓSTICO Y RIESGO RELATIVO

Una de las aplicaciones más útiles de las probabilidades en el campo médico o biológico está en el área de los tests de diagnóstico. Un test de diagnóstico es un test para detectar la presencia de alguna condición específica en una unidad experimental. En el campo médico, generalmente intentamos detectar la presencia de una enfermedad, un factor genético o alguna otra condición específica en un ser humano. Sería deseable que estos tests fueran seguros en el sentido de que siempre detectaran la condición cuando de hecho está presente y nunca la indicaran cuando el individuo no tiene dicha condición. Desgraciadamente, esto no es así. Sólo podemos esperar que los tests existentes actualmente no den resultados erróneos muy a menudo.

En un test de diagnóstico, o bien cada sujeto es realmente positivo, lo que significa que la condición para la que está diseñado el test está presente, o bien el sujeto en realidad es negativo. El test, en sí mismo, puede dar positivo, lo que significa que se ha detectado la presencia de la condición, o no darlo. Esto garantiza que cada sujeto entrará exactamente en una de entre cuatro categorías. Éstas son:

1. La condición está presente y el test detecta su presencia. Es decir, un sujeto realmente positivo da positivo. En este caso no se ha cometido ningún error.
2. La condición está presente pero el test no detecta su presencia. Cuando un sujeto realmente positivo da negativo, decimos que hemos obtenido un resultado *falso negativo.* En este caso se ha cometido un error.
3. La condición no está presente pero el test detecta su presencia. Cuando un sujetó realmente negativo da positivo, decimos que se ha obtenido un resultado *falso positivo.* Se ha cometido un error.
4. La condición no está presente y el test no indica su presencia. Un sujeto realmentje negativo da negativo. No se ha cometido ningún error.

Obsérvese que es posible cometer dos tipos de errores. Esperamos no cometer ninguno pero cualquiera de ellos es posible. Un test ideal es aquel en el que la probabilidad de cometer cualquier error es pequeña. A continuación se definen estas probabilidades, llamadas coeficientes de error.

Definición 3.4.1. Coeficiente de falsos positivos. El *coeficiente de falsos positivos* de un test se denota por α (alfa) y viene dado por

$$\alpha = P[\text{el test resulta positivo} \mid \text{el sujeto es realmente negativo}]$$

Definición 3.4.2. Coeficiente de falsos negativos. El *coeficiente de falsos negativos* de un test se denota por β (beta) y viene dado por

$$\beta = P[\text{el test resulta negativo} \mid \text{el sujeto es realmente positivo}]$$

La Tabla 3.1 resume la terminología introducida hasta ahora.

Tabla 3.1. Terminología asociada con las pruebas de diagnóstico

		Estado real	
		Condición ausente (-)	**Condición presente (+)**
Resultados del test	Condición encontrada (+)	Realidad - pero test + Falso positivo Pffalso positivo] = α	Realidad + y test + No hay error
	Condición no encontrada (-)	Realidad - y test - No hay error	Realidad + pero test - Falso negativo P[falso negativo] = β

Al realizar un test para detectar una enfermedad, un error causado por un alto coeficiente de falsos positivos puede ocasionar inconvenientes y gastos a la persona implicada. A ésta se le detecta una enfermedad que no está presente y como resultado probablemente busque un tratamiento para un problema inexistente. Un error resultante de un alto coeficiente de falsos negativos es potencialmente peligroso. En este caso, el sujeto desconoce una condición existente y, por lo tanto, no buscará el tratamiento que necesita. Con una tabla de frecuencias pueden hallarse las aproximaciones de estos coeficientes mediante la técnica descrita en el Ejemplo 3.4.1.

Ejemplo 3.4.1. El suero de una mujer embarazada puede ser analizado por medio de un procedimiento llamado electroforesis en gel de almidón. Este procedimiento permite detectar la presencia de una zona proteínica llamada *zona de embarazo,* la cual se supone que es un indicador de que el niño es una hembra. Para investigar las propiedades de este test se seleccionaron 300 mujeres para su estudio. En la Tabla 3.2 se dan los resultados del test y los sexos de los niños nacidos. Obsérvese que, en este caso, el único valor de la tabla predeterminado o fijado por el experimentador es el tamaño total de la muestra. Todos los otros son aleatorios, los totales por filas, los totales por columnas y las frecuencias de las celdas. Por definición el coeficiente de falsos positivos es

$$\alpha = P[\text{test} + \mid \text{realidad} -]$$

Para estimar esta probabilidad condicionada debemos estimar P[realidad -] y P[test + y realidad -] . Utilizando el método de la frecuencia relativa para hallar la probabilidad, P[realidad -] $\simeq$ 147/300 y P[test + y realidad -] $\simeq$ 51/300. La definición de la probabilidad condicionada nos conduce a

$$\alpha \cong \frac{51/300}{147/300} = \frac{51}{147} = 0.3469$$

Tabla 3.2

	Sexo		
Zona de embarazo	**Varón (realidad -)**	**Mujer (realidad +)**	
Presente (test +)	51	78	129 (aleatorio)
Ausente (test -)	96	75	171 (aleatorio)
	147 (aleatorio)	153 (aleatorio)	300 (fijo)

Este resultado puede obtenerse por observación directa de la Tabla 3.2 considerando que, una vez que sabemos que el sujeto es realmente negativo, la atención deberá centrarse inmediatamente en los 147 casos de la columna 1. De éstos, 51 dieron positivo. Por tanto, el sentido común apunta a 51/147 como el coeficiente de falsos positivos estimado. Para calcular β), obsérvese que de los 153 sujetos realmente positivos, 75 dieron negativo. Por lo tanto

$$\beta \cong \frac{75}{153} = 0.4902$$

Dado que son coeficientes de error estimados, el test no parece ser efectivo para determinar el sexo de un niño. (Basado en datos registrados en *Human Heredity,* vol. 20, 1970, pág. 530.)

También pueden considerarse otros dos coeficientes, la *especificidad* y la *sensibilidad.* Estos coeficientes dan la probabilidad de tomar las decisiones correctas en la elaboración de un diagnóstico. Se definen en los Ejercicios 5 y 6 de esta sección.

La técnica descrita en el Ejemplo 3.4.1 puede emplearse para estimar las probabilidades condicionadas en temas distintos de los tests de diagnóstico. Sin embargo, es preciso hacer una llamada de atención. Si todos los totales por filas y columnas son aleatorios, puede hallarse la aproximación de *cualquier* probabilidad condicionada. De lo contrario, las únicas aproximaciones de probabilidades que pueden hallarse son aquellas en las que el investigador *fija* los tamaños de la muestra para los *sucesos dados.* La razón se explica en el Ejemplo 3.4.2

Ejemplo 3.4.2. Supongamos que se ha desarrollado un nuevo test doméstico para detectar el embarazo. Se realiza un experimento para hallar la aproximación de los coeficientes de falsos positivos y de falsos negativos del test. Para participar en el estudio se seleccionaron cinco mujeres que se sabía que estaban embarazadas y 10 mujeres no embarazadas. Se efectuó el nuevo test en cada una de ellas y el resultado se indica en la Tabla 3.3 (los datos son ficticios)

A partir de estos datos, es posible hallar una aproximación fiable de algunas probabilidades condicionadas mientras que para otras, no. Por ejemplo, en el caso de

P[está embarazada | test indica embarazo]

no puede hallarse la aproximación, pero para

P[test indica embarazo | está embarazada]

sí es posible. ¿Cuál es la diferencia entre ambos? Simplemente ésta: en la primera, se debe hallar la aproximación de la probabilidad de que una persona seleccionada aleatoriamente esté embarazada y el test indique embarazo a partir de los datos. Dado que el experimentador ha fijado en cinco el número dé mujeres embarazadas en el experimento, la probabilidad aproximada de este suceso ha sido forzada a ser como máximo de 5/15. Debido a esta limitación artificial, la probabilidad de que una persona cuyo test ha dado como resultado que está

Tabla 3.3

	Estado real		
	No embarazadas (realidad -)	**Embarazadas (realidad +)**	
Embarazadas (+)	5	1	6 (aleatorio)
No embarazadas (-)	5	4	9 (aleatorio)
	10 (fijo)	5 (fijo)	15 (fijo)

embarazada, lo esté realmente, no puede aproximarse a partir de este experimento. Sin embargo, para hallar la última, las cinco mujeres embarazadas seleccionadas aleatoriamente por el experimentador pueden entenderse como una muestra aleatoria de la población de todas las mujeres embarazadas. Por lo tanto, podemos utilizar el método de la frecuencia relativa para encontrar que la aproximación de la probabilidad de que una mujer embarazada dé positivo en el test es de $\frac{1}{5}$.

Hemos visto que los coeficientes de falsos positivos y de falsos negativos para un test de diagnóstico pueden aproximarse a partir de una tabla con los totales de todas las filas y columnas, los cuales pueden variar. El Ejemplo 3.4.2 muestra que también pueden aproximarse cuando los totales de las filas (o columnas) son fijos siempre que los totales fijos se refieran al número de sujetos realmente positivos y realmente negativos del estudio.

Riesgo relativo

Algunos estudios se diseñan para investigar un factor que el investigador cree que puede estar asociado con el desarrollo de una enfermedad o condición específica. Este factor se denomina factor *de riesgo.* Para realizar el estudio se seleccionan dos muestras. Una muestra, representada mediante *E,* consiste en sujetos que han estado expuestos al factor de riesgo; los demás, representados mediante *E',* no han estado expuestos al factor de riesgo.

En un momento dado, se clasifica cada sujeto según tenga la enfermedad, *D,* o no la tenga, *D'.* Hay dos probabilidades condicionadas de interés. Estas son: la probabilidad de que la enfermedad esté presente dado que el sujeto estuvo expuesto al riesgo, $P[D \mid E]$, y la probabilidad de que la enfermedad esté presente aunque el sujeto no estuvo expuesto al riesgo $P[D \mid E']$. Puesto que los tamaños muestrales para los sucesos E y E' son fijos, puede hacerse un cálculo aproximado de cada una de estas probabilidades utilizando la idea demostrada en el Ejemplo 3.4.2. Puede realizarse un cálculo aproximado de una medida del impacto del factor de riesgo a partir de estas probabilidades condicionadas. Esta medida, denominada *riesgo relativo* (RR), se calcula mediante:

$$\boxed{\text{RR} \cong \frac{P[D \mid E]}{P[D \mid E']}}$$

Recuérdese que, dado que las probabilidades utilizadas en el cálculo son aproximadas, el riesgo obtenido sólo es una estimación del riesgo relativo verdadero. Si RR = 1, significa que no existe asociación entre el factor de riesgo y el desarrollo de la enfermedad. Si RR > 1, se supone que un individuo expuesto al riesgo tiene más probabilidades de desarrollar la enfermedad que uno que no esté expuesto al riesgo. Un valor de RR < 1 significa que un individuo expuesto al riesgo es menos probable que desarrolle la enfermedad que uno no expuesto al riesgo. El Ejemplo 3.4.3 muestra esta idea.

Ejemplo 3.4.3. Se ha realizado un estudio sobre la edad de la madre en el momento del nacimiento de su hijo como factor de riesgo en el desarrollo del síndrome de la muerte súbita del lactante (SMSL). Se seleccionaron para el estudio un total de 7330 mujeres que estaban por debajo de los 25 años en el momento del nacimiento del niño. De ellas, 29 tuvieron niños afectados de SMSL. De las 11 256 mujeres seleccionadas para el estudio que tenían 25 años o más en el momento del nacimiento de sus hijos, 15 tuvieron niños con SMSL. Estos datos se muestran en la Tabla 3.4. Partiendo de esta tabla podemos observar que

$$P[D \mid E] = \frac{29}{7330} \quad \text{y} \quad P[D \mid E'] = \frac{15}{11\,256}$$

Tabla 3.4. La edad como factor de riesgo del desarrollo de SMSL

		SMSL		
		Sí	No	
Edad	Menos de 25 años	29	7 301	7 330 (fijo)
	25 años o más	15	11241	11 256 (fijo)

El riesgo relativo estimado es

$$\text{RR} \cong \frac{P[D \mid E]}{P[D \mid E']} = \frac{29/7330}{15/11\,256} = 2.96$$

Podemos sacar la conclusión de que un niño de una madre joven (menos de 25 años) tiene aproximadamente 2.96 veces más probabilidades de sufrir un SMSL que uno nacido de una madre de más edad. (Basado en los datos registrados por Norman Lewak, Bea van der Berg y Bruce Beckwith, en «Sudden Infant Death Syndrome Risk Factors: Prospective Data Review», *Clinical Pediatrics,* vol. 18, 1979, págs. 404-411.)

Dado que se puede hacer un cálculo aproximado de $P[D \mid E]$ y $P[D \mid \text{E}']$ a partir de los datos de las tablas, ya que todos los totales por filas y columnas son aleatorios, puede hacerse un cálculo aproximado del riesgo relativo a partir de dichas tablas. El Ejercicio 14 de esta sección es un ejemplo de ello.

EJERCICIOS 3.4

1. En un estudio de 300 pares de gemelos se planteaba la cuestión de si eran realmente idénticos. Se consideraban indicadores tales como los grupos sanguíneos ABO, MN o el factor Rh. Basándose en estos indicadores, los gemelos se clasificaban en idénticos (+) o no idénticos (-). La última clasificación realizada se consideraba correcta. El propósito del estudio es averiguar la capacidad de los gemelos para autoclasificarse. Los resultados se muestran en la Tabla 3.5. Los datos marginales se obtienen por medio de un proceso aleatorio. Calcular aproximadamente los coeficientes de falsos positivos y de falsos negativos del procedimiento de autoclasificación.
2. Se proyecta un estudio para conocer la asociación entre color y olor en azaleas silvestres de los montes Great Smoky. Se selecciona un área de 5 acres de terreno y se encuentra que contiene 200 brotes de esta planta. Cada uno de ellos se clasifica en función de que tenga o no color y presencia o ausencia de olor. Los resultados se muestran en la Tabla 3.6. Haciendo uso de estos datos, aproximar, si es posible, cada una de las siguientes

Tabla 3.5

	Clasificación verdadera		
Autoclasificación	**No idénticos (-)**	**Idénticos (+)**	
+	**12**	**54**	
	130	**4**	
			200

Tabla 3.6

		Color		
		Sí	No	
Fragancia	Sí	12	118	
	No	50	20	
				200

probabilidades. Si no fuera posible aproximar a partir de estos datos alguna probabilidad en particular, explicar por qué.

a) P[una azalea seleccionada aleatoriamente tenga olor].
b) P[una azalea seleccionada aleatoriamente tenga color].
c) Pfuna azalea seleccionada aleatoriamente tenga color y olor].
d) P[una azalea seleccionada aleatoriamente tenga color dado que tiene olor].
e) P[una azalea seleccionada aleatoriamente tenga olor dado que tiene color].

3. Los resultados descritos en la Tabla 3.7 se obtuvieron en un estudio diseñado para averiguar la capacidad de un cirujano anatomopatólogo para codificar correctamente biopsias quirúrgicas en malignas o benignas. Aproximar α y β a partir de estos datos.
4. Se ha realizado un estudio para poner a prueba un procedimiento de detección de enfermedades renales en pacientes con hipertensión. Aplicando el nuevo procedimiento, los experimentadores detectan 137 pacientes hipertensos. A continuación se determinó de nuevo la presencia o ausencia de enfermedad renal por otro método. Los datos obtenidos se recogen en la Tabla 3.8. Utilizando estos datos, aproximar los coeficientes de falsos positivos y de falsos negativos del test.
5. *Definición:* La *especificidad* de un test es la probabilidad de que el resultado del test sea negativo supuesto que el sujeto sea ciertamente negativo. Aproximar la especificidad del test del Ejercicio 1. En general, ¿es de desear que la especificidad de un test sea alta, o baja? Explicarlo.

	Estado real		
Informe del anatomopatólogo	Benigno (-)	Maligno (+)	
+	7	79	
–	395	19	
			500

Tabla 3.8

	Estado real		
Enfermedad detectada	Enfermedad ausente (-)	Enfermedad presente (+)	
Sí(+)	23	44	
No(-)	60	10	
			137

6. *Definición:* La *sensibilidad* de un test es la probabilidad de que dicho test conduzca a un resultado positivo supuesto que el sujeto sea efectivamente positivo. Aproximar la sensibilidad del test del Ejercicio 1. En general, ¿es de desear que la sensibilidad de un test sea alta, o baja? Explíquese.
7. Se sometió a 100 pacientes y 75 sujetos normales a un test de diagnóstico de la orina. En un 60 % de los casos, el diagnóstico fue positivo. Hubo también ocho falsos negativos. ¿Cuál es el coeficiente de falsos positivos aproximado?
8. Aproximar la especificidad y la sensibilidad del test del Ejemplo 3.4.1. En general, ¿qué relación existe entre la especificidad y el coeficiente de falsos positivos? ¿Qué relación existe entre la sensibilidad y el coeficiente de falsos negativos?
9. Se ha realizado un estudio de una técnica de inmunoensayo de enlace de enzimas (EIA) para examinar a donantes de sangre con el fin de detectar anticuerpos frente al VIH. Los sujetos se someten a la técnica EIA, y la presencia o ausencia de anticuerpos se confirma en una fecha posterior. En la Tabla 3.9 se proporcionan los datos.
   a) Estimar el coeficiente de falsos positivos del test. Utilizarlo para hallar la especificidad de la prueba.
   b) Estimar el coeficiente de falsos negativos del test. Utilizarlo para hallar la sensibilidad del test. (Basado en la información hallada en Richard Eisenstaedt y Thomas Getzen, «Screening Blood Donors for HIV Antibody: Cost Benefit Analysis», *American Journal of Public Health,* vol. 78, núm. 4, abril de 1988, págs. 450-4541)
10. El *valor predictivo positivo* de un test se define como la probabilidad de que un individuo sea realmente positivo dado que el resultado del test ha sido positivo. Puede hallarse el valor aproximado a partir de una tabla en la que todos los totales por filas y columnas sean susceptibles de variar. Hallar la aproximación del valor predictivo positivo del test de autoclasificación del Ejercicio 1.
11. El *valor predictivo negativo* de un test se define como la probabilidad de que un individuo sea realmente negativo dado que el resultado del test ha sido negativo. Puede hallarse el valor aproximado a partir de una tabla en el que todos los totales por filas y columnas sean susceptibles de variar. Hallar la aproximación del valor predictivo negativo del test de autoclasificación del Ejercicio 1.
12. Hallar la aproximación de los valores predictivos positivo y negativo del test para la enfermedad renal del Ejercicio 4.
13. Se ha realizado un estudio para determinar los síntomas clínicos que ayudan a la identificación de la tos ferina. Un síntoma investigado es la tos aguda de cualquier duración. Los datos obtenidos sobre 233 niños estudiados se muestran en la Tabla 3.10. Hallar lia aproximación del coeficiente de falsos positivos y el valor real positivo del test. ¿Puede parecer que sólo la presencia de tos aguda es un buen indicador de la presencia de la tos ferina? Explicarlo. (Basado en la información hallada en Peter Patriaca et al., «Sensiti-

Tabla 3.9

	Estado real	
Test EIA	Anticuerpos ausentes (-)	Anticuerpos presentes (+)
+	1000	30
–	98 969	1
		100 000

Tabla 3.10

Tos presente	Estado real		
	Tos ferina ausente (-)	**Tos ferina presente (+)**	
Sí (+)	**83**	116	
No (-)	**32**	**2**	
			233

vity and Specifity of Clinical Case Definition of Pertussis», *American Journal of Public Health,* vol. 78, núm. 7, julio de 1988, págs. 833-835.)

14. En 1985, muchas familias estadounidenses adoptaron a niños asiáticos. Algunos de estos niños habían estado expuestos al virus de la hepatitis B y eran hipotéticos transmisores del virus a otros. En un estudio del riesgo implicado, se obtuvieron los datos de la Tabla 3.11. Los valores de las celdas representan el número de familiares cercanos a los que se les ha detectado el virus y todos los totales por filas y columnas son aleatorios. Hallar la aproximación del riesgo relativo. (Basado en la información hallada en Andrew Friede et al., «Transmission of Hepatitis B Virus from Adopted Asian Children to Their American Families», *American Journal of Public Health,* vol. 78, núm. 1, enero de 1988, págs. 26-29.)
15. Se sabe que los pacientes con SIDA a menudo presentan tuberculosis. Se llevó a cabo un estudio de los factores de riesgo asociados con el desarrollo de esta enfermedad en los pacientes. Uno de los factores considerados fue la adicción a drogas intravenosas. De los 1992 pacientes del estudio, 307 habían abusado de las drogas por vía intravenosa. Cuarenta y seis de los pacientes tenían tuberculosis y, de ellos, 11 eran adictos a drogas por vía intravenosa. (Basado en los datos registrados en Timothy Cote et al., «The present and the Future of AIDS and Tuberculosis in Illinois», *American Journal of Public Health,* vol. 80, núm. 8, agosto de 1990, págs. 950-953.)
    a) Construir una tabla de 2 x 2 para visualizar estos datos.
    b) Hallar e interpretar el riesgo relativo.
16. En un estudio sobre la relación entre el uso regular de tinte para el cabello y el desarrollo de la leucemia, fueron seleccionados 577 pacientes con leucemia y 1245 personas sin la enfermedad (controles) y fueron consultados en relación con el uso de dicho tinte. Cuarenta y tres pacientes y 55 controles dijeron haber estado bastante expuestos al tinte. (Basado en la información hallada en Kenneth Cantor et al., «Hair Dye Use and Risk of Leukemia and Lymphoma», *American Journal of Public Health,* vol. 78, núm. 5, mayo de 1988, págs. 570-571.)
    a) Completar la Tabla 3.12.

Tabla 3.11

		Virus presente	
		Sí	**No**
Expuesto al riesgo	Sí	**7**	**70**
	No	**4**	228

Tabla 3.12

		Leucemia presente	
		Sí	**No**
Utiliza tinte para cabello	Sí	43	55
	No		
		577 (fijos)	1245 (fijos)

b) En este caso, ¿es posible hallar la aproximación del riesgo relativo usando la definición dada en esta sección? Explicarlo.

c) Es posible hacerse una idea de la repercusión del uso de tinte para cabello, considerando el cociente

$$\frac{P[E \mid D]}{P[E \mid D']}$$

donde E es el suceso de que el individuo estuvo expuesto al riesgo y D es el suceso de que la leucemia está presente. ¿Puede estimarse cada una de las probabilidades condicionadas implicadas en este cociente? Si es así, evaluarlo e interpretarlo.

3.5. INDEPENDENCIA

Pueden existir, fundamentalmente, dos relaciones entre sucesos. La primera, ser mutuamente excluyentes, ha sido tratada en la Sección 3.1; la segunda, ser *independientes,* se expone en ésta. El término matemático tiene prácticamente el mismo significado que el lingüístico Webster define objetos *independientes* como objetos que actúan «con independencia el uno del otro». De este modo, dos sucesos son independientes si uno puede producirse con independencia del otro. Es decir, la realización o no realización de uno no tiene efecto alguno sobre la realización o no del otro. En numerosos casos, podemos determinar sobre una base puramente intuitiva, si dos sucesos son independientes. Por ejemplo, los sucesos A_1, el paciente tiene sinovitis, y A_2, el paciente tiene apendicitis, son intuitivamente independientes. El hecho de que el paciente tenga apendicitis nada tiene que ver con que padezca o no sinovitis y viceversa.

En algunos casos, no obstante, la delimitación no es tan evidente. Necesitamos entonces una definición matemática precisa del concepto, para poder determinar sin la menor duda si dos sucesos son, de hecho, independientes. La definición es fácil de justificar. Por ejemplo, supongamos que, basándonos en los síntomas descritos, podamos admitir que la probabilidad de que un paciente tenga apendicitis sea de 0.9 (A_2). Supongamos que se nos dé ahora la información adicional de que el paciente tiene sinovitis (A_1). ¿Cuál es la probabilidad de que el paciente tenga apendicitis? ¡Obviamente, la respuesta sigue siendo 0.9! Dado que A_x y A son independientes, la nueva información es irrelevante y no afecta para nada a la probabili dad original. De este modo la independencia entre dos sucesos A_1 y A_2 implica que la probabi lidad condicionada $P[A_1 \mid A_2]$ ha de ser igual a la asignada originalmente a A_2. Esta caracterización se adopta como definición del término *sucesos independientes.*

Definición 3.5.1. Sucesos independientes. Sean A, y A_2 dos sucesos tales que $P[A_1] \neq 0$. Estos sucesos son *independientes* si y sólo si

$$P[A_2 \mid A_1] = P[A_2]$$

Ejemplo 3.5.1. Se estima que entre la población total de Estados Unidos, el 55 % padece obesidad (A,), el 20% es hipertenso (A_2) y el 60% es obeso o hipertenso. ¿Es, de hecho, independiente el que una persona sea obesa de que padezca hipertensión? La respuesta a esta pregunta no es obvia. Haciendo uso del principio general de la adición, se tiene

$$P[\text{A}_1 \text{ y } \text{A}_2] = P[A_1] + \text{P}[\text{A}_2] - \text{P}[\text{A}_1 \text{ o } \text{A}_2]$$

En este caso

$$P[A_1 \text{ y } \text{A}_2] = 0.55 + 0.20 - 0.60 = 0.15$$

Así que

$$P[A_2 \mid A_1] = \frac{P[A_1 \text{ y } A_2]}{P[A_1]}$$

$$= \frac{0.15}{0.55} = \frac{15}{55} = 0.27$$

Puesto que $\text{P}[A_2|\text{A}_1] = 0.27 \neq 0.20 = \text{P}[\text{A}_2]$, puede concluirse que los sucesos no son independientes. Hablando en términos prácticos, el hecho de que una persona tenga exceso de peso aumenta la probabilidad de que tenga hipertensión.

Obsérvese que estamos suponiendo que las probabilidades del Ejemplo 3.5.1 se basan en datos de la población y, por lo tanto, son exactas. Aquí puede utilizarse esta Definición 3.5.1 para probar la independencia de dos sucesos. En la práctica, esta situación surge muy raras veces. En cambio, habitualmente estaremos tratando con probabilidades de frecuencias relativas obtenidas de muestras extraídas de la población. En este caso, no puede utilizarse la Definición 3.5.1 para probar la independencia. No obstante, en el Capítulo 12 se desarrollará un test apropiado para las muestras.

La Definición 3.5.1 es lógica y fácil de comprender. No obstante, no es la que normalmente se emplea para el término *sucesos independientes.* La definición usual puede derivarse de lo siguiente:

$P[A_2 \mid A_1] = \dfrac{P[A_1 \text{ y } A_2]}{P[A_1]}$ es siempre cierta en tanto que $P[A_1] \neq 0$

$P[A_2 \mid A_1] = P[A_2]$ si $P[A_1] \neq 0$ y los sucesos son independientes.

Así que, *si A_1 y A_2* son independientes, ambas ecuaciones se verifican simultáneamente. Tenemos pues para *$P[A_2| A_1]$,* dos expresiones que conducen a

$$\frac{P[A_1 \text{ y } A_2]}{P[A_1]} = P[A_2]$$

Multiplicando ambos miembros de la ecuación obtenida por P[A], obtenemos: *$P[A_1$* y *$A_2]$* = $= \text{P}[\text{A}_1]\text{P}[\text{A}_2]$, que es la definición usual del término *sucesos independientes.*

Definición 3.5.2. Sucesos independientes. Sean A_1 y A_2 dos sucesos. *A_1 y A_2* son *independientes* si y sólo si *$P[A_1$* y $\text{A}_2] = P[A_1]\, P[A_2]$.

Obsérvese que cuando los sucesos son independientes la probabilidad de que ambos ocurran simultáneamente se obtiene por multiplicación. Así, como se indicó en la Sección 3.2, la palabra y es la clave de que las probabilidades se deben multiplicar.

Ejemplo 3.5.2. Estudios de genética de poblaciones indican que el 39 % de los genes que gobiernan la información del factor Rh determinan que éste sea negativo. Basándose en ello, ¿cuál es la probabilidad de que un individuo seleccionado aleatoriamente tenga Rh negativo? El factor Rh negativo se presenta si, y sólo si, el individuo implicado posee dos genes determinativos de información negativa. Dado que cada gen se hereda de uno de los padres, puede suponerse que el tipo de cada uno de los genes es independiente del otro. Por lo tanto, la probabilidad de que un individuo tenga dos genes negativos es $(0.39)(0.39) \cong 0.15$. (Basado en la información de William Keeton y Carol McFadden, *Elements of Biological Science* W.W. Norton, Nueva York, 1983.)

La idea de independencia puede extenderse a más de dos sucesos. Un conjunto de sucesos se dice que es independiente siempre que cualquier subconjunto de sucesos satisfaga la propiedad de que la probabilidad de la aparición simultánea sea igual al producto de las probabilidades individuales de cada suceso. El Ejemplo 3.5.3 demuestra esta idea en el contexto de un problema que utiliza un diagrama en árbol. Obsérvese que ahora estamos en disposición de calcular probabilidades de trayectorias en el caso de que éstas no sean equiprobables.

Ejemplo 3.5.3. A lo largo de un día, se pone a prueba un determinado diagnóstico con tres pacientes que no guardan relación alguna entre ellos. El diagnóstico es fiable en un 90 % de los casos tanto cuando se da como cuando no se da la condición para cuya detección se ha diseñado la prueba. ¿Cuál es la probabilidad de que exactamente dos de los tres resultados de la prueba sean erróneos?

Un diagrama de árbol nos ayudará a resolver la cuestión. En él, C representa una decisión correcta, y E un error. En la Figura 3.11 aparecen las probabilidades correspondientes a cada alternativa. Las ramas representan alternativas interesantes. Cada trayectoria completa representa la realización simultánea de tres sucesos diferentes. Por ejemplo, la trayectoria *EEC* representa que se dieron simultáneamente, un error con el primer paciente (E_1), un error con el segundo (E_2) y una decisión correcta con el tercero (C_3). Puesto que las pruebas se llevan a cabo sobre pacientes diferentes y con independencia las unas de las otras, podemos suponer que los resultados son independientes. De acuerdo con la Definición 3.5.2, la probabilidad a lo largo de cada trayectoria se calculará *multiplicando* las probabilidades que aparecen en el recorrido. Así, en este caso, se tendrá $P[E_1 \text{ y } E_2 \text{ y } C_3] = P[E_1]P[E_2]P[C_3] = (0.1)(0.1)(0.9) = 0.009$. Dado que hay tres trayectorias en las que aparecen exactamente dos errores, la probabilidad de obtener exactamente dos errores en cualquier orden es $3(0.009) = 0.027$.

La Definición 3.5.2 debe utilizarse con cuidado. Se debe estar seguro de que es razonable suponer que los sucesos son independientes antes de aplicar la definición para calcular la probabilidad de que se produzca una serie de sucesos. En el Ejemplo 3.5.4 se ilustra el peligro de una independencia erróneamente supuesta.

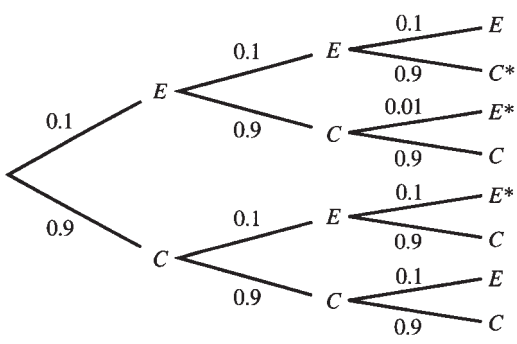


Figura 3.11. Resultados de una prueba de diagnóstico (tres pacientes).

Ejemplo 3.5.4. Un estudio del Comité de Energía Atómica, WASH 1400, informó de que la probabilidad de que se produzca un accidente nuclear, parecido al que ocurrió en Three Mile Island en marzo de 1978, es de 1 en 10 millones. No obstante, el accidente se produjo. Según Mark Stephens, «la metodología del WASH 1400 utilizó árboles de sucesos, secuencias de acciones que eran necesarias para que se produjera el accidente. Estos árboles de sucesos no suponían ninguna interrelación entre sucesos, que podían haber sido causados por el mismo error de juicio o como parte de la misma acción equivocada. Los profesionales de la estadística que asignaron las probabilidades al proyecto WASH 1400 dijeron, por ejemplo, que existía un riesgo del 1 por 1000 de que una de las válvulas auxiliares de control de entrada de agua (de las doce) estuviera cerrada. Y si existe una probabilidad de que dos válvulas estén cerradas, es una milésima parte de ésta, es decir, de una en un millón. Pero las doce fueron cerradas por la misma persona el 26 de marzo y una nunca fue cerrada con la otra». Los sucesos A_1: la primera válvula está cerrada, y A_2: la segunda válvula está cerrada, no eran independientes. Sin embargo, fueron tratados como tales al calcular la probabilidad de un accidente. Esto, entre otras cosas, llevó a una subestimación del potencial de accidentalidad (de *Three Mile Island* de Mark Stephens, Random House, 1980).

EJERCICIOS 3.5

1. ¿Cuáles de los siguientes pares de sucesos cree usted que son independientes? ¿Cuáles son mutuamente excluyentes?
   A_1. Una madre padece rubéola durante los tres primeros meses de embarazo.
   B_1. Un hijo nace muerto o deforme.
   A_2: Un hombre es estéril.
   B_2: Un hombre padece parotiditis en la edad adulta.
   A_3: Una rata hembra y una rata macho son enjauladas juntas.
   B_3: La rata hembra es estéril.
   A_4: Un niño es miope.
   B_4: Un niño es hipermétrope.
   A_5: Un terreno ha sido drenado.
   B_5: El terreno experimenta frecuentes inundaciones.
   A_6: Un conejo es inoculado con virus de poliomielitis.
   B_6: La sangre del conejo contiene anticuerpos de poliomielitis.
   A_7: Un conejo es inoculado con virus de poliomielitis.
   B_7: La sangre del conejo contiene anticuerpos de sarampión.
2. Argumentar, de forma intuitiva, que si dos sucesos que no son imposibles son mutuamente excluyentes, no pueden ser independientes. Probarlo matemáticamente. *Sugerencia:* demostrar que, con estas condiciones, no se puede satisfacer la Definición 3.5.2.
3. Los contaminantes más comunes de las aguas son de origen orgánico. Puesto que la mayor parte de los materiales orgánicos se descompone por acción de bacterias que requieren oxígeno, un exceso de materia orgánica puede significar una disminución en la cantidad de oxígeno disponible. Ello afecta eventualmente a otros organismos presentes en el agua. La demanda de oxígeno por parte de una bacteria se llama *demanda biológica de oxígeno* (DBO). Un estudio de las corrientes acuáticas que circulan en las proximidades de un complejo industrial revela que el 35 % tiene una alta DBO, el 10 % muestra una acidez elevada y un 4 % presenta ambas características. ¿Son independientes los sucesos «la corriente tiene una alta DBO» y «la corriente posee una acidez elevada»? Calcular la probabilidad de que la corriente tenga una acidez elevada, dado que presenta una alta DBO.
4. El 50 % de la población aproximadamente corresponde a varones, el 68 % bebe con cierto exceso, y el 38.5 % bebe y es varón. Dado que una determinada persona aleatoria-

mente seleccionada es varón, hallar la probabilidad de que beba. ¿Es la condición de bebedor independiente del sexo?

5. La probabilidad de contraer hepatitis a partir de una unidad de sangre, es de 0.01. Un paciente recibe dos unidades de sangre durante su estancia en un hospital. ¿Cuál es la probabilidad de que no contraiga hepatitis como consecuencia de ello?
6. Aunque el tétanos es infrecuente en Estados Unidos, es mortal en el 70 % de los casos. Si tres personas contraen el tétanos en el período de un año, ¿cuál es la probabilidad de que mueran al menos dos de los tres? *(Sugerencia:* Úsese un diagrama de árbol.)
7. Considere el diagrama de árbol del Ejercicio 5, Sección 2.2. Supongamos que las probabilidades de que un paciente tenga dolor de cabeza, fiebre, malestar corporal o dolor muscular son de 0.7, 0.8, 0.1 y 0.2, respectivamente. Supongamos también que las apariciones de estos síntomas son independientes entre sí.
   a) Determinar la probabilidad para cada una de las 16 trayectorias del árbol.
   b) Determinar la probabilidad de que el diagnóstico pueda ser alergia sistémica a los alimentos.
   c) Determinar la probabilidad de que el diagnóstico pueda ser gripe.
   d) Determinar la probabilidad de que el diagnóstico no pueda ser ni alergia a lo: alimentos ni gripe.
8. *Principio de Hardy-Weinberg.* El principio de Hardy-Weinberg, en genética de poblaciones, recibió este nombre de G. H. Hardy, un matemático inglés, y de G. Weinberg, un médico alemán. Este principio establece básicamente que una población es genéticamente estable en las sucesivas generaciones. Los fundamentos matemáticos de este principio se apoyan sobre la noción de independencia en dos aspectos: apareamiento independiente y herencia independiente por parte de los hijos del gen de cada padre. Consideremos la distribución de un simple par de genes *A y a.* Cada miembro de la población portará dos de estos genes. Tendremos, pues, tres genotipos diferentes: *AA, Aa y aa.* Supongamos que estos genotipos están presentes en la población en las proporciones $\frac{1}{4}$**AA,** $\frac{1}{2}$**Aa,** $\frac{1}{4}$**aa.** Si admitimos que los miembros de la población se aparean aleatoriamente, habría nueve posibles tipos de cruzamiento, que son los que aparecen en la Tabla 3.13. Cada tipo de cruzamiento induce uno o más genotipos en la descendencia. Dada la independencia, las primeras filas son como aparece en la tabla. Completar la tabla. Una vez hecho, comprobar que un cuarto de la descendencia es de genotipo *AA,* la mitad de *Aa y* un cuarto de *aa,* tal como asegura el principio de Hardy-Weinberg.
9. Algunos caracteres en animales se dice que están sexualmente influenciados. Por ejemplo, la aparición de cornamenta en la oveja está gobernada por un par de alelos, *H y h.* El alelo *H* para la presencia de cornamenta es dominante en los machos, pero recesivo en las hembras. El alelo *h* para la ausencia de cornamenta es dominante en las hembras, pero recesivo en los machos. Por tanto, dados un macho y una hembra heterocigotos *(Hh),* el macho tendrá cornamenta y la hembra no. Supongamos que tales animales se aparean.
   a) Dibujar un diagrama de árbol para representar los posibles genotipos relativos a la aparición de cornamenta.
   b) Supongamos que cada cría de este cruce tenga exactamente la misma posibilidad de ser macho que de ser hembra. Calcular la probabilidad de que dada una cría, sea macho y tenga cornamenta. Calcular la probabilidad de que dada una cría, sea hembra y tenga cornamenta.
   c) Hallar la probabilidad de que una cría dada tenga cornamenta. Demostrar que el suceso *A,* la cría es macho y *B,* la cría tiene cornamenta, no son independientes.

Tabla 3.13

Tipo de apareamiento		Probabilidad de cruzamiento	Genotipo filial posible	Genotipo filial probable	Probabilidad de la trayectoria
Varón	**Mujer**				
AA	***AA***	$\frac{1}{4} \cdot \frac{1}{4}$	***AA***	1	$\frac{1}{16}$
AA	***Aa***	$\frac{1}{4} \cdot \frac{1}{2}$	***AA***	$\frac{1}{2}$	$\frac{1}{16}$
			Aa	$\frac{1}{2}$	$\frac{1}{16}$
AA	***aa***	$\frac{1}{4} \cdot \frac{1}{4}$	***Aa***	1	$\frac{1}{16}$
Aa	***AA***				
Aa	***Aa***				
Aa	***aa***				
aa	***AA***				
aa	***Aa***				
aa	***aa***				

10. Verificar que la probabilidad de que un individuo, seleccionado aleatoriamente, sea homocigoto Rh positivo (++), es aproximadamente 0.37 y que la probabilidad de que sea heterocigoto Rh positivo (+ - o - +) es aproximadamente 0.48.
11. El grupo sanguíneo de un individuo (A, B, AB, 0) es independiente del factor Rh.
    a) Determinar la probabilidad de que un individuo seleccionado aleatoriamente sea del grupo AB negativo dado que dicho individuo es un norteamericano de raza blanca (véanse los Ejemplos 3.5.2 y 3.1.2).
    b) Determinar la probabilidad de que un individuo seleccionado aleatoriamente sea del grupo AB negativo dado que dicho individuo es un norteamericano de raza negra (véase el Ejemplo 3.5.2 y el Ejercicio 12 de la Sección 3.1).
    c) ¿El hecho de tener sangre del grupo AB negativo es independiente del grupo racial, blanco o negro, al que pertenece el individuo? Explicarlo.
    d) ¿El hecho de tener sangre del grupo A negativo es independiente del grupo racial, blanco o negro, al que pertenece el individuo? Explicarlo.
12. Considerar el riesgo relativo definido en la Sección 3.4. Probar que si RR = 1, los sucesos *D,* la enfermedad está presente, y *E,* el paciente está expuesto a riesgo, son independientes. *Sugerencia:* Establezcamos que *P[D |E]* es igual a *P[D* | E'] y apliquemos la definición de probabilidad condicionada a cada lado de la ecuación. Recordemos que *P[E'] = 1 - P[E].* Demostrar que *P[D* y *E] = P[D]P[E].*
13. Un médico solicita 10 pruebas de diagnóstico independientes para que sean realizadas en un mismo paciente. El coeficiente de falsos positivos de cada test es 0.05. ¿Cuál es la probabilidad de que al menos se obtenga un resultado positivo erróneo?
14. Si el coeficiente de falsos positivos de cada test de un grupo de tests es 0.05, ¿cuántos tests independientes pueden incluirse en el grupo si deseamos que la probabilidad de obtener al menos un resultado falso positivo sea como máximo 0.20?

3.6. LA REGLA DE LA MULTIPLICACIÓN

Podemos ahora calcular $P[A_1$ y $A_2]$, si los sucesos son independientes. Además, si la información de que disponemos lo permite, es posible hacerlo mediante la regla general de la adición.

¿Existe algún otro procedimiento para hallar la probabilidad de que dos sucesos se produzcan simultáneamente, si éstos no son independientes? La respuesta es afirmativa, y el método utilizado es fácil de deducir. Sabemos que

$$P[A_2 \mid A_1] = \frac{P[A_1 \text{ y } A_2]}{P[A_1]}$$

al margen de que los sucesos sean o no independientes. Multiplicando cada miembro de la igualdad por $P[A_1]$ obtenemos la fórmula siguiente, llamada *regla de multiplicación:*

$$P[A_1 \text{ y } A_2] = P[A_2 \mid A_1]P[A_1] \quad \text{regla de la multiplicación}$$

Su utilización se describe en el Ejemplo 3.6.1.

Ejemplo 3.6.1. La denominada prospección geobotánica se basa en el estudio de las plantas que aparecen en depósitos de minerales. Una pequeña planta de menta con una flor de color malva es un indicador del cobre. Supongamos que, en una región dada, existe un 30 % de probabilidad de que el suelo tenga un alto contenido de cobre y un 23 % de que la menta esté presente en ese lugar. Si el contenido de cobre es alto, existe un 70 % de probabilidad de que la menta esté presente. ¿Cuál es la probabilidad de que el contenido de cobre sea alto y de que esté presente la menta? Si representamos con A_1 el suceso de que el contenido de cobre sea alto y con A_2 el suceso de que la menta esté presente deberemos determinar $P[A_1 \text{ y } A_2]$. Tenemos que $P[A_1] = 0.30$, $P[A_2] = 0.23$, y $P[A_2 \mid A_1] = 0.70$. Mediante la regla de la multiplicación

$$\begin{aligned} P[A_1 \text{ y } A_2] &= P[A_2 \mid A_1]P[A_1] \\ &= 0.70(0.30) \\ &= 0.21 \end{aligned}$$

En el Ejemplo 3.6.2 se ilustra el uso de la regla de la multiplicación en genética.

Ejemplo 3.6.2. Si una madre es Rh negativo y su hijo es Rh positivo, existe una incompatibilidad sanguínea que puede conducir a una eritroblastosis fetal, consistente en que la madre crea un anticuerpo contra el Rh del feto que conduce a la destrucción de los hematíes del feto. ¿Cuál es la probabilidad de que un niño seleccionado aleatoriamente corra este riesgo?

Una forma de que el niño tenga este problema es que el padre sea heterocigoto Rh positivo (+ - o - +) y pase un gen positivo al niño mientras que la madre sea Rh negativo. Para determinar la probabilidad de esta combinación de sucesos, debemos hallar $P[(A_1 \text{ y } A_2)$ y $A_3]$ donde A, representa que el padre sea heterocigoto Rh positivo, A_2 que el padre trasmita un gen positivo al niño y A_3 que la madre sea Rh negativo. Obsérvese que los sucesos A, y A_2 no son independientes. El hecho de que el padre sea heterocigoto Rh positivo está presente en la posibilidad de que el niño obtenga un gen positivo de esta fuente. A través de la regla de la multiplicación,

$$P[A_1 \text{ y } A_2] = P[A_2 \mid A_1]P[A_1]$$

Por el Ejercicio 10 de la Sección 3.5, sabemos que $P[A_1] \cong 0.48$. Dado que un gen se hereda aleatoriamente del padre, $P[A_2 \mid A_1] = 0.5$. Por lo tanto

$$P[A_1 \text{ y } A_2] \cong 0.5(0.48) = 0.24$$

Puesto que el genotipo de la madre no tiene efecto sobre el padre o sobre su capacidad de transferir un gen positivo al niño, A_3 es independiente de A_1 y A_2. Por el Ejemplo 3.5.2 sabemos que $P[A_3] \simeq 0.15$. Así pues, por la definición de independencia,

$$P[(A_1 \text{ y } A_2) \text{ y } A_3] \cong 0.24(0.15) = 0.0360$$

Existen otras formas de que esté presente la condición. El Ejercicio 1 lo señala y permite calcular la probabilidad de que un niño contraiga el problema de cualquier procedencia.

EJERCICIOS 3.6

1. Un niño tendrá eritroblastosis fetal si la madre es Rh negativo y el padre es homocigoto Rh positivo (++). Utilizar la información del Ejercicio 10 de la Sección 3.5 para hallar la probabilidad de que esto ocurra. Determinar la probabilidad de que un niño seleccionado aleatoriamente tenga la condición, combinando este resultado con el obtenido en el Ejemplo 3.6.2.
2. Ciertos estudios indican que el 82 % de los profesionales varones bebe. De los que beben, el 18% corresponde a grandes bebedores. ¿Cuál es la probabilidad de que, seleccionando aleatoriamente a un profesional, beba y sea un gran bebedor?
3. De todos los pacientes de cáncer, en el 52% son mujeres. El 40% de los pacientes sobrevive al menos cinco años desde el momento del diagnóstico. No obstante, esta tasa de supervivencia es válida solamente para el 35 % de las mujeres. ¿Cuál es la probabilidad de que un paciente de cáncer seleccionado aleatoriamente sea mujer y sobreviva, al menos, cinco años?
4. La probabilidad de que una unidad de sangre proceda de un donante remunerado es 0.67. Si el donante es remunerado, la probabilidad de que la unidad contenga el suero de la hepatitis es 0.0144. Si el donante es desinteresado, esta probabilidad es 0.0012. Un paciente recibe una unidad de sangre. ¿Cuál es la probabilidad de que contraiga hepatitis como consecuencia de ello?
5. El 2% de la población en general padece diabetes. De ellos, solamente la mitad lo sabe. Si se selecciona aleatoriamente a un individuo. ¿Cuál es la probabilidad de que padezca diabetes pero no sea consciente de padecerla?
6. Se sabe que el coeficiente de falsos positivos de un test para una determinada enfermedad es del 4 % y que el coeficiente de falsos negativos es del 6 %. El test muestra que el 15 % de las personas da positivo. ¿Cuál es la probabilidad de que un individuo aleatoriamente seleccionado tenga efectivamente la enfermedad? *Sugerencia:* Sea x = P[realmente positivo] y $1 - x$ = P[realmente negativo]. Obsérvese que

$$\begin{aligned} \text{P[test positivo]} = {} & \text{P[test positivo y realmente positivo]} \\ & + \text{P[test positivo y realmente negativo]} \end{aligned}$$

7. En la replicación del DNA, a veces se presentan errores que pueden dar lugar a mutaciones observables en el organismo. En ocasiones, tales errores están inducidos químicamente. Se expone un cultivo de bacterias a la presencia de un producto químico que tiene un 0.4 de probabilidad de inducir a error. Sin embargo, el 65 % de los errores es «silencioso», en el sentido de que no dan lugar a una mutación observable. ¿Cuál es la probabilidad de que se observe una colonia mutada? *Sugerencia:* Hallar P[error y observable].
8. En la ciencia es importante la capacidad de observar y recordar datos. Desgraciadamente, el poder de la sugestión puede distorsionar la memoria. Se realizó un estudio

sobre los recuerdos: a los sujetos se les muestra una película en la que un coche pasa por una carretera. En la película no sale ningún granero. A continuación, se les formula una serie de preguntas relacionadas con la película a los sujetos. A la mitad se les preguntó: «¿Con qué velocidad se mueve el coche cuando pasa por el granero?» A la otra mitad de los sujetos no se les hizo esta pregunta. Más tarde, se le preguntó a cada uno de ellos: «¿Sale algún granero en la película?» Entre los que se les formuló la primera pregunta concerniente al granero, el 17 % respondió «sí»; sólo el 3 % de los restantes respondió «sí». ¿Cuál es la probabilidad de que un participante en este estudio, seleccionado aleatoriamente, haya dicho ver el granero inexistente? ¿Decir que se ha visto el granero es independiente de que se le haya formulado la primera pregunta sobre el mismo? Sugerencia:

$$\begin{aligned} P[\text{sí}] = {} & P[\text{sí y se le ha preguntado acerca del granero}] \\ & + P[\text{sí y no se le ha preguntado acerca del granero}] \end{aligned}$$

(Basado en un estudio registrado en McGraw-Hill Yearbook of Science and Technology, 1981, págs. 249-251.)

Método aleatorizado de respuesta para obtener respuestas honestas a preguntas comprometidas. Es un método que se utiliza para garantizar que un individuo que responde a cuestiones comprometidas mantenga el anonimato, animándole así a dar una respuesta verdadera. Funciona de la forma siguiente: se plantean dos preguntas A y B, una de las cuales es referente a temas comprometidos y la otra no. Debe conocerse la probabilidad de recibir un sí como respuesta a la pregunta no comprometida. Por ejemplo, se podría preguntar

A: ¿Su número de la Seguridad Social termina en un dígito impar? (No comprometida)

B: ¿Alguna vez ha cursado intencionadamente una reclamación de seguro fraudulenta? (Comprometida)

Sabemos que P[responde sí | ha respondido a A] $= \frac{1}{2}$. Deseamos hallar la aproximación de P[responde sí | ha respondido a B]. Se le pide al sujeto que tire una moneda y responda a A si en la moneda sale cara y a B si sale cruz. De esta forma, el entrevistador no sabe a qué pregunta está respondiendo el sujeto. Así, una respuesta afirmativa no es incriminativa. No existe forma alguna de que el entrevistador sepa si el sujeto está diciendo «Sí, mi número de la Seguridad Social termina en un dígito impar» o «Sí, he cursado intencionadamente una reclamación fraudulenta». El porcentaje de sujetos del grupo de individuos que han respondido sí, se utiliza para calcular P[responde sí].

a) Utilizar el hecho de que el suceso «responde sí» es el suceso «responde sí y ha respondido a A» o «responde sí y ha respondido a B» para demostrar que P[responde sí | ha respondido a B] es igual a

$$\frac{P[\text{responde sí}] - P[\text{responde sí} \mid \text{ha respondido a A}]\, P[\text{ha respondido a A}]}{P[\text{ha respondido a B}]}$$

b) Si se prueba esta técnica en 100 sujetos y 60 responden sí, hallar la probabilidad aproximada de que una persona del grupo, seleccionada aleatoriamente, haya cursado intencionalmente una reclamación fraudulenta.

10. En un estudio sobre estudiantes de bachillerato, a cada sujeto se le pide que tire un dado y luego una moneda. Si en la moneda sale cara, el sujeto debe responder a la pregunta A, de lo contrario, a la pregunta B.

A: ¿El dado ha sacado un número par?
B: ¿Ha fumado alguna vez marihuana?

En un grupo de 50 sujetos, 35 respondieron sí. Utilizar esta información para hallar la aproximación de la probabilidad de que un estudiante de este grupo seleccionado aleatoriamente haya fumado marihuana.

3.7. TEOREMA DE BAYES

El objeto de esta sección es el teorema formulado por el reverendo Thomas Bayes (1761). Está relacionado con la probabilidad condicionada. El teorema de Bayes se utiliza para hallar *P[A|B]* cuando la información de que se dispone no es directamente compatible con la que se requería en la Definición 3.3.1. Es decir, se utiliza para hallar *P[A | B]* cuando *P[A yB] y P[B]* no se conocen de inmediato.

Los problemas de Bayes pueden resolverse con la ayuda de un diagrama de árbol. Ilustraremos la idea antes de formular formalmente el teorema.

Ejemplo 3.7.1. Se ha desarrollado un procedimiento para detectar un tipo particular de artritis en individuos de alrededor de cincuenta años de edad. A partir de una investigación realizada a nivel nacional, se sabe que, aproximadamente, el 10 % de los individuos de esta edad sufre esta forma de artritis. Se aplica el procedimiento propuesto a individuos con enfermedad artrítica confirmada, y su resultado es correcto en el 85 % de los casos. Cuando el procedimiento se pone a prueba con individuos de la misma edad que, se sabe, están libres de la enfermedad, se obtiene un coeficiente de falsos positivos del 4%.

Para que este test sea utilizado como detector de la artritis es necesario que sea un fuerte indicador de que la enfermedad está presente. Sea *D* el suceso que denote la presencia de la enfermedad y *T+* el suceso que alude al resultado positivo para el test. Pretendemos hallar *P[D | T+]* y que sea alta. Puesto que esta probabilidad es condicionada, lo primero que se nos ocurriría hacer sería aplicar la Definición 3.3.1. Sin embargo, no tenemos *P[D y T+],* la probabilidad de que exista la enfermedad y el test dé positivo, ni tampoco *P[T+],* la probabilidad de resultado positivo para el test. Así que la Definición 3.3.1 no puede emplearse directamente; se necesita otro método para calcular la probabilidad deseada.

Para resolver el problema, obsérvese que se dan las probabilidades *(T—* denota el hecho de que el resultado del test sea negativo):

$$P[D] = 0.10 \quad P[T+ \mid D] = 0.85 \quad P[T+ \mid D'] = 0.04$$
$$P[D'] = 0.90 \quad P[T- \mid D] = 0.15 \quad P[T- \mid D'] = 0.96$$

Dado que conocemos *P[D] y P[D'],* empezamos el árbol enumerando estos sucesos junto con sus probabilidades correspondientes. Si la enfermedad está presente, podemos asignar

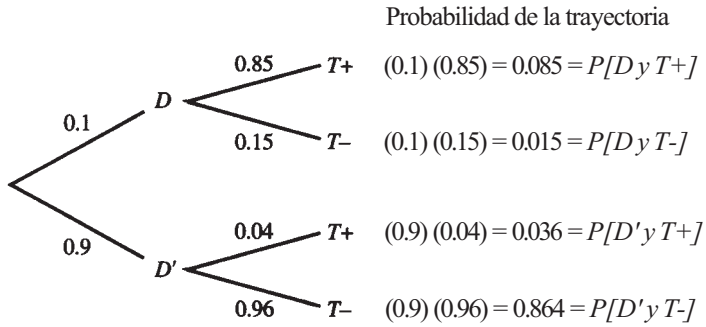


Figura 3.12. Trayectorias y probabilidades de las trayectorias.

probabilidades de 0.85 y 0.15 a los sucesos *T+,* el resultado del test es positivo, y *T-,* el resultado del test es negativo, respectivamente. Si la enfermedad no está presente, estas probabilidades condicionadas son, respectivamente, 0.04 y 0.96. Todas estas probabilidades se muestran en la Figura 3.12. Obsérvese que la probabilidad de la primera trayectoria es *P[D] P[T+ | D]* lo que, mediante la regla de la multiplicación, nos da *P[D* y *T+].*

Para hallar *P[D | T+],* el coeficiente predictivo positivo, aplicamos la Definición 3.3.1 para obtener

$$P[D \mid T+] = \frac{P[D \text{ y } T+]}{P[T+]}$$

En el árbol, vemos que *P[D y T+]* = 0.085. El suceso *T+* se representa mediante las trayectorias 1 y 3, y, por lo tanto, *P[T+]* = 0.085 + 0.036 = 0.121. Mediante sustitución obtendremos

$$P[D \mid T+] = \frac{0.085}{0.121} = 0.70$$

Esto es, si el test es positivo, hay un 70 % de posibilidades de que, en efecto, la enfermedad exista.

Resolviendo el Ejemplo 3.7.1 con un diagrama de árbol, estamos aplicando de forma completamente natural el teorema de Bayes. En sí mismo, el teorema es una afirmación formal de la técnica empleada al utilizar un árbol. Supongamos que existe un conjunto de sucesos mutuamente excluyentes $A_1, A_2, \ldots, A_n$, tales que $P[A_1], P[A_2], \ldots, P[A_n]$ son conocidas y $\sum_{i=1}^{n} P[A_i] = 1$. Dicho conjunto se denomina una *partición del espacio muestral.* Estos sucesos producen la primera ramificación del diagrama de árbol. Supongamos que se produce otro suceso *B* y que conocemos *P[B | A_i]* para cada *i.* Este suceso produce la ramificación de la segunda etapa del árbol. Queremos hallar la probabilidad de que ocurra un suceso específico de la partición A_j dado que ha ocurrido *B.* Por la Definición 3.3.1,

$$P[A_j \mid B] = \frac{P[A_j \text{ y } B]}{P[B]}$$

En la formulación del teorema de Bayes, el numerador y el denominador se expresan en forma alternativa aplicando a cada uno de ellos la regla de la multiplicación. El numerador corresponde a la probabilidad de la trayectoria j'-ésima; el denominador es la suma de las probabilidades de las trayectorias que corresponden al suceso *B.* La formulación formal del teorema se da en el Teorema 3.7.1. Su demostración está indicada en el Ejercicio 6.

Teorema 3.7.1. Teorema de Bayes. Sea $A_1, A_2, A_3, \ldots, A_n$ una colección de sucesos que forman una partición de 5. Sea *B* un suceso tal que $P[B] \neq 0$. Entonces, cualquiera que sea el suceso A_j, $J = 1, 2, 3, \ldots, n$,

$$P[A_j \mid B] = \frac{P[B \mid A_j]P[A_j]}{\sum_{i=1}^{n} P[B \mid A_i]P[A_i]}$$

El teorema de Bayes es mucho más fácil de manejar en la práctica que de establecer formalmente. Para verlo vamos a reconsiderar el Ejemplo 3.7.1 y a resolverlo sin utilizar el diagrama de árbol.

Ejemplo 3.7.2. En el Ejemplo 3.7.1, hemos calculado $P[D \mid T+]$, donde D es el suceso que significa «tenga artritis» y $T+$ es el suceso representativo del hecho de que el test sea positivo. Los sucesos D y D' forman una partición de 5. (Un individuo o tiene o no tiene artritis.) El suceso $T+$ tiene asociada una probabilidad distinta de cero. Se nos da

$$P[D] = 0.10 \quad P[T+ \mid D] = 0.85 \quad P[T+ \mid D'] = 0.04$$
$$P[D'] = 0.90 \quad P[T- \mid D] = 0.15 \quad P[T- \mid D'] = 0.96$$

Aplicando el teorema de Bayes, obtenemos

$$P[D \mid T+] = \frac{P[T+ \mid D]P[D]}{P[T+ \mid D]P[D] + P[T+ \mid D']P[D']}$$
$$= \frac{(0.85)(0.10)}{(0.85)(0.10) + (0.04)(0.90)} \cong 0.70$$

Obsérvese que el resultado es el mismo que el obtenido por medio del diagrama de árbol.

El Ejemplo 3.7.3 nos muestra el manejo del teorema de Bayes cuando S está dividido por una partición de más de dos sucesos.

Ejemplo 3.7.3. Se cree que la distribución de los grupos sanguíneos en Estados Unidos en la Segunda Guerra Mundial era: tipo A, 41 %; tipo B, 9 %; tipo AB, 4 %; y tipo 0,46 %. Se estima que en esa época, el 4 % de las personas pertenecientes al tipo 0 fue clasificado como del tipo A; el 88 % de los del tipo A fue correctamente clasificado; el 4 % de los del tipo B se clasificó como del tipo A, y el 10 % de los del tipo AB fue, igualmente, clasificado como del tipo A. Un soldado fue herido y conducido a la enfermería. Se le clasificó como del tipo A. ¿Cuál es la probabilidad de que tal grupo sea ciertamente el suyo?
Sean los sucesos:

A_1: Es del tipo A.
A_2: Es del tipo B.
A_3: Es del tipo AB.
A_4: Es del tipo 0.
B: Es clasificado como del tipo A.

Deseamos calcular $P[A_1 \mid B]$. Los datos de que disponemos son:

$$P[A_1] = 0.41 \quad P[B \mid A_1] = 0.88$$
$$P[A_2] = 0.09 \quad P[B \mid A_2] = 0.04$$
$$P[A_3] = 0.04 \quad P[B \mid A_3] = 0.10$$
$$P[A_4] = 0.46 \quad P[B \mid A_4] = 0.04$$

En la Figura 3.13 se muestra el diagrama de árbol utilizado para responder a esta pregunta. Obsérvese que, según la Definición 3.3.1, $P[A_1 \mid B] = P[A_1 \text{ y } B]/P[B]$. El numerador de esta probabilidad es la probabilidad de la trayectoria 1, es decir, 0.3608. El denominador es la suma de las probabilidades de las trayectorias 1, 3, 5 y 7, es decir, 0.3868. Por tanto, $P[A_1 \mid B] = 0.3608/0.3868 \cong 0.93$. Por el teorema de Bayes,

$$P[A_1 \mid B] = \frac{P[B \mid A_1]P[A_1]}{\sum_{i=1}^{4} P[B \mid A_i]P[A_i]}$$
$$= \frac{(0.88)(0.41)}{(0.88)(0.41) + (0.04)(0.09) + (0.10)(0.04) + (0.04)(0.46)}$$
$$\cong 0.93$$

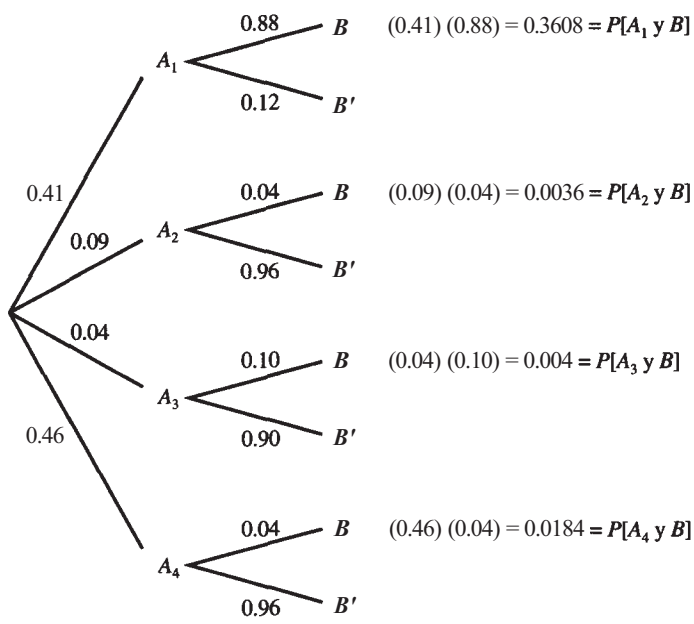


Figura 3.13. $P[A_1 \text{ y } B] = 0.3608$; $P[B] = 0.3608 + 0.0036 + 0.004 + 0.0184 = 0.3868$; $P[A_1 \mid B] = 0.3608/0.3868 \cong 0.93$

En la práctica, esto significa que hay un 93 % de posibilidades de que, si se le clasificó como del tipo A, su grupo sanguíneo sea efectivamente el A. Hay un 7 % de posibilidades de que, habiendo sido clasificado como del tipo A, pertenezca en realidad a otro.

1. Las estadísticas indican que en Estados Unidos la probabilidad de que una madre muera durante el parto es 0.00022. Si no es de raza negra, la probabilidad de muerte es 0.00017, mientras que si lo es, esta probabilidad aumenta a 0.00064. Supongamos que el 10 % de los partos corresponde a mujeres negras.
   a) Dibujar un diagrama de árbol describiendo las probabilidades dadas, y hallar las probabilidades correspondientes a las trayectorias en cada uno de los cuatro casos. (Sea *D* el suceso que denota que la madre muere y *B* el que alude a que es de raza negra.)
   b) Utilizar el árbol del apartado *a* para calcular la probabilidad de que una madre que muere en el parto sea de raza negra.
   c) Haciendo uso del teorema de Bayes, hallar la probabilidad de que una madre que muere en el parto sea de raza negra, y comparar el resultado con el obtenido en el apartado *b.*
2. Un test diseñado para diagnosticar el cáncer de cuello uterino tiene un coeficiente de falsos negativos y falsos positivos de 0.05, cada uno. De una cierta población de mujeres, el 4 % está afectado por este tipo de cáncer. ¿Cuál es la probabilidad de que una mujer de la población elegida aleatoriamente tenga cáncer de cuello uterino, dado que su resultado con el test es positivo?
3. Un paciente de cáncer está siendo tratado con una combinación de tres fármacos. Se observa que, cuando se utilizan simultáneamente, a menudo dos de los tres fármacos se inhibirán de forma que, de hecho, sólo uno será activo frente al tumor. Suponga que cuando esto ocurra, la probabilidad de que el fármaco A actúe solo es la misma que la del fármaco B y la del C, es decir $\frac{1}{3}$. La efectividad de cada fármaco, con respecto a producir

una remisión del tumor, es diferente. El fármaco A se ha mostrado efectivo en un 50 % de los casos; el fármaco B, en un 75 %, y el fármaco C, en un 60 % . La enfermedad remite en el paciente. ¿Cuál es la probabilidad de que el responsable de ello sea el fármaco B?

4. La distrofia muscular de Duchenne es una enfermedad de los músculos que afecta a los jóvenes. La naturaleza de esta enfermedad es tal que no se transmite desde los varones afectados, sino que se propaga a partir de mujeres portadoras que rara vez exhiben síntoma alguno de tener la enfermedad. Considérese una mujer que es hija de una portadora detectada de la enfermedad. Ésta tiene tres hijos completamente normales. Emplear el teorema de Bayes para hallar la probabilidad de que la mujer sea portadora. Es decir, calcular P[portadora | tres hijos normales].
5. Se nos dice que el valor predictivo positivo de un test está más influenciado por la especificidad que por la sensibilidad. (Véanse Ejercicios 5, 6 y 10 de la Sección 3.4.) Para demostrarlo, calcular el valor predictivo positivo de cada uno de los conjuntos dados en los apartados *a, b, dy e:*
   a) Sensibilidad = 0.95
   Prevalencia (P[realidad +]) = 0.10
   Especificidad =1.00
   b) Sensibilidad = 0.95
   Prevalencia = 0.10
   Especificidad = 0.50 ,
   c) ¿Cuál es la diferencia entre los coeficientes predictivos positivos a medida que la especificidad disminuye de 1.0 a 0.5?
   d) Especificidad = 0.95
   Prevalencia = 0.10
   Sensibilidad = 1.00
   e) Especificidad = 0.95
   Prevalencia = 0.10
   Sensibilidad = 0.50
   f) ¿Cuál es la diferencia entre los coeficientes predictivos positivos a medida que la sensibilidad disminuye de 1.00 a 0.5?
   (Basado en la información hallada en Victoria Wells, William Halperin y Michael Thun, «Estimated Predictive Value of Screening for Illicit Drugs in the Workplace», *American Journal of Public Health,* vol. 78, n.° 7, julio de 1988, págs. 817-823.)
6. Para deducir el teorema de Bayes, consideremos la Figura 3.14.
   a) Determinar la expresión para *P[B].*
   b) Utilizar la regla de la multiplicación para hallar las expresiones para P[A_1 y *B], P[*A_2 y *B],* ..., *P[*A_n y *B]* en las cuales A_1, A_2, ..., A_n son los sucesos dados.

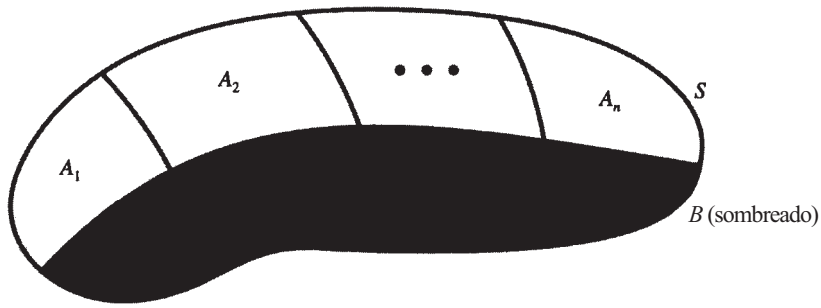
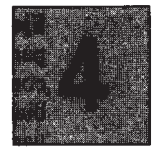


Figura 3.14. Sucesos A_1, A_2,..., A_n, partición de *S.*

c) Utilizar el apartado *b* para hallar una expresión alternativa de $P[B]$.
d) Aplicar la Definición 3.3.1 para hallar una expresión de $P[A_j \mid B]$.
e) Efectuar una sustitución en la expresión del apartado *d* para obtener el teorema de Bayes.



Variables aleatorias discretas

En el Capítulo 1 hemos considerado parte de los métodos que generalmente se utilizan para describir conjuntos de datos. Algunos de estos conjuntos representan el total de la población estudiada; otros constituyen una muestra extraída de una población más grande. En este último caso sólo podemos intentar sacar conclusiones aproximadas acerca de la población de la que procede la muestra. En los Capítulos 2 y 3 hemos tratado la teoría de la probabilidad. Vimos cómo pueden utilizarse los axiomas y los teoremas de probabilidad para resolver muchas cuestiones de naturaleza moderadamente complicada. No hemos considerado, sin embargo, las implicaciones de la teoría de la probabilidad en el análisis de datos. Es decir, todavía no hemos empezado a mostrar cómo puede emplearse la teoría de la probabilidad para sacar conclusiones precisas acerca de una población, con base en una muestra extraída de ella. Para hacer esto, primero hemos de dirigir nuestra atención hacia un tema que constituye el eslabón entre la teoría de la probabilidad y la estadística aplicada. En particular, desarrollaremos la noción de *variable aleatoria.* Este concepto ya fue introducido en el Capítulo 1, y vimos que las variables aleatorias de interés para nosotros pueden clasificarse en dos grandes categorías: discretas o continuas. Comenzaremos con el repaso de estas definiciones.

4.1. VARIABLES DISCRETAS Y CONTINUAS

El concepto de variable aleatoria no es difícil. De hecho, en muchos de los problemas que se han presentado con anterioridad subyacen variables aleatorias aunque entonces no se haya utilizado explícitamente este término. Intuitivamente, una variable aleatoria es una variable cuyo valor numérico se determina al azar. Las variables aleatorias se representarán por letras mayúsculas, y sus valores numéricos observados, por letras minúsculas.

Ejemplo 4.1.1. Consideremos la variable Y, número de pérdidas de petróleo que afectan por año a las aguas costeras de Estados Unidos. Se trata de una variable aleatoria. No toma exactamente el mismo valor cada año. Puede variar enormemente de año en año, y su variación es debida al azar. Si en un año determinado hay cinco pérdidas escribiremos $y = 5$.

Ejemplo 4.1.2. Sea Z el número de cm^3 de un fármaco que debe prescribirse a un paciente para controlar ataques epilépticos. Esta variable cambia de valor de paciente a paciente como resultado de diferencias metabólicas y de otros tipos. De hecho, su valor puede cambiar en el mismo paciente a lo largo del tiempo; de aquí que deba ser considerada como variable aleatoria.

Ejemplo 4.1.3. Los cefalópodos («pies en la cabeza») son los moluscos más desarrollados. Uno de ellos, el pulpo, contrariamente a la creencia popular, no es grande. La variable *D,* diámetro del cuerpo de un pulpo adulto, es una variable aleatoria. No todos los pulpos son del mismo tamaño. La variabilidad en el tamaño es debida a factores tanto genéticos como del medio.

Ejemplo 4.1.4. Consideremos la variable *B,* número de niños nacidos en una maternidad concreta antes del nacimiento de los primeros gemelos siameses. *A priori,* no puede establecerse firmemente un límite superior para el conjunto de los posibles valores de *B.* Posiblemente, *B* puede asumir cualquiera de los valores {0, 1, 2, 3, 4, …}. Se trata de una variable aleatoria, ya que el azar juega el papel más importante en el nacimiento de siameses.

Hay dos tipos de variables aleatorias fácilmente identificados, discretas y continuas. En todo lo que viene a continuación el lector debe ser capaz de distinguir los dos tipos, ya que matemáticamente se manejan de forma algo diferente.

Definición 4.1.1. Una variable aleatoria *X* es *discreta* si puede tomar un número finito, p infinito numerable de valores puntuales posibles.

En los Ejemplos 4.1.1 y 4.1.4, *Y,* número de pérdidas de petróleo por año que afectan a las aguas costeras de Estados Unidos, y *B,* número de niños nacidos en una determinada maternidad antes del nacimiento de la primera pareja de gemelos siameses, son discretas. La variable *Y* es discreta, dado que el número de valores posibles para *Y* es finito. Los valores posibles pueden razonablemente variar desde 0 hasta, quizás, 50, un conjunto finito de números. La variable *B* es discreta, ya que el conjunto de valores posibles {0, 1, 2, 3, 4, …}, es infinito numerable. Generalmente, las variables discretas surgen en cuestiones relacionadas con datos obtenidos por conteo.

Definición 4.1.2. Una variable aleatoria *X* es *continua* si puede tomar cualquier valor en algún intervalo (o intervalos) del conjunto de los números reales y la probabilidad de que tome uno determinado es 0.

Las variables Z, número de cm^3 de fármaco para controlar ataques, que debería prescribirse a un paciente, y *D,* diámetro del cuerpo de un pulpo adulto, son ambas continuas. La cantidad de fármaco a prescribir no se restringe a cualquier colección finita de posibles valores prefijados. Puede tomar cualquier valor entre cero y, digamos, 0.3 cm^3. Es decir, los valores de Z están en el intervalo [0,0.3]. Análogamente, el valor del diámetro del cuerpo de un pulpo adulto puede situarse en cualquier valor dentro de límites razonables, digamos, de 10 a 30 centímetros. Es decir, los valores de *D* están en el intervalo [10,30]. Establecer que la probabilidad de que una variable continua tome cualquier valor específico es 0 es *esencial* para la definición. Las variables discretas no están sujetas a tal restricción. La restricción es intuitivamente aceptable, puesto que si nosotros preguntamos, *antes* de que la selección se haya hecho, ¿cuál es la probabilidad de que un determinado pulpo tenga un diámetro corporal de *exactamente* 12.981 321 069 217 031 2 centímetros? la respuesta es 0. Es prácticamente imposible encontrar un pulpo con ese diámetro precisamente, sin la más mínima desviación

Lo discutiremos más adelante desde un punto de vista matemático, en la Sección 4.2. Las variables continuas generalmente surgen cuando se trabaja con datos de medida.

EJERCICIOS 4.1

En cada uno de los ejercicios siguientes, identificar la variable como discreta o continua:

1. A: El número de septos o tabiques en cámaras de una concha de nautilo.
2. V: El volumen de orina producido por hora.
3. *B;* La cantidad de sangre perdida por un paciente durante el transcurso de una operación.
4. *H:* El número de horas de luz por día necesarias para que una planta florezca.
5. C: El número de abejas obreras en una colonia de abejas productoras de miel.
6. *R:* La cantidad de lluvias recibidas por día en una región concreta.
7. *S:* El nivel en suero de bilirrubina en un niño, en miligramos por decilitro.
8. *W:* El peso ganado por una mujer durante el embarazo.
9. *T:* El tiempo mínimo necesario para que una plaga de abejas asesinas avance 1000 millas.
10. *X:* El número de pruebas necesarias que permita conseguir el primer injerto realizado con éxito, de un tallo de cornejo rosa sobre un tronco de cornejo blanco.
11. C: Donde $C = 1$, si el árbol muestra es de tamaño adecuado para madera, y $C = 0$ en caso contrario.
12. *P:* La tensión arterial sistólica de un paciente con hipertensión.
13. *L:* El tiempo que la enfermedad de un paciente de leucemia ha estado en remisión.
14. *E:* La altitud a la que se sitúa el límite de arbolado en una montaña

4.2. FUNCIONES DE DENSIDAD DISCRETA Y ESPERANZA

Cuando estamos tratando con una variable, no basta con considerar que es aleatoria. Necesitamos ser capaces de predecir, en algún sentido, el valor que la variable adoptará en cualquier momento. Puesto que el comportamiento de una variable aleatoria está gobernado por el azar, las predicciones deberán hacerse con un serio tratamiento de la incertidumbre. Lo más conveniente es describir el comportamiento de la variable en términos de probabilidades. Para ello se utilizan dos funciones, la función de densidad y la función de distribución acumulada.

La función de densidad, para una variable aleatoria discreta, nos da la probabilidad de que la variable aleatoria X adopte un valor numérico x determinado; la distribución acumulada proporciona la probabilidad de que X tome un valor por debajo de x, incluyendo éste. Definimos la densidad discreta en la Definición 4.2.1.

Definición 4.2.1. Densidad discreta. Siendo *X discreta.* La *densidad f para X* es

$$f(x) = P[X = x]$$

para x real.

Hay varias cuestiones a tener en cuenta relativas a la densidad en el caso discreto. Primero, está definida sobre toda la recta real, y para cualquier número real *x, f(x)* es la probabilidad de que la variable aleatoria X tome el valor x. Por ejemplo,/(2) es la probabilidad de que

la variable aleatoria X tome el valor numérico 2. Segundo, puesto que $f(x)$ es una probabilidad, $f(x) > 0$, independientemente del valor de x. Es decir, f nunca es negativa. Tercero, si sumamos f sobre todos los valores físicos posibles de X, la suma deberá ser 1. Esto es,

$$\sum_{\text{todo } x} f(x) = 1$$

Para todos los valores reales de x que son físicamente imposibles, $f(x) = 0$. Este hecho debería ser interpretado correctamente en los próximos ejercicios, aunque no esté especificado siempre. Cualquier función que es no negativa y toma el valor 1 cuando se suma sobre su conjunto de valores posibles, puede considerarse como la densidad para una variable aleatoria discreta.

El Ejemplo 4.2.1 ilustra el uso de un diagrama de árbol para generar una densidad. Como se verá más adelante, la variable estudiada en este ejemplo es una variable aleatoria que sigue una distribución binomial. Esta distribución se presenta frecuentemente en la práctica, por lo que será tratada con detalle en la Sección 4.4.

Ejemplo 4.2.1. El mimetismo batesiano fue descrito por primera vez por el naturalista británico H. W. Bates, en 1862. En el mimetismo batesiano, un imitador inocuo engaña a sus depredadores imitando modelos que, porque pican o tienen mal sabor, el depredador ha aprendido a evitar. En un experimento sobre mimetismo se obtiene un modelo artificial bañando a gusanos de la harina en una solución de quinina para darles un sabor amargo. Después se marcan los gusanos con una banda verde de pintura celulosa para darles un aspecto anormal, y se les da como alimento a tres estorninos enjaulados. Los estorninos aprenden a asociar el mareaje especial con el sabor amargo. A continuación se presenta a cada estornino un gusano de harina que no ha sido sumergido en quinina, pero que ha sido pintado para que se parezca al modelo. La probabilidad de que el imitador no sea comido por el estornino, que normalmente come vorazmente los gusanos de harina, es 0.8. Sea X el número de gusanos que escapan de ser detectados. Puesto que sólo pueden tomar los valores 0, 1, 2, 3, X es discreta.

Para responder a cuestiones de probabilidad relativas a X, debemos encontrar su densidad, $f(x)$. Esto se hace mediante un diagrama de árbol. Para construir el árbol indicaremos con e un mimetismo que escapa y por c uno que es capturado. El diagrama de árbol de la Figura 4.1 tiene tres ramificaciones, representando el experimento en el que se utilizaron los tres estorninos. Dado que el comportamiento de un estornino no tiene efecto sobre los demás, la probabi-

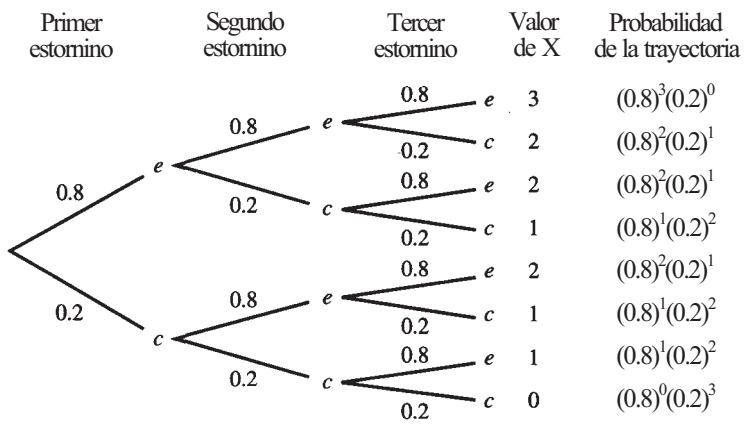


Figura 4.1

lidad de que escapen y la de que sean capturados siguen siendo 0.8 y 0.2, respectivamente, durante todo el experimento. Estas probabilidades quedan reflejadas en el diagrama. Para hallar la probabilidad de conseguir una secuencia de sucesos tal como *eee,* los tres gusanos escapan, multiplicamos las probabilidades a lo largo de la trayectoria 1 para así obtener $(0.8)^3$. Si queremos resaltar el hecho de que esta secuencia de sucesos puede asegurar que ninguno de los gusanos ha sido capturado, volvemos a escribir la probabilidad de la trayectoria en la forma $(0.8)^3$ (0.2)°. Podemos hallar las probabilidades de otras trayectorias del mismo modo. Para hallar la probabilidad de que X tome un valor concreto, sumaremos las probabilidades de las trayectorias correspondientes a ese valor. Por tanto, la densidad puede ser leída directamente del árbol y se resume del siguiente modo:

x	0	1	2	3
$f(x) = P[X = x]$	$(0.2)^3$	$3(0.8)^1(0.2)^2$	$3(0.8)^2(0.2)$	$(0.8)^3$

x	0	1	2	3
f(x)	$\frac{8}{1000}$	$\frac{96}{1000}$	$\frac{384}{1000}$	$\frac{512}{1000}$

Se entiende que f(x) = 0 para valores de X diferentes que 0, 1, 2 ó 3.

Observemos que, como era de esperar, cada valor de la fila 2 de la tabla es no negativo y entre todos suman 1. Esta tabla puede utilizarse para responder a cualquier cuestión relevante que se plantee. Por ejemplo, la probabilidad de que escapen exactamente dos a la detección está dada por: $P[X = 2] = /(2) = 0.384$. La probabilidad de que escapen a lo sumo dos viene dada por

$$\begin{aligned} P[X \leqslant 2] &= P[X = 0] + P[X = 1] + P[X = 2] \\ &= f(0) + f(1) + f(2) \\ &= 0.008 + 0.096 + 0.384 = 0.488 \end{aligned}$$

Obsérvese también que $P[X < 2] = 0.104 \neq P[X \leqslant 2]$. La inclusión o la exclusión de un punto extremo en el caso *discreto* puede afectar al valor numérico de la respuesta.

El ejemplo anterior pretende representar el concepto de una densidad discreta. En la práctica, la densidad teórica deriva con frecuencia de los datos muestrales. La aproximación viene dada por la distribución de frecuencias relativas de la variable aleatoria, ya comentada en la Sección 1.3.

Esperanza

La función densidad de una variable aleatoria describe totalmente el comportamiento de una variable en el sentido de una población ideal. Al considerar una visión general de una población, podemos definir unas constantes o «parámetros» descriptivos asociados a cualquier variable aleatoria. El conocimiento de los valores numéricos de estos parámetros proporciona al investigador una rápida visión de la naturaleza de las variables. En el Capítulo 1, mencionábamos tres parámetros: la media μ, la varianza σ^2 y la desviación típica σ. Si la densidad

exacta de la variable que nos interesa es conocida, entonces el valor numérico de cada parámetro puede ser hallado a partir de consideraciones matemáticas. Este es el tema que tratamos en este apartado. Si el investigador dispone sólo de algunas observaciones con respecto a la variable aleatoria (o serie de datos), los valores de estos parámetros no pueden ser hallados con exactitud, sino mediante aproximaciones utilizando técnicas estadísticas. Éste es el tema tratado en el Capítulo 1 y en gran parte del resto de este texto.

Para poder entender el razonamiento seguido en la mayoría de los métodos estadísticos, necesitamos familiarizarnos principalmente con un concepto general, la idea de esperanza matemática o valor esperado. Este concepto se utiliza en la definición de la mayoría de los parámetros estadísticos, y confiere una lógica básica para muchos de los métodos de inferencia estadística que presentaremos posteriormente. Comenzaremos con una definición intuitiva del término.

> **Definición 4.2.2.** Valor intuitivo de la esperanza. Sea X una variable aleatoria. La *esperanza* de X, representada por $E(X)$, es el promedio teórico de X a largo plazo.

Recuerde del Capítulo 1, que la media teórica de una variable aleatoria X se representa con la letra griega μ, que hace referencia a la media poblacional. Este valor promedio recibe también el nombre de esperanza de X, representado por $E(X)$. Así, $E(X)$ y μ son símbolos intercambiables para el mismo concepto. A efectos de conseguir una mayor comprensión del concepto de esperanza, consideramos un experimento en el cual un único dado equilibrado qs lanzado repetidas veces y cada vez se anota el número obtenido, X. Una secuencia de observaciones podría ser

$$2,\ 1,\ 6,\ 4,\ 2,\ 5,\ 1,\ 6,\ 3,...$$

Esta secuencia generará la correspondiente secuencia de medias aritméticas:

$$2,\ \frac{2+1}{2},\ \frac{2+1+6}{3},\ \frac{2+1+6+4}{4},\ \frac{2+1+6+4+2}{5},\ \frac{2+1+6+4+2+5}{6},...$$

o bien,

$$2,\ 1.5,\ 3,\ 3.25,\ 3,\ 3.33,...$$

Estas medias no son obviamente constantes. Varían durante el experimento pero, por lo general, la magnitud de los cambios disminuye a medida que se realizan más observaciones. En realidad, cuanto más aumente el número de lanzamientos del dado, la secuencia de medias tenderá a concentrarse en torno a un mismo valor numérico. Este valor es la esperanza, o «media teórica a largo plazo», para X. Nuestro trabajo consiste en determinar estas esperanzas a partir de la densidad de X, sin tener que recurrir a la experimentación. En algunos casos, esto se puede conseguir mediante observación, como lo demuestra a continuación el Ejemplo 4.2.2.

> **Ejemplo 4.2.2.** Considere un experimento en el cual lanzamos un dado. La variable k representa el valor obtenido en cada lanzamiento. Esta variable está distribuida uniformemente con una densidad de la siguiente forma:

x	**1**	**2**	**3**	**4**	**5**	**6**
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Observe la simetría de la densidad. A largo plazo, esperaríamos obtener por lo menos tantos seises como unos, tantos doses como cincos, y tantos treses como cuatros. Cada uno de estos pares tiene una media de 3.5. El sentido común nos indica que 3.5 es la esperanza de *X.* Debemos tener en cuenta también que 3.5 no puede ser un valor *X.* No hay necesidad de que la esperanza de una variable aleatoria esté entre los posibles valores de *X.*

Si la densidad no es simétrica, significa que no puede ser hallada mediante una simple observación. En este caso, para calcular *E(X),* se emplea la Definición 4.2.3.

Definición 4.2.3. El valor esperado, discreto. Sea *X* una variable aleatoria *discreta* con una densidad *f*(x). La *esperanza* o *valor medio* de *X* es

$$\mu = E(X) = \sum_{\text{todo } x} xf(x)$$

Hallemos la *E(X)* en el caso del problema planteado con el dado en el Ejemplo 4.2.2. Utilizando la Definición 4.2.3, tenemos que

$$\begin{aligned}\mu = E(X) &= \sum_{\text{todo } x} xf(x) \\ &= 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} \\ &= 3.5\end{aligned}$$

Como esperábamos, este resultado concuerda con el que hemos hallado mediante observación en el Ejemplo 4.2.2.

Dado que la densidad calculada en el Ejemplo 4.2.1 no es simétrica, el número esperado de mimetismos que escapan a la detección, no puede ser hallado mediante inspección. Sin embargo, *E(X)* puede calcularse mediante la Definición 4.2.3.

Ejemplo 4.2.3. Sea la variable aleatoria *X,* número de mimetismos que escapan al control, en el experimento realizado sobre mimetismos batesianos del Ejemplo 4.2.1. Definimos la densidad de *X,*

x	0	1	2	3
f(x)	$\frac{8}{1000}$	$\frac{96}{1000}$	$\frac{384}{1000}$	$\frac{512}{1000}$

Esta densidad no es simétrica, por lo que es imposible predecir *E(X)* mediante observación. Utilizando la Definición 4.2.3, tenemos

$$\begin{aligned}\mu = E(X) &= 0 \cdot \tfrac{8}{1000} + 1 \cdot \tfrac{96}{1000} + 2 \cdot \tfrac{384}{1000} + 3 \cdot \tfrac{512}{1000} \\ &= \tfrac{2400}{1000} = \tfrac{12}{5} = 2.4\end{aligned}$$

Así, durante varias pruebas con el mismo experimento, esperaríamos que la media de los mimetismos que escapan a la detección sea de 2.4.

Dada la densidad para *X,* podemos hallar la esperanza de las funciones de *X* tales como X^2 o $(X - \mu)^2$. Estas son especialmente importantes puesto que nos permiten calcular σ^2, la

varianza de *X,* a partir del conocimiento de la densidad. Los Ejercicios 7 a 10 plantean está cuestión.

EJERCICIOS 4.2

1. La siguiente tabla muestra la densidad para la variable aleatoria *X,* número de personas por día que solicitan un tratamiento innecesario en el servicio de urgencias de un pequeño hospital.

x	0	1	2	3	4	5
$f(x)$	0.01	0.1	0.3	0.4	0.1	?

   a) Encontrar $f(5)$. ¿Qué probabilidad representa en el contexto del problema?
   b) Encontrar $P[X < 2]$. Interpretar esta probabilidad en el contexto del problema.
   c) Encontrar $P[X < 2]$.
   d) Encontrar $P[X > 3]$.

2. La siguiente tabla muestra la densidad para la variable aleatoria *X,* número de aleteos por segundo de una especie de polillas grandes mientras vuelan.

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	?

   a) Encontrar $f(10)$.
   b) Encontrar $P[X < 8]$. Interpretar esta probabilidad en el contexto del problema.
   c) Encontrar $P[X < 8]$.
   d) Encontrar $P[X > 7]$.
   e) Encontrar $P[X > 7]$.

3. Se desarrolla un compuesto para aliviar las cefaleas migrañosas. El fabricante afirma que es eficaz en un 90 % de los casos. Se prueba sobre cuatro pacientes. Sea X el número de pacientes que obtienen alivio. Utilice un diagrama de árbol para resolver estos apartados:
   a) Encontrar la densidad para *X,* suponiendo que la afirmación del fabricante sea correcta.
   b) Encontrar $P[X < 1]$. Interpretar esta probabilidad en el contexto del problema.
   c) Si el compuesto no alivia a ninguno de los pacientes, ¿es ésa una razón para sospechar de la afirmación de la compañía relativa a que el compuesto es eficaz en el 90 % de los casos? Razonar sobre la base de la probabilidad implicada.

4. Un brote de parotiditis está extendiéndose entre los niños de la escuela primaria. El 10 % de ellos está afectado. Un pediatra atiende a tres niños de esta edad durante la primera hora de su día de trabajo. Sea X el número de los que tienen parotiditis. Suponga independencia y use un diagrama de árbol para hallar la densidad de X. Utilícela para calcular la probabilidad de que ninguno de los tres niños tenga parotiditis, y la de que sólo uno las tenga.

5. Sea la densidad:

x	−2	−1	0	1	2
$f(x)$	0.1	0.2	0.3	0.2	0.2

Hallar $E(X)$ y μ.

6. La tabla siguiente nos muestra la densidad para la variable aleatoria X, número de hembras adultas en un grupo de monos aulladores:

x	1	2	3	4	5
$f(x)$	0.1	0.15	0.5	0.15	0.1

Hallar el número promedio de hembras adultas, por grupo.

7. *La esperanza de las funciones de X.* Si X es una variable aleatoria, entonces X^2, X - 1, $\sqrt{X}$, y muchas otras variables que podemos expresar en función de X, son también variables aleatorias. Cada una tendrá una esperanza o promedio teórico a largo plazo. Si $H(X)$ es una función de X, entonces su esperanza viene expresada por,

$$E[H(X)] = \sum_{\text{todo } x} H(x)\, f(x)$$

Ilustraremos este concepto hallando $E(X^2)$, sólo para el caso del problema del dado del Ejemplo 4.2.2. Para este ejercicio,

$$\begin{aligned} E[X^2] &= \sum_{\text{todo } x} x^2 f(x) \\ &= 1^2 \cdot \tfrac{1}{6} + 2^2 \cdot \tfrac{1}{6} + 3^2 \cdot \tfrac{1}{6} + 4^2 \cdot \tfrac{1}{6} + 5^2 \cdot \tfrac{1}{6} + 6^2 \cdot \tfrac{1}{6} \\ &= \tfrac{91}{6} \end{aligned}$$

a) Hallar $E(X^2)$ para la variable aleatoria del Ejercicio 1.
b) Hallar $E(X^2)$ para la variable aleatoria del Ejercicio 2.
c) Hallar $E(X^2)$ para la variable aleatoria del Ejercicio 3.
d) Hallar $E(X^2)$ para la variable aleatoria del Ejercicio 4.
e) Hallar $E[(X - \mu)^2]$ para la variable aleatoria del Ejercicio 5.
f) Hallar $E[(X - \mu)^2]$ para la variable aleatoria del Ejercicio 6.

8. *Varianza.* Recordemos de la Sección 1.5 que la medida de variabilidad de la población más usual es la varianza poblacional. Representábamos este parámetro mediante la letra griega σ^2. Su valor, como μ, no puede hallarse a partir de una muestra aunque puede

estimarse mediante s^2. No obstante, si disponemos de la densidad real para X, podemos calcular σ^2, definida por $\sigma^2 = \text{Var}\, X = E[(X - \mu)^2]$. Sea la densidad

x	1	2	3
f(x)	0.4	0.2	0.4

Aquí, por observación, $\mu = 2$. Y de ello deducimos,

$$\begin{aligned}\sigma^2 &= E[(X - 2)^2] \\ &= (1 - 2)^2(0.4) + (2 - 2)^2(0.2) + (3 - 2)^2(0.4) \\ &= 0.8\end{aligned}$$

a) ¿Cuál es la varianza de la variable aleatoria X en el Ejercicio 5?
b) ¿Cuál es la varianza de la variable aleatoria X en el Ejercicio 6?
c) Hallar la varianza de la variable aleatoria X del Ejemplo 4.2.2.
d) Hallar la varianza de la variable aleatoria X del Ejemplo 4.2.3.

9. *Método simplificado para calcular la* σ^2. En el Ejercicio 8, se define la varianza de una variable aleatoria X. Es fácil calcular σ^2 a partir de la definición cuando μ es un número entero y X no toma muchos valores. En este ejercicio, se presenta un método alternativo para calcular cómodamente σ^2. Esta fórmula simplificada es

$$\sigma^2 = E[(X - \mu)^2] = E[X^2] - (E[X])^2$$

Para ilustrar el método, consideremos la densidad del Ejercicio 8. En este caso

$$\begin{aligned}E[X^2] &= (1)^2(0.4) + (2^2)(0.2) + 3^2(0.4) \\ &= 4.8\end{aligned}$$

y

$$\sigma^2 = 4.8 - (2)^2 = 0.8$$

a) Utilizar el método simplificado para verificar las respuestas de los apartados *a* a d del Ejercicio 8.
b) Considerar la densidad del Ejercicio 1. Encontrar Var X.
c) Considerar la densidad del Ejercicio 2. Hallar σ^2.

10. *Desviación típica.* Recuerde de la Sección 1.5 que la desviación típica muestral es la raíz cuadrada positiva de la varianza muestral. Teóricamente, la desviación típica σ se define como la raíz cuadrada positiva de σ^2. Es decir, $\sigma = \sqrt{\sigma^2}$.
a) Considerar la densidad del Ejercicio 1. Hallar σ.
b) Considerar la densidad del Ejercicio 2. Hallar σ.
c) ¿Qué magnitud física podemos asociar a σ en el apartado a? ¿Y en el *b?*

11. Tres pacientes reciben inyecciones de desensibilización contra las picaduras de insect tos. Se calcula que este suero tiene una eficacia del 95 %. Sea X el número de pacientes desensibilizados.
a) Utilice un diagrama de árbol para obtener la tabla de $f(x)$.
b) Hallar e interpretar $E[X]$.
c) Hallar μ.
d) Hallar $E[X^2]$.
e) Hallar Var X y σ.

12. Algunos genes experimentan una desviación tan extrema de su estructura normal, que el organismo es incapaz de sobrevivir. Estos genes reciben el nombre de genes letales. Un ejemplo de esto es el gen que produce un pelaje amarillo en los ratones, llamémoslo *Y.* Este gen es dominante con respecto al que expresa el color gris, *y.* La teoría genética habitual postula que, para dos ratones amarillos, heterocigóticos para este carácter *(Yy),* \ de las crías serán grises y $\frac{3}{4}$ serán amarillas. Los biólogos han observado que estas proporciones previsibles no se dan en la realidad y que los verdaderos porcentajes obtenidos son $\frac{1}{3}$ grises y $\frac{2}{3}$ amarillos. Esta variación tiene una explicación por el hecho de que de los embriones, aquellos que son homocigotos para el amarillo *(YY),* no llegan a desarrollarse. Con esto, nos quedan sólo dos genotipos, *Yy y yy,* con una relación de 2 es a 1, siendo el primero un ratón de pelo amarillo. Por este motivo se dice que el gen *Y* es letal.
    a) En el caso de que dos ratones heterocigotos amarillos se acoplen, utilice un diagrama de árbol para comprobar que la teoría genética es capaz de prever una relación de 3 a 1, para ratones amarillos y grises.
    b) Se lleva a cabo un experimento de reproducción en el que una pareja de ratones amarillos heterocigotos se acopla. Consideremos tres crías. Con *X* indicaremos el número de ratones amarillos que encontramos. La densidad para *X* es

x	0	1	2	3
$f(x)$	$\frac{1}{27}$	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{8}{27}$

    Compruebe los valores de esta tabla.
    c) Halle el número esperado de ratones amarillos entre las crías, en una carnada de tamaño 3. Halle también la varianza y la desviación típica de *X.*
13. *La desigualdad de Chebyshev.* Esta desigualdad nos indica otra propiedad muy útil de la desviación típica. En particular nos dice que «la probabilidad de que cualquier variable aleatoria *X* caiga a una distancia máxima de su media de *k* desviaciones típicas, es por lo menos de $1 - 1/k^2$». Por ejemplo, si sabemos que *X* tiene de media 3, con desviación típica 1, podemos llegar a la conclusión de que la probabilidad de que *X* caiga entre 1 y 5 *(k = 2* desviaciones típicas de la media) es por lo menos de $1 - 1/2^2 = 0.75$.
    a) Sea *X* la variable que representa la cantidad de lluvia caída de una semana en una región determinada. Supongamos que $\mu = 1.00$ y $a = 0.25$ pulgadas. ¿Sería extraño que esta región registre más de 2 pulgadas de agua durante una semana? Razónelo basándose en la desigualdad de Chebyshev.
    b) Sea *X* el número de casos de rabia registrados a la semana en un determinado Estado. Supongamos que $\mu = \frac{1}{2}$ y $\sigma^2 = \frac{1}{25}$. ¿Podría considerarse infrecuente registrar dos casos en una misma semana? Razónelo basándose en la desigualdad de Chebyshev.
14. Se lleva a cabo un estudio comparativo de dos fármacos destinados a mantener un ritmo cardíaco constante en pacientes que ya han sufrido un infarto. Sea *X* el número de latidos por minuto, registrado mediante la utilización del fármaco A, e *Y* el número de latidos registrado con el fármaco B. Utilice las densidades hipotéticas siguientes:

x	40	60	68	70	72	80	100
$f(x)$	0.01	0.04	0.05	0.80	0.05	0.04	0.01

y	40	60	68	70	72	80	100
$f(y)$	0.40	0.05	0.04	0.02	0.04	0.05	0.40

a) Mediante observación, halle el ritmo cardíaco medio para cada fármaco. ¿Existe alguna diferencia entre los ritmos cardíacos provocados por los dos fármacos?
b) Mediante observación, ¿cuál de los dos fármacos provocará una mayor variación en el ritmo cardíaco? Compruebe su respuesta calculando la Var X y la Var Y, y comparando los valores de estos dos parámetros.
c) Halle σ_X y σ_Y. ¿Qué magnitud física podemos asociar a estas desviaciones típicas?
d) Utilizando la desigualdad de Chebyshev, ¿entre qué valores oscilará el ritmo cardíaco del 75 % de los pacientes tratados con el fármaco A? ¿Qué valores obtendremos con el fármaco B? (Véase el Ejercicio 13.)

4.3. LA FUNCIÓN DE DISTRIBUCIÓN ACUMULADA

La segunda función que utilizaremos en el cálculo de probabilidades es la función de distribución acumulada F. Esta es el equivalente teórico de la distribución de frecuencias relativas acumuladas, ya comentada en la Sección 1.3. En el caso discreto, podemos hallarla sumando los valores de la tabla de densidades. Es importante entender esta función ya que las tablas que se utilizan a lo largo del texto son tablas acumuladas. Daremos a continuación la definición de F.

Definición 4.3.1. La función de distribución acumulada. Sea X una variable aleatoria con densidad f La *función de distribución acumulada* de X, representada por F, se define como

$$F(x) = P[X \leq x] \qquad \text{para } x \text{ real}$$

Tomemos un valor real concreto x_0. En el caso discreto, $P[X \leqslant x_0] = F(x_0)$ se calcula sumando la densidad f para todos los valores de X menores o iguales al valor x_0. Así,

$$F(x_0) = \sum_{x \leqslant x_0} f(x)$$

Ejemplo 4.3.1. Sea la variable aleatoria X, número de mimetismos que escapan a la detección del Ejemplo 4.2.1. La densidad f para X viene dada por

x	0	1	2	3
$P[X=x]=f(x)$	$\frac{8}{1000}$	$\frac{96}{1000}$	$\frac{384}{1000}$	$\frac{512}{1000}$

La función de distribución acumulada (o función de distribución) para X se expresa por,

X	0	1	2	3
$P[X \leqslant x] = F(x)$	$\frac{8}{1000}$	$\frac{104}{1000}$	$\frac{488}{1000}$	$\frac{1000}{1000}$

Supongamos que queremos utilizar la función de distribución acumulada para calcular la probabilidad de encontrar entre 1 y 3 mimetismos que escapan a la detección. Esto es, queremos hallar $P[1 \leqslant X \leqslant 3]$. Para ello, volveremos a escribir la expresión como sigue:

$$P[1 \leqslant X \leqslant 3] = P[X \leqslant 3] - P[X \leqslant 0]$$

De esta forma, es evidente que

$$\begin{aligned} P[1 \leqslant X \leqslant 3] &= F(3) - F(0) \\ &= \frac{1000}{1000} - \frac{8}{1000} = \frac{992}{1000} = 0.992 \end{aligned}$$

Para hallar $P[X \geqslant 2]$, volveremos a escribir la expresión de esta forma,

$$\begin{aligned} P[X \geqslant 2] &= 1 - P[X \leqslant 1] \\ &= 1 - F(1) \\ &= 1 - \frac{104}{1000} = \frac{896}{1000} \end{aligned}$$

Observe que F realiza lo que su nombre, función de distribución acumulada, implica: sumando o acumulando probabilidades hasta alcanzar e incluir el valor de interés.

EJERCICIOS 4.3

1. La tabla siguiente muestra la densidad para la variable aleatoria X, número de aleteos por segundo de ciertas especies de grandes polillas mientras vuelan.

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	0.1

   a) Hallar la tabla para la función de distribución acumulada F.
   b) Utilizar F para calcular $P[X \leqslant 8]$.
   c) Utilizar F para calcular $P[X > 7]$.
   d) Utilizar F para calcular $P[7 \leqslant X \leqslant 9]$

2. La densidad para la variable aleatoria X, número de personas por día que buscaron tratamiento innecesario en un servicio de urgencias de un pequeño hospital, viene dada por:

x	0	1	2	3	4	5
$f(x)$	0.01	0.1	0.3	0.4	0.1	0.09

   a) Construir la tabla para la función de distribución acumulada.
   b) Hallar $P[X \leqslant -2]$.
   c) Utilizando F, calcular $P[2 \leqslant X \leqslant 4]$.

d) Hallar $P[X \leqslant 6]$.
e) Calcular $P[X = 3]$.
f) Utilizando F, hallar la probabilidad de que más de dos recurrieran innecesariamente al auxilio de un servicio de urgencias.

3. Se clasifican las células situadas en secciones de tejidos dañados, examinados con el microscopio, de acuerdo con la extensión del daño mediante la siguiente escala: 0, no dañadas; 1, débilmente dañadas; 2, moderadamente dañadas; 3, extremadamente dañadas. Las células de un tejido, expuestas a 20 minutos de anoxia (un suministro de oxígeno por debajo del normal) antes de ser preparadas para un estudio con el microscopio, dan lugar a la siguiente densidad, donde x es el valor de clasificación según el daño:

x	0	1	2	3
$f(x)$	0.15	0.25	0.50	0.10

a) Construir la tabla para la función de distribución acumulada F.
b) Utilizar F para hallar la probabilidad de que una célula aleatoriamente seleccionada esté sólo ligeramente dañada o no dañada.
c) Utilizar F para obtener la probabilidad de que se observe al menos un daño moderado en una célula seleccionada aleatoriamente.

4. Los injertos, unión del tronco de una planta con el tronco o la raíz de otra, se utilizan comercialmente con gran frecuencia para hacer crecer el tronco de una variedad que produce fruta fina, sobre el sistema radical de una variedad más robusta. La mayor parte de las naranjas dulces de Florida crece en árboles injertados a la raíz de una variedad de naranja ácida. Se lleva a cabo un experimento con cinco injertos de este tipo. La densidad de X, número de injertos que fracasan, viene dada por:

x	0	1	2	3	4	5
$f(x)$	0.7	0.2	0.05	0.03	0.01	0.01

a) Construir la tabla para F.
b) Utilizar F para hallar la probabilidad de que fracasen como máximo tres injertos.
c) Utilizar F para hallar la probabilidad de que fracasen al menos dos injertos.
d) Utilizar F para comprobar que la probabilidad de tres fracasos es de 0.03.

5. Sea X el número de casos nuevos de SIDA diagnosticados en un importante hospital, durante un día. La distribución acumulada para X se supone que es

x	0	1	2	3	4	5	6
$F(x)$	0.1	0.2	0.3	0.6	0.8	0.9	1.00

a) Hallar la probabilidad de que en un día cualquiera,
   i. Sean diagnosticados tres casos nuevos, a lo sumo,
   ii. Por lo menos sea diagnosticado un caso nuevo.

iii. Ningún caso nuevo sea diagnosticado.
iv. Sean diagnosticados entre dos y cuatro casos nuevos, ambos inclusive.

b) Hallar la densidad para X.
c) Calcular la media de casos diagnosticados al día.
d) Hallar σ^2.
e) Calcular la desviación típica de X. ¿Qué magnitud física podemos asociar a σ?

6. Sea F la función de distribución acumulada para una variable aleatoria discreta X y x_0 el mayor valor que X puede tomar. ¿Cuánto vale $F(x_0)$? Explicar el razonamiento.

4.4. LA DISTRIBUCIÓN BINOMIAL

En las Secciones 4.1 a 4.3 hemos estudiado con algún detalle las propiedades generales de las variables aleatorias discretas. Ahora desarrollaremos un tipo específico de variable discreta, la variable aleatoria binomial. Las variables binomiales surgen en conexión con experimentos que, a primera vista, pueden parecer de naturaleza completamente diferente. En cualquier caso, un minucioso examen demuestra ciertos rasgos comunes subyacentes. Observemos los siguientes experimentos.

Ejemplo 4.4.1

a) Un varón y una mujer, cada uno con un gen recesivo (azul) y uno dominante (marrón) para el color de los ojos, son padres de tres hijos. ¿Cuál es la distribución de probabilidades para el número de hijos con los ojos azules?

b) Un portador de tuberculosis tiene un 10 % de posibilidades de transmitir la enfermedad a alguien que no haya estado previamente expuesto a ella y con el que entre en contacto directo. Durante el transcurso de un día, un portador entra en contacto con diez de tales individuos. ¿Cuántos se espera que contraigan la enfermedad de este modo?

c) Se está desarrollando una nueva variedad de maíz en una extensión de experimentación agrícola. Se espera que tenga una tasa de germinación del 90 %. Para verificar esto, se plantan 20 semillas en suelos de idéntica composición y se les dedican los mismos cuidados. Si la cifra 90 % es correcta, ¿cuántas semillas se espera que germinen? Si sólo germinan 15 o menos, ¿hay razón para sospechar de la cifra 90%?

d) Se está llevando a cabo un estudio de opinión pública relativo a la conveniencia de construir una presa para controlar inundaciones en New River Valley. Hay que elegir aleatoriamente y preguntar a quince residentes del área. Si resulta que un 80 % de la gente que vive en el área se opone a la presa, ¿cuál es la probabilidad de que una mayoría de las personas preguntadas esté en contra? ¿Cuál es la probabilidad de que entre 10 y 14, ambos inclusive, se opongan a la construcción?

¿Qué tienen en común estos experimentos, que aparentemente no guardan relación alguna? Hay cuatro puntos esenciales que debemos observar:

1. *Se puede considerar cada uno como compuesto por un número fijo de pruebas idénticas* n. En el Ejemplo 4.4. la, el nacimiento de un niño sería una prueba, $n = 3$. En el Ejemplo 4.4. 1*b,* la prueba consistiría en observar a un individuo que entre en contacto con un portador de tuberculosis para ver si contrae la enfermedad, $n = 10$. Una prueba en el Ejemplo 4.4.1c consistiría en la observación de una semilla de maíz para ver si germina, $n = 20$. En el Ejemplo 4.4.1d, una prueba consistiría en determinar la opinión de un residente con respecto a la construcción de la presa, $n = 15$.
2. *El resultado de cada prueba puede clasificarse como «éxito» o «fracaso».* Generalmente, el «éxito» se define como la constatación de la característica que se esté contabilizando. De este modo, en el Ejemplo 4.4.1 el éxito consistiría en obtener un

niño con ojos azules, en observar que un individuo contrae la tuberculosis, en ver que una semilla de maíz germina, y en encontrar a un residente de New River Valley que se oponga a la construcción de la presa.

3. *Las pruebas son independientes en el sentido de que el resultado de una prueba no tiene efecto sobre el resultado de cualquier otra prueba, y la probabilidad de éxito* p *continúa siendo la misma de una prueba a otra.* En el Ejemplo 4.4. 1*a,* estas condicionnes se satisfacen obviamente con $p = \frac{1}{4}$. Puesto que el hecho de que una persona sea susceptible de contraer la tuberculosis no tendrá influencia sobre la susceptibilidad dé otra; en el Ejemplo 4.4.1*b* la independencia puede darse por admitida con $p = 0.1$. Si suponemos que las semillas del Ejemplo 4.4.1c se plantan de modo que el crecimiento de una de ellas no impida el de cualquiera de las otras, las pruebas serán independientes con $p = 0.9$. En el Ejemplo 4.4. 1*d* hay ciertos aspectos discutibles. Los sondeos de opinión suponen generalmente muestreo sin reemplazamiento. Una vez que un individúo ha sido encuestado se le extrae de la población. De este modo, la composición de la población cambia, y la probabilidad de éxito cambia algo de prueba a prueba. En todo caso, si el grupo que esté siendo muestreado es grande, como es generalmente el caso, entonces el cambio es tan insignificante como para no tenerlo en cuenta. Concluiremos que en la práctica tenemos independencia con $p = 0.8$.
4. *La variable de interés es el número de éxitos en* n *pruebas.*

Los cuatro puntos que acabamos de tratar son los supuestos generales subyacentes al *modelo binomial.* Cualquier variable aleatoria X que represente el número de éxitos en n pruebas idénticas e independientes, con probabilidad de éxito p, constante de una prueba a otra, se llama *variable aleatoria binomial* con parámetros n y p. Obsérvese que, para que una variable aleatoria binomial quede definida, el número de pruebas y la probabilidad del éxito deben ser conocidos. Esta información es necesaria para utilizar las tablas binomiales o para encontrar probabilidades binomiales por medio de paquetes estadísticos o calculadoras estadísticas. Obsérvese además que p, la probabilidad de éxito, también representa la proporción de los ensayos que se espera que den lugar a éxito.

Antes de estudiar la densidad para una variable aleatoria binomial, repasaremos primero dos conceptos desarrollados en la Sección 2.5. Tenemos antes que recordar el significado del término *n factorial* (Definición 2.5.1), así como la fórmula empleada en el recuento de ordenaciones hechas en los objetos, cuando éstos son indistinguibles (Teorema 2.5.1).

Definición 4.4.1. Sea n un entero positivo. Mediante *n factorial,* representado por n!, queremos decir $n(n - 1)(n - 2) \ldots 3 \cdot 2 \cdot 1$. *Cero factorial,* representado por 0!, se define como 1.

Por ejemplo, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ y $7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$.

Consideremos ahora una secuencia de n objetos de los cuales x son de un tipo y el resto, $n - x$, pertenece a otro. Expresaremos a continuación el número de maneras de ordenar estos n objetos formando diferentes conjuntos reconocibles,

$$\frac{n!}{x!(n - x)!}$$

Tomaremos como ejemplo un experimento realizado con 10 pacientes a los que se les somete a un análisis de colesterol. Si a indica un resultado con un valor de colesterol alto, y n un resultado con un valor normal, entonces la secuencia *annaaanaaa* representa una de

las formas en las que pueden presentarse 7 resultados altos y 3 normales. Este mismo resultado puede conseguirse con la secuencia *nnaaanaaaa.* La fórmula arriba mencionada implica que hay

$$\frac{10!}{7!\ 3!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{7!\,3 \cdot 2 \cdot 1} = 120$$

de tales secuencias. Este concepto se utilizará en el desarrollo de la densidad binomial.

Para contestar a las cuestiones planteadas en el Ejemplo 4.4.1 o cualquier otro problema probabilístico concerniente a este tipo de variables, debemos contar con las densidades adecuadas. Estudiaremos con detalle el Ejemplo 4.1.4a, cuyo número de pruebas es pequeño, para ver de qué densidades se trata. Si podemos hallar un modelo, podrá ser generalizado y nos permitirá calcular las densidades para las variables de los demás experimentos.

Ejemplo 4.4.2. Un varón y una mujer, cada uno con un gen recesivo y uno dominante para el color de los ojos, son padres de tres hijos. ¿Cuál es la distribución de probabilidades para *X,* número de hijos con ojos azules?

Esto se puede considerar como un proceso en tres etapas. El espacio muestral y la densidad de *X* se hallan considerando el árbol de la Figura 4.2. En el árbol, *b* representa el nacimiento de un niño con ojos azules y *B,* el nacimiento de un niño con ojos marrones. Puesto que el color de ojos de un niño no influye sobre el que pueda tener cualquier otro, las pruebas serán independientes, y la probabilidad de la trayectoria se hallará multiplicando las probabilidades que aparecen a lo largo de ella. Así, la densidad para *X* viene dada por:

x	0	1	2	3
$f(x)$	$1\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^3$	$3\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^2$	$3\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^1$	$1\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^0$

Para expresar esta densidad por medio de una ecuación, solamente es preciso buscar modelos. Es evidente que las proporciones de $\frac{1}{4}$ para el éxito y $\frac{3}{4}$ para el fracaso aparecen en cada probabilidad enunciada, siendo x el exponente para la proporción de éxito. También la suma de los dos exponentes en cada caso es 3, número de pruebas. De este modo, en gene-

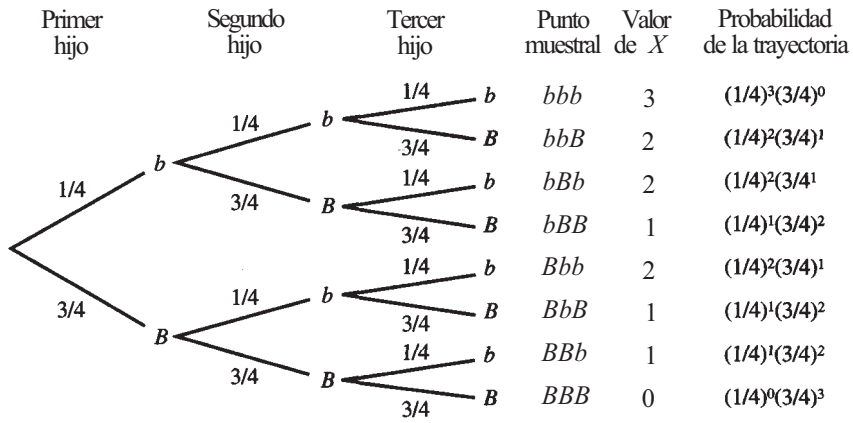


Figura 4.2. Color de los ojos en una familia de tres hijos.

ral, el exponente asociado con la proporción de fracaso es 3 - *x.* Por lo tanto, la forma general de la densidad será

$$f(x) = k(x)\left(\tfrac{1}{4}\right)^x\left(\tfrac{3}{4}\right)^{3-x} \qquad x = 0,\ 1,\ 2,\ 3$$

donde *k(x)* es un coeficiente cuyo valor depende del valor de *X* en cuestión. Es decir, los coeficientes no son necesariamente iguales para cada valor de *x.* En el ejemplo anterior, cuando $x = 0$ ó $x = 3$, $k(x) = 1$; cuando $x = 1$ ó $x = 2$, $k(x) = 3$. La única pregunta que hay que responder es, ¿qué es *k(x)?* ¿Existe una fórmula para calcularlo? Este coeficiente cuenta el número de trayectorias del árbol en las que se incluye un valor concreto de *X.* Una trayectoria es simplemente una permutación de tres letras, de las cuales *x* son *b* y 3 - *x* son *B.* Si utilizamos la fórmula para hallar el número de permutaciones de objetos indistinguibles, veremos que

$$k(x) = \frac{3!}{x!(3-x)}$$

De este modo, la expresión de la densidad para *X* es

$$f(x) = \frac{3!}{x!(3-x)!}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x} \qquad x = 0,\ 1,2,\ 3$$

El método utilizado para obtener la densidad de *X* en el Ejemplo 4.4.2 es independiente del número de pruebas implicadas y del valor numérico de la tasa de éxito. Dado cualquier número de pruebas *n* y cualquier tasa de éxito *p,* se puede emplear un razonamiento análogo para obtener la densidad. Esta tendrá la misma forma general que la del Ejemplo 4.4.2, sólo que reemplazando el número de pruebas, 3, por *n y* la tasa de éxitos, $\frac{1}{4}$, por *p.* Esto se recoge: en el Teorema 4.4. 1.

Teorema 4.4.1. Sea *X* una variable aleatoria binomial con parámetros *n y p.* La densidad para *X* viene dada por

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad x = 0,\ 1,\ 2,\ 3,\ ...,\ n$$

Esperanza y varianza: binomial

Consideremos la segunda de las preguntas que se plantean en el Ejemplo 4.4.1. Es decir, dada una variable aleatoria binomial *X,* con parámetros *n y p,* ¿cuál es el valor esperado de X? Una vez más nos referiremos a un ejemplo numérico para orientarnos.

Ejemplo 4.4.3. Diez individuos, cada uno de ellos propenso a la tuberculosis, entran en contacto con un portador de la enfermedad. La probabilidad de que la enfermedad se contagie del portador a un sujeto cualquiera es de 0.10. ¿Cuántos se espera que contraigan la enfermedad?

Puesto que cada individuo tiene un 10% de posibilidades de contraer la enfermedad, el sentido común nos conduce a esperar que se contagiará un 10 % de las personas expuestas. Es decir, el sentido común señala 10(0.1) = 1 como número esperado de los que contraen tuberculosis. Obsérvese que este valor se ha obtenido multiplicando el número de pruebas, 10, por la tasa de éxitos, 0.1.

Una vez más, el razonamiento empleado para responder a la pregunta es independiente del número real de pruebas implicadas o del valor numérico de la tasa de éxito. Ello sugiere al menos una parte del Teorema 4.4.2. Que la varianza es como se ha establecido no podría deducirse desde un razonamiento intuitivo. Dado que el desarrollo de la fórmula de la varianza no es realmente instructivo, utilizaremos el resultado sin prueba. El Ejercicio 11 nos pide que comprobemos, de forma numérica, que la fórmula es correcta.

Teorema 4.4.2. Sea X binomial con parámetros n y p. Entonces E[X]=np y Var X=np(1 -*p).*

Cálculo de probabilidades binomiales: distribución acumulada

La densidad para una variable aleatoria discreta proporciona la probabilidad de que la variable aleatoria tome un valor específico. En el caso de una variable aleatoria binomial X hay una fórmula para la densidad. En la práctica, se necesita frecuentemente conocer la probabilidad de que X tome un valor mayor o igual, o menor o igual que uno dado. Esto es, $P[X \leqslant x]$ o $P[X \geqslant x]$. Estas probabilidades se calculan hallando la densidad para cada x y sumándole las probabilidades individuales de la Sección 4.2. Sin embargo, hay un camino más fácil utilizando la función de distribución acumulada. Recuerde que esta función, F, es análoga a la noción de frecuencia relativa acumulada dada en la Sección 1.3. y proporciona la probabilidad de que X tome valores menores o iguales que otro especificado x. Esta probabilidad se halla sumando la densidad de X para todos los valores menores o iguales que el especificado.

Puesto que la variable aleatoria binomial es de gran aplicación, existen tablas y programas para ordenadores y calculadoras con los que podemos calcular F. De este modo, la primera misión de los investigadores es darse cuenta de que están tratando con una variable binomial, a fin de que formulen sus preguntas en términos de probabilidades y, a continuación, utilizar correctamente las tablas disponibles para contestar a tales preguntas. La Tabla I del Apéndice B es una tabla binomial abreviada. Muestra valores de la función de distribución acumulada para variables binomiales con $n = 5$ a 20 y $p = 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9$. Hay tablas más extensas. Son todas semejantes a la presentada y el saber manejar la Tabla I que aparece en el Apéndice B facilita el manejo de tablas más completas con poca dificultad. Exponemos su uso en el Ejemplo 4.4.4.

Ejemplo 4.4.4. Se está desarrollando una nueva variedad de maíz en una estación agrícola experimental. Se espera que germine el 90 % de las semillas. Para verificar eso, se plantan 20 semillas en un suelo de idéntica composición y se le dedican los mismos cuidados. Si la cifra 90 % es correcta, ¿cuántas semillas se espera que germinen?

Esta pregunta no es nueva. Se nos pide que encontremos $E[X]$, en donde X es el número de semillas que germinan. Del Teorema 4.4.2, $E[X] = np = 20(0.9) = 18$. Si la cifra 90% es correcta, esperaríamos obtener alrededor de 18 semillas germinadas. Si germinan 15 semillas por lo menos, ¿hay razón para sospechar de la cifra 90 %? En principio, parece haber razones para dudar, porque 15 es algo menos de lo que se esperaba, 18. La cuestión es, si la tasa de 90 % para el éxito es correcta, ¿cuál es la probabilidad de ver 15 ó menos semillas germinadas? Si esta probabilidad es grande, entonces no hay razón para sospechar del valor 90 %. En todo caso, si esta probabilidad es pequeña, hay dos posibles explicaciones. O se ha presentado un suceso raro, o la tasa real de germinación es menor que la que defendíamos. Por lo tanto, nuestra decisión se basa en la determinación e interpretación de $P[X \leqslant 15]$. Utilizando la función de distribución acumulada, tenemos:

$$P[X \leqslant 15] = F(15) = \sum_{x=0}^{15} f(x) = \sum_{x=0}^{15} \frac{20!}{x!(20-x)!} (0.9)^x (0.1)^{20-x}$$

El cálculo de esta probabilidad entraña una exagerada cantidad de cálculos aritméticos. Así que volvemos a la Tabla I del Apéndice B. Puesto que la Tabla I nos da una relación directa de los valores de la función de distribución acumulada, podemos encontrar la respuesta a nuestra pregunta mirando al grupo de valores de $n = 20$. La probabilidad buscada es de 0.0432 y se halla en la columna 0.9 y en la fila 15. Es decir:

$$P[X \leqslant 15] = F(15) = 0.0432$$

¿Este valor es grande o pequeño? La respuesta no está claramente definida. Muchos investigadores tenderían a considerarlo pequeño y llegarían a la conclusión de que la tasa de germinación de 0.9 que establecimos es demasiado alta.

La Tabla I no puede responder directamente a preguntas que no sean de la forma $P[X \leqslant x]$. Este tipo de preguntas deberán reescribirse primero en términos de función de distribución acumulativa. Este punto se explica en el Ejemplo 4.4.5.

Ejemplo 4.4.5. Se está llevando a cabo un sondeo para determinar la opinión con respecto a la construcción de una presa para controlar inundaciones en New River Valley. Hay que elegir aleatoriamente y estudiar a quince residentes del área. Si resulta que un 80 % de la gente que vive en el área se opone a la presa, ¿cuál es la probabilidad de que la mayoría de las estudiados esté en contra? Puesto que ocho o más individuos representan mayoría, lo que se nos pide es que calculemos $P[X \geqslant x]$.

La probabilidad buscada está representada en la Figura 4.3. Fíjese que lo que buscamos es la probabilidad asociada a los puntos marcados. Para ello, nos basaremos en el hecho de que la probabilidad total asociada a los puntos 0, 1, 2, 3,…, 15 es 1. Podemos hallar la probabilidad deseada restando a 1 la probabilidad asociada a los puntos no deseados, de 0 a 7 incluidos.

En términos de la función de distribución acumulada,

$$P[X \geqslant 8] = 1 - P[X \leqslant 7] = 1 - F(7)$$

En la Tabla I con $n = 15$ y $p = 0.8$ puede verse que $F(7) = 0.0042$. De este modo, la probabilidad de que la mayoría de los muestreados se oponga a la construcción de la presa es de $1 - 0.0042 = 0.9958$. ¿Cuál es la probabilidad de que entre 10 y 14, ambos inclusive, se opongan a la construcción? Expresando esta pregunta en términos de la función de distribución acumulada, vemos que:

$$\begin{aligned} P[10 \leqslant X \leqslant 14] &= P[X \leqslant 14] - P[X \leqslant 9] \\ &= F(14) - F(9) \\ &= 0.9648 - 0.0611 \\ &= 0.9037 \end{aligned}$$

EJERCICIOS 4.4

1. En cada uno de los siguientes puntos, se describe una variable aleatoria. Decir en cada caso si la variable es binomial, aproximadamente binomial o no binomial. Si la variable

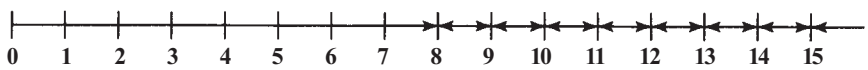


Figura 4.3. La probabilidad total es 1. $P[X \geqslant 8]$ es la probabilidad asociada a los puntos señalados, $P[X \geqslant 8] = 1 - P[X \leqslant 7]$.

es binomial o aproximadamente binomial, determinar los valores numéricos de n y p. Si la variable no es binomial, ¿cuál de los supuestos binomiales se contradice? (Por *aproximadamente binomial* entendemos que, si bien p puede variar un poco de prueba a prueba, el cambio es tan pequeño que puede despreciarse. De modo que las probabilidades calculadas utilizando la densidad binomial, aunque no son exactas, son buenas aproximaciones a las efectivamente implicadas.)

a) Se ha producido un accidente y se necesitan unidades de sangre AB negativa. No hay ninguna en el laboratorio. Se consiguen cinco empleados no emparentados, como posibles donantes, y se determinan sus grupos sanguíneos. La probabilidad de que un individuo elegido al azar tenga su grupo sanguíneo AB negativo es 0.006. X es el número de empleados con sangre AB negativo.

b) En el código RNA, UGG codifica el triptófano y UGA codifica una parada. En un determinado segmento, aparece cinco veces la palabra UGA. Supóngase que los nucleótidos U y G no sufren mutación, pero que el nucleótido A (adenina) muta a G (guanina) el 0.1 % de las veces. X es el número de mutaciones en la secuencia en que la palabra parada (UGA) muta a triptófano (UGG).

c) Una determinada reacción química presenta un rendimiento habitual del 70 %. Se idea un nuevo proceso que incrementará dicho rendimiento. Los que lo proponen reivindican que en el 90 % de los casos el nuevo proceso produce un rendimiento mayor que el anterior. Se prueba el nuevo proceso en diez ocasiones, y se registran los rendimientos. La variable X es el número de veces que el rendimiento ha mejorado sobre la cifra 70 %.

d) Un biólogo dispone de ocho plantas para experimentación. El experimento a realizar requiere el empleo de cuatro plantas solamente. El biólogo ignora que tres de las plantas están enfermas. Selecciona aleatoriamente cuatro plantas. La variable X es el número de plantas enfermas seleccionadas.

e) En el estudio de los hábitos migratorios de los gansos canadienses, se ha anillado aproximadamente el 5 % de la población total de estas aves. En un día determinado se capturan ocho gansos. X es el número de los que están anillados.

f) Una pareja se ha propuesto tener una hija. Decide continuar teniendo hijos hasta que nazca una hija, momento en el que ya no tendrá más. X es el número de hijos nacidos antes del nacimiento de la primera hija.

2. Considérese la variable aleatoria del Ejemplo 4.2.1.
   a) Justificar que X es binomial.
   b) ¿Cuáles son los valores de n y p?
   c) ¿Cuál es la fórmula para la densidad de X?
   d) Utilizar la fórmula de la densidad del apartado *c* para verificar las probabilidades dadas en el Ejemplo 4.2.1.
   e) Hallar $E[X]$ y comparar el resultado con el obtenido en el Ejemplo 4.2.3.
3. Una compañía de petróleos dispone de diez tanques distribuidos a lo largo de una extensa área del Golfo de México. Los oficiales creen que, en condiciones normales, cada tanque tiene sólo un 1 % de posibilidades de tener una pérdida de petróleo en todo el año. Sea X el número de tanques que han experimentado pérdidas durante el año.
   a) Fundamentar que X es binomial.
   b) Encontrar la expresión para la densidad.
   c) Hallar $E[X]$, Var $[X]$ y σ.
   d) Si los tanques están muy próximos y se produce alguna eventualidad (tal como un huracán o un terremoto), ¿es correcto suponer que X es binomial? Razonar la respuesta.

4. Para cada una de las variables X binomiales o binomiales aproximadas del Ejercicio 11, hallar:
   a) La expresión de la densidad.
   b) La media de X.
   c) La varianza de X.
   d) La desviación típica de X.
5. Sea X binomial con n - 4 y p = 0.2.
   a) Hallar una expresión para la densidad.
   b) Utilizar la densidad para hallar $P[X = 0]$.
   c) Utilizar la densidad para hallar $P[X \leqslant 1]$ ¿Por qué motivo no podemos hallar esta probabilidad a partir de la Tabla I del Apéndice B?
6. Un ingeniero de montes está estudiando la difusión del tizón en los pinos de las Great Smoky Mountains. Después de supervisar cinco áreas clave, el ingeniero anota 0 si no encuentra tizón y 1 si la enfermedad se ha presentado. Se introducen los datos en un ordenador que presenta una probabilidad de 0.001 de invertir un dígito al transmitir (leyendo un 0 como 1 o viceversa). X es el número de errores en la transmisión.
   a) Hallar la expresión de la densidad de X.
   b) Hallar $E[X]$, σ^2 y σ.
   c) Utilizar la densidad de X para hallar la probabilidad de que no se cometan errores de transmisión.
   d) Utilizar la densidad de X para hallar $P[X \leqslant 1]$.
7. Sea X binomial con n = 10 y p = 0.4. En cada caso esbozar un diagrama similar al de la Figura 4.3 y evaluar la probabilidad indicada, utilizando la Tabla I del Apéndice B.
   a) $P[X \leqslant 4]$.
   b) $P[X < 4]$.
   c) $P[X = 4]$.
   d) $P[X \geqslant 5]$.
   e) $P[X > 6]$.
   f) $P[3 \leqslant X \leqslant 6]$.
   g) $P[4 \leqslant X \leqslant 7]$.
   h) $P[3 \leqslant X < 6]$.
   i) $P[A < X \leqslant 7]$.
8. Los genetistas han identificado dos cromosomas sexuales, *R e Y, en* los seres humanos. Todo individuo tiene un cromosoma *R, y* la presencia de un cromosoma *Y* distingue al individuo como varón. Así que los dos sexos se caracterizan por *RR* (hembra) y *RY* (macho). El daltonismo es causado por un alelo recesivo en el cromosoma *R,* que denotamos por r. El cromosoma *Y* no tiene conexión alguna con el daltonismo. De modo que, en función de tal deficiencia, hay tres genotipos para las mujeres y dos para los varones:

Mujeres	Varones
RR (normal)	*RY* (normal)
Rr (portador)	*rY* (daltónico)
rr (daltónico)	

Cada hijo hereda al azar un cromosoma sexual de cada progenitor,

a) Un portador de daltonismo es padre de un hijo varón normal. Construir un diagrama de árbol para representar los posibles genotipos del hijo.

b) ¿Cuál es la probabilidad de que un determinado hijo nacido de la pareja sea un varón daltónico?

c) Si la pareja tiene tres hijos, ¿cuál es la probabilidad de que exactamente dos sean varones daltónicos?

d) Si la pareja tiene cinco hijos, ¿cuál es el número esperado de varones daltónicos? ¿Cuál es la probabilidad de que a lo sumo los dos sean varones daltónicos? ¿Cuál es la probabilidad de que tres o más sean varones daltónicos?

9. Se va a construir una planta nuclear y se quiere conocer la opinión de los vecinos de la localidad. Se selecciona una muestra aleatoria de 20 individuos y se realiza un sondeo. Se piensa que el 60 % de los habitantes del lugar estará a favor del proyecto. Si esto es verdad, ¿cuántos piensa usted que expresarían una opinión favorable? Si sólo nueve o menos son de tal opinión, ¿piensa usted que es razón de peso para poner en duda la cifra 60 %? Explicarlo sobre la base de la probabilidad implícita.
10. Para estudiar la regulación hormonal de una línea metabólica, se inyecta a ratas albinas un fármaco que inhibe la síntesis de proteínas del organismo. En general, 4 de cada 20 ratas mueren a causa del fármaco antes de que el experimento haya concluido. Si se trata a 10 animales con el fármaco, ¿cuál es la probabilidad de que al menos 8 lleguen vivos al final del experimento?
11. Utilice los ejercicios 7 y 9 de la Sección 4.2 para verificar que la varianza de la variable aleatoria X del Ejemplo 4.4.2 es $3 \cdot \frac{1}{4} \cdot \frac{3}{4} = \frac{9}{16}$, como se indica en el Teorema 4.4.2.

4.5. DISTRIBUCIÓN DE POISSON (OPCIONAL)

La segunda familia discreta considerada es la familia de Poisson, llamada así en honor al matemático francés Simeón Denis Poisson (1781-1840). Las variables aleatorias de Poisson surgen en conexión con los llamados *procesos de Poisson.* Estos implican la observación de un conjunto discreto de sucesos en un «intervalo» continuo de tiempo, longitud o espacio. Utilizamos la palabra *intervalo* en la descripción del proceso general de Poisson, bien entendido que no estamos tratando con un intervalo en el sentido matemático usual. Por ejemplo, podemos observar el número de leucocitos en una gota de sangre. El suceso de interés es la observación de un leucocito, mientras que el «intervalo» continuo implicado es una gota de sangre. Podemos observar también el número de veces que una planta de energía nuclear emite gases radiactivos en un período de tres meses. El suceso es la emisión de gases radiactivos. El intervalo continuo es el período de tres meses. Podríamos contabilizar el número de llamadas de emergencia recibidas cada noche por una brigada de rescate. En este caso, el suceso de valor discreto que queremos estudiar es la llegada de una llamada. El tiempo de observación es de una hora.

La variable aleatoria de interés en un proceso de Poisson es X, número de sucesos en un intervalo de tamaño s unidades. Para hallar la densidad de X, nos plantearemos tres preguntas:

1. ¿Cuál es la unidad de medida básica en este problema?
2. ¿Cuál es la media del número de ocurrencias del suceso por unidad? λ representa este valor.
3. ¿Cuál es el tamaño del intervalo de observación? Este valor está representado por s.

Podemos utilizar técnicas de cálculo para demostrar que la densidad de X viene representada por

$$f(x) = \frac{e^{-\lambda s}(\lambda s)^x}{x!} \qquad x = 0, 1, 2, 3,...$$

donde $e \simeq 2.7183$, y su esperanza es λs. Podemos comprobar, una vez que los valores λ y de s han sido hallados a partir del contexto físico del problema tratado, que las probabilidades son las calculadas utilizando la densidad ya mencionada. Una variable aleatoria cuya densidad adopta esta forma, recibe el nombre de *variable aleatoria de Poisson.* El Ejemplo 4.5.1 ilustra esta idea.

Ejemplo 4.5.1. El recuento de leucocitos de un individuo sano puede presentar en promedio un valor mínimo de hasta 6000 por mm^3 de sangre. Para detectar una deficiencia de leucocitos, se toma una gota de sangre de 0.001 mm^3 y se halla el número X de leucocitos. ¿Cuántos leucocitos cabe esperar en un individuo sano? Si a lo sumo se encuentran dos, ¿hay signos de una deficiencia de leucocitos?

Este experimento puede considerarse un proceso de Poisson. El suceso discreto de interés es encontrar un leucocito, el intervalo continuo es una gota de sangre. Sea el mm^3 la unidad de medida; así, s = 0.001 y λ, la media de veces que tendremos un suceso por cada unidad, es 6000. Por lo tanto, X es una variable aleatoria de Poisson con $E[X] = \lambda s$ ' = 6000(0.001) = 6. Para una persona con buena salud, esperaríamos observar un promedio de, por lo menos, seis leucocitos. ¿Sería extraño encontrar dos células como máximo? ¿Lo expresaremos por $P[X \leqslant 2 \mid \lambda s = 6]$?

Para contestar a esta pregunta utilizaremos la densidad

$$f(x) = \frac{e^{-6}6^x}{x!} \qquad x = 0, 1, 2, 3,...$$

Por sustitución,

$$P[X \leqslant 2] = \sum_{x=0}^{2} f(x) = \sum_{x=0}^{2} \frac{e^{-6}6^x}{x!}$$
$$= \frac{e^{-6}6^0}{0!} + \frac{e^{-6}6^1}{1!} + \frac{e^{-6}6^2}{2!}$$

La evaluación de este tipo de expresión requiere algunos cálculos aritméticos.

Aquí también, debido a la extensa demanda del modelo de Poisson, se han tabulado los valores de la función de distribución acumulada, para valores seleccionados del parámetro ,Ls. La Tabla II, en el Apéndice B, es una de esas tablas. La probabilidad buscada, 0.062, se calcula mirando la columna $\lambda s = 6$ y la fila 2. ¿Hay signos de una deficiencia de leucocitos? Puesto que 0.062 es un valor moderado, parece que la respuesta no está claramente determinada.

Mencionaremos otra importante aplicación de la densidad de Poisson. Considérese una variable binomial con n grande y p pequeña. En este caso, se puede demostrar que la densidad de Poisson con parámetro $\lambda s = np$ da una buena aproximación a la binomial. En la práctica, la aproximación de Poisson es buena si $n > 20$ y $p \leqslant 0.05$ y muy buena si $n \geqslant 100$ y np $\leqslant 10$. Puesto que la aproximación se utiliza siempre que la probabilidad p del suceso es pequeña, es frecuente aludir a la densidad de Poisson como función de distribución de sucesos «raros».

Ejemplo 4.5.2. En *Escherichia coli,* una bacteria que aparece con frecuencia en el tracto digestivo humano, una célula de cada 10^9 muta de ser sensible a la estreptomicina a ser resistente a ella. Esta mutación puede dar lugar a que el individuo implicado se vuelva resistente a la estreptomicina. Observando $2 \cdot 10^9$ de tales células, ¿cuál es la probabilidad de que ninguna mute? ¿Cuál es la probabilidad de que al menos una mute?

Este problema es efectivamente binomial, con $n = 2 \times 10^9$ y $p = 1/10^9$. Como $1/10^9$ es extremadamente pequeño, la mutación de una célula es un suceso muy raro. De modo que X,

número de células que mutan, puede considerarse como aproximadamente de Poisson con $\lambda s = np = (2 \times 10^9)(1/10^9) = 2$. De la Tabla II, en el Apéndice B, $P[X = 0] = 0.135$. La probabilidad de que se produzca al menos una mutación es $P[X \geq 1]$. Esta probabilidad se halla por sustracción. Es decir, $P[X \geq 1] = 1 - P[X = 0] = 1 - 0.135 = 0.865$.

EJERCICIOS 4.5

1. Una variable aleatoria de Poisson X tiene parámetro $\lambda s = 10$.
   a) Hallar $E[X]$.
   b) Encontrar la expresión de la densidad de X.
   c) Calcular $P[X \leq 4]$.
   d) Hallar $P[X \geq 6]$.
   e) Calcular $P[4 \leq X \leq 12]$.
   f) Hallar $P[X = 9]$.
2. En el estudio del sueño en los seres humanos se reconocen cinco fases (somnolencia, ligero, intermedio, profundo, REM) por medio del electroencefalograma. El sueño intermedio se caracteriza por la presencia de ondas de gran amplitud, que aparecen en un promedio de alrededor de dos ondas por segundo. ¿Cuál es la probabilidad de que, durante un sueño intermedio, no se presente ninguna de estas ondas durante un período de cinco segundos? ¿Cuál es la probabilidad de que aparezca un máximo de 15 de tales ondas en un período de cinco segundos? Si aparecieran 20 ó más de tales ondas durante un período de cinco segundos, ¿podría sospecharse que el sujeto no está en la etapa de sueño intermedio? Razonar sobre la base de la probabilidad implícita.
3. Una determinada planta nuclear desprende una cantidad detectable de gases radiactivos, un promedio de dos veces al mes. Hallar la probabilidad de que no se produzcan tales emisiones durante un período de tres meses. Hallar la probabilidad de que haya, como máximo, cuatro de tales emisiones durante ese período. ¿Cuál es el número esperado de emisiones durante un período de tres meses? Si han sido detectadas 12 ó más emisiones, ¿piensa usted que habría que dudar del promedio de dos al mes? Razonar sobre la base de la probabilidad implícita.
4. En una cierta población, se ha observado un número medio anual de muertes por cáncer de pulmón de 12. Si el número de muertes causadas por la enfermedad sigue una distribución de Poisson, ¿cuál es la probabilidad de que durante el año en curso:
   a) Haya exactamente diez muertes por cáncer de pulmón?
   b) Quince personas o más mueran a causa de la enfermedad?
   c) Diez personas o menos mueran a causa de la enfermedad?
5. En cierto cultivo, el número medio de células de *Rickettsia typhi* (células que causan el tifus) es de cinco por 20 micrómetros cuadrados (1/10 000 cm). ¿Cuántas células de ese tipo se espera encontrar en un cultivo de 16 micrómetros cuadrados? ¿Cuál es la probabilidad de que no se encuentre ninguna en un cultivo de 16 micrómetros cuadrados? ¿Cuál es la probabilidad de que al menos nueve de tales células se encuentren en un cultivo de este tamaño?
6. Se sospecha que muchas muestras de agua, todas del mismo tamaño y tomadas del Hillbank River, han sido contaminadas por operarios irresponsables de una planta de tratamiento de aguas. Se contó el número de microorganismos conformes en cada muestra. El número medio de microorganismos por muestra fue de 15. Suponiendo que el número de microorganismos se distribuye según una distribución de Poisson, calcular la probabilidad de que:
   a) La siguiente muestra contenga al menos 17 microorganismos.
   b) La siguiente muestra contenga 18 o menos microorganismos.
   c) La siguiente muestra contenga exactamente dos microorganismos.

7. Algunas especies de paramecios producen y secretan partículas «letales» que causarían la muerte de un individuo susceptible, si éste se pusiera en contacto con ellas. Todos los paramecios que no son capaces de producir partículas letales son susceptibles a ellas. E1 número medio de partículas letales emitidas por un paramecio de tales especies es de uno cada cinco horas. ¿Cuál es la probabilidad de que uno de estos paramecios no emita tales partículas en un período de dos horas y media? ¿Cuál es la probabilidad de que emita al menos una partícula letal?
8. El portavoz de una brigada de rescate afirma que ésta recibe un promedio de 3 llamadas por hora. Si la brigada recibiese un máximo de 6 llamadas, durante un intervalo de 5 horas, ¿podríamos sospechar que esta persona está exagerando el promedio de llamadas recibidas por hora? Explique el hecho basándose en la probabilidad de que esto ocurra.
9. Sea X binomial con $n = 20$ y $p = 0.1$. Utilice la tabla binomial para completar la segunda fila de la tabla mostrada en la Figura 4.4. Observe que, en este caso, no se sigue el método de aproximación de una distribución binomial por distribución de Poisson. Sin embargo, hallar la aproximación utilizando la tabla de Poisson con $\lambda s = 2$ para completar la fila 3 de la Figura 4.4. ¿Es ésta una buena aproximación?
10. En las moscas de la fruta, cuatro de cada 10^5 espermatozoides presentan una mutación del color rojo de los ojos a blanco, o viceversa. ¿Cuántas mutaciones esperaría usted que se produjesen en 200 000 espermatozoides? ¿Cuál es la probabilidad de que se produzca un máximo de 10? ¿Cuál es la probabilidad de que se produzcan entre 6 y 10, ambas inclusive?
11. En los seres humanos, se producen mutaciones por la enfermedad de Huntington en aproximadamente cinco de cada 10^6 gametos ¿Cuál es la probabilidad de que en dos millones de gametos haya al menos una mutación?
12. Se estima que sólo uno de cada 50 loros capturados en la cuenca del Amazonas, para su utilización como animales domésticos, sobrevive al cambio. Se capturan 700 pájaros en un día. ¿Cuál es el número esperado de supervivientes? ¿Cuál es la probabilidad de que sobreviva un máximo de 10 pájaros? Se capturan diariamente 700 pájaros durante un período de tres días, ¿cuál es la probabilidad de que en cada uno de los tres días sobreviva un máximo de 10 pájaros?
13. Dañando los cromosomas del óvulo o del espermatozoide pueden causarse mutaciones que conducen a abortos, defectos congénitos u otras deficiencias genéticas. La probabilidad de que tal mutación se produzca por radiación es de 0.10. De las 150 mutaciones próximas causadas por cromosomas dañados, ¿cuántas se espera que sean debidas a radiaciones? ¿Cuál es la probabilidad de que solamente 10 se deban a radiaciones?
14. La probabilidad de que un niño, aleatoriamente seleccionado, sea albino es de 1/20 000. En los próximos 40 000 niños nacidos, ¿cuál es la probabilidad de que ninguno sea albino? ¿Cuál es la probabilidad de que al menos uno sea albino?

x	0	1	2	3	4	5	6	7	8	9	10	▪■■
Distribución binomial acumulada	0.1216	0.3917	0.6769	?	?	?	?	?	?	?	?	?
Aproximación de Poisson	0.135	0.406	0.677	?	?	?	?	?	?	9	?	?

Figura 4.4. Comparación entre la distribución binomial con $n = 20$ y p = 0.1, y la distribución de Poisson con $\lambda s = np = 2$.

HERRAMIENTAS COMPUTACIONALES

TI83

La calculadora TI83 evalúa la densidad y la distribución acumulada para las distribuciones binomial y de Poisson. En ambos casos el usuario debe proporcionarle los valores numéricos de los parámetros que identifican a la distribución y el valor x de interés.

VII. Densidad binomial

Para mostrar el uso de la TI83 en el cálculo de la probabilidad binomial, volvamos al Ejemplo 4.4.5. En él, $n = 15$ y $p = 0.8$. Vamos a calcular la probabilidad de que $x = 12$. Recordemos que para hacerlo a mano hay que hallar el valor de la densidad binomial en $x = 12$. Así,

$$P[X = 12] = \frac{15!}{12!\ 3!}(0.8)^{12}\ (0.2)^3 = f(12)$$

La siguiente secuencia de pulsaciones evalúa la expresión anterior:

Tecla/Comando de la TI83	**Propósito**
1. 2^{ND} DISTR Ø	1. Muestra en pantalla la densidad binomial, binomial pdf (.
2. 15	2. Inserta el valor numérico de *n.*
3. 0.8 ,	3. Inserta el valor numérico de *p.*
4. 12) ENTER	4. Inserta el valor numérico de *x;* calcula /(12) = 0.2501388953.

VIII. Distribución acumulada binomial

Aquí vamos a calcular $P[10 \leqslant X \leqslant 14]$. En el Ejemplo 4.4.5, se hace restando F(9) a F(14). Con la calculadora, esta probabilidad acumulada es fácil de obtener.

Tecla/Comando de la TI83	**Propósito**
1. 2^{ND} DISTR ALPHA A	1. Muestra en pantalla la distribución binomial acumulada cdf (.
2. 15	2. Inserta el valor numérico de *n.*
3. 0.8	3. Inserta el valor numérico de *p.*
4. 14) ENTER	4. Inserta el primer valor de *x* deseado; calcula y presenta en pantalla F(14) = 0.9648156279.
5. CLEAR	5. Borra la pantalla.

6.	2^{ND} DISTR ALPHA A	6. Muestra nuevamente en pantalla la distribución binomial.
7.	15	7. Inserta el valor numérico de *n*.
8.	0.8	8. Inserta el valor numérico de *p*.
9.	9) ENTER	9. Inserta el segundo valor de *x* deseado; calcula y presenta en pantalla *F(9)* = 0.0610514296.

Para completar el cálculo, halle con la calculadora F(14) - F(9).

IX. Densidad de Poisson

La densidad de Poisson puede calcularse indicando los valores numéricos de λs y de *x,* del problema en cuestión. En el Ejemplo 4.5.1, $\lambda s = 6$. Calcularemos $P[X = 0]$ con la TI83.

Tecla/Comando de la TI83	**Propósito**
1. 2^{ND} DISTR ALPHA B	1. Muestra en pantalla la densidad de Poisson; Poisson pdf (.
2. 6 ,	2. Inserta el valor numérico de λs.
3. 0) ENTER	3. Inserta el valor numérico de *x;* calcula y muestra *f*(0) = 0.0024787522.

X. Distribución acumulada de Poisson

Vamos a calcular $P[X \leqslant 2]$. Si lo hacemos a mano, este valor viene dado por

$$F(2) = P[X \leqslant 2] = /(0) + /(1) + /(2)$$

El resultado, 0.062, ya fue obtenido en el Ejemplo 4.5.1.

Tecla/Comando de la TI83	**Propósito**
1. 2^{ND} DISTR ALPHA C	1. Muestra en pantalla la distribución acumulada de Poisson; Poisson cdf (.
2. 6	2. Inserta el valor numérico de λs.
3. 2) ENTER	3. Inserta el valor numérico de *x;* calcula y muestra F(2) = 0.0619688044.

Paquete estadístico SAS

VI. Distribución acumulada binomial

Se puede utilizar el SAS para generar las tablas acumuladas recogidas en un apéndice de este texto. Mostraremos la forma de generar las probabilidades acumuladas recogidas en la Tabla I, Apéndice B, para $n = 10$ y $p = 0.5$. Este programa puede modificarse para generar otras probabilidades.

Código SAS	Propósito
OPTIONS LS = 80 PS = 60 NODATE;	Establece las especificaciones de impresión.
DATA BINOMIAL;	Designa el conjunto de datos.
DO *X* = 0 a 10;	Genera la distribución acumulativa binomial para una variable aleatoria
P = PROBBNML (0.5, 10, X);	binomial con $p = 0.5$ y $n = 10$, mediante la función de SAS llamada PROBBNML.
CUMPROB = ROUND (P, 0.0001);	Redondea las probabilidades generadas a cuatro cifras decimales, como en la tabla del texto.
OUTPUT;	Produce la salida de los datos de *P* y de CUMPROB.
END;	Fin del bucle «DO».
PROC PRINT;	Imprime los resultados.

Los resultados de este programa se recogen a continuación. La variable OBS es una variable producida por SAS, que indica el número de la observación. La variable *X* es el valor de la variable aleatoria *X* cuya probabilidad acumulada se está calculando. *P* es la probabilidad acumulada calculada por la función PROBBNML, de SAS; CUMPROB es la probabilidad redondeada a cuatro cifras decimales de precisión.

El sistema SAS

OBS	X	P	CUMPROB
1	0	0.00098	0.0010
2	1	0.01074	0.0107
3	2	0.05469	0.0547
4	3	0.17188	0.1719
5	4	0.37695	0.3770
6	5	0.62305	0.6230
7	6	0.82812	0.8281
8	7	0.94531	0.9453
9	8	0.98926	0.9893
10	9	0.99902	0.9990
11	10	1.00000	1.0000

Variables aleatorias continuas

Recuerde que una variable aleatoria continua es una variable aleatoria que, de por sí, puede tomar *cualquier* valor en cierto intervalo o secuencia de números reales y *no* exclusivamente en puntos aislados. Vimos ya algunos ejemplos de variables aleatorias continuas en la Sección 4.1 y pudimos observar con anterioridad que la probabilidad de que la variable aleatoria tome un valor específico cualquiera es 0. Dado que esto no es cierto para el caso discreto, esta propiedad nos permite distinguir claramente una variable aleatoria continua de otra discreta. Por último, también pudimos darnos cuenta de que, en la práctica, las variables aleatorias continuas surgen en conexión con el tipo de medida de los datos.

En este capítulo, haremos un estudio comparativo en paralelo de los conceptos presentados en el Capítulo 4. En particular, definiremos la noción de densidad continua y explicaremos cómo debemos emplearla para calcular probabilidades. Representaremos geométricamente el valor esperado de X y consideraremos su función de distribución acumulativa también desde un punto de vista geométrico; además, presentaremos la familia de variables aleatorias normales, siendo ésta una de las más habituales dentro de las distribuciones continuas.

5.1 FUNCIONES DE DENSIDAD CONTINUA Y ESPERANZA

En el caso discreto, las funciones de densidad se representan frecuentemente mediante tablas. Estas tablas nos proporcionan una lista de los posibles valores de X junto con/(x) o probabilidad de que X tome el valor de x. El caso continuo es más complejo, debido a que, como una variable aleatoria continua puede tomar infinitos valores, resulta imposible enumerarlos todos. Además, no nos interesa encontrar una ecuación que, al ser aplicada a un determinado valor de x, nos dé la probabilidad de que X tome el valor de x, puesto que ya sabemos que su probabilidad es 0. Sin embargo, sí necesitamos disponer de una expresión que nos permita calcular probabilidades, ya que para el caso continuo, nos interesa conocer la probabilidad de que X esté comprendida en un intervalo de valores específico.

Por ejemplo, consideremos la variable aleatoria T, número de horas transcurridas entre la percepción de temblores de tierra y la próxima erupción del volcán Kilauea. Nos gustaría encontrar la probabilidad de que T sea menor de 24 horas $(P[T< 24])$. En el caso continuo, el

cálculo de las probabilidades puede hacerse geométricamente igualando probabilidades a áreas; la Definición 5.1.1 nos muestra cómo hacerlo.

Definición 5.1.1. Densidad continua. Sea X una variable aleatoria *continua.* La *densidad* de X es una función/definida en todos los números reales tal que

1. $f(x) \geqslant 0$ (no negativa).
2. El área comprendida entre la gráfica de f y el eje x es igual a 1.
3. Para cualquier valor real de los números a y b, $P_{\iota} \leqslant X \leqslant b]$ viene representada por el área comprendida entre la gráfica de f, las rectas $x = a$, $x = b$, y el eje x.

Esta definición parece maravillosa, pero en realidad no lo es, como veremos a continuación. Observe que existen similitudes entre las Definiciones 5.1.1 y 4.2.1. En ambos casos, las funciones implicadas son positivas, ambas «suman» 1 de algún modo, y pueden ser utilizadas en el cálculo de probabilidades. La Figura 5.1 nos muestra la representación gráfica de algunas densidades típicas.

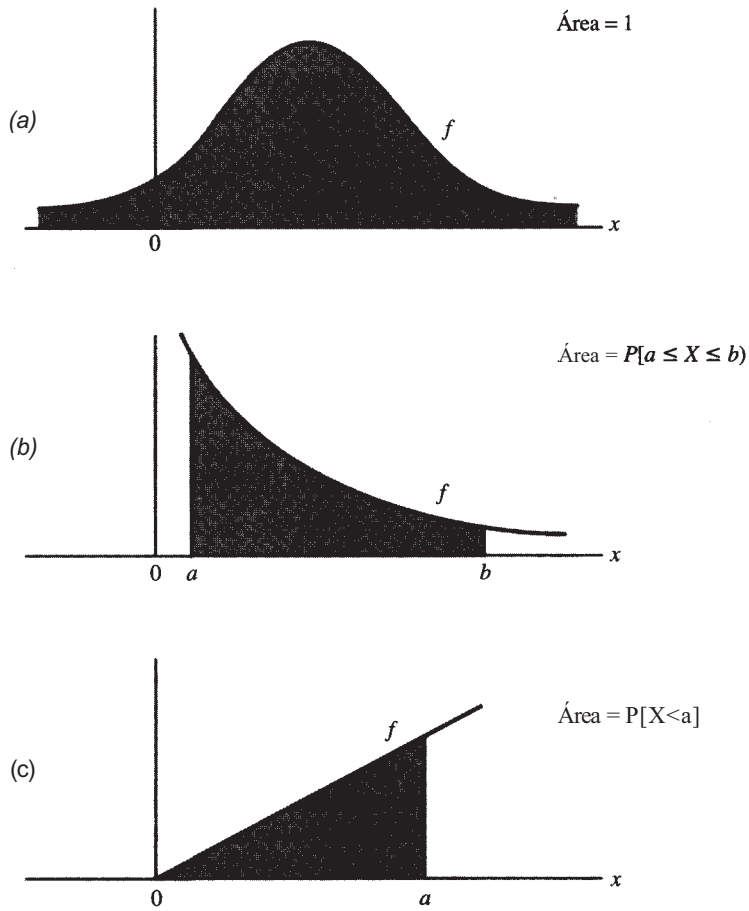


Figura 5.1. Algunas densidades continuas, (a) El área total limitada por f es 1. (b) El área que aparece bajo la gráfica de f entre los valores a y b representa la probabilidad de que X esté entre a y b. (c) El área que encontramos bajo la gráfica de f a la izquierda de a nos da la probabilidad de que X alcance como máximo el valor de a.

La primera tarea de un científico experimental es la de determinar la densidad adecuada de la variable aleatoria que esté considerando. Si esta variable aleatoria ha sido ampliamente estudiada en el pasado, se supone que su densidad puede haber sido ya desarrollada por otros y podremos utilizarla para contestar a las preguntas que se nos planteen. En cambio, si la variable es estudiada por primera vez, su densidad debe ser hallada a partir de datos experimentales. En el Capítulo 1 presentamos dos métodos para hacer esto: los diagramas de tallo y hojas y los histogramas. Por ejemplo, la representación gráfica de los datos en centímetros del perímetro craneal de un niño varón al nacer (Ejemplo 1.2.2), forma aproximadamente una campana. Esto implica que una densidad como la que mostramos en la Figura 5. la, podría utilizarse en el futuro para calcular las probabilidades relacionadas con esta variable. Los datos registrados durante el terremoto del Ejemplo 1.2.1 sugieren una densidad que tomará la forma de la Figura *5.1b.* En este capítulo nos centraremos en la forma de utilizar una densidad para el cálculo de probabilidades, una vez que conozcamos ya su forma.

Ejemplo 5.1.1. La experiencia pasada acerca de las erupciones volcánicas nos indica que *T,* tiempo transcurrido entre los temblores de tierra y la erupción del volcán, tiene una densidad con forma de campana y valor centrado en 36 horas. La densidad está representada en la Figura *5.2a.* Sabemos que el área sombreada toma el valor 1. Para encontrar la probabilidad de que T tenga un valor menor de 24 horas, $P[T < 24]$, debemos hallar el área de la región sombreada representada en la Figura *5.2b.* Tengamos en cuenta que, por definición, $P[T = 24] = 0$. Esta afirmación puede comprobarse en su representación geométrica: puesto que el área en cuestión es la delimitada por la línea de puntos en $T = 24$ de la Figura 5.2c y las líneas tienen longitud pero no anchura, su área es 0; además, dado que toda probabilidad corresponde a área determinada, podremos decir también que $P[T= 24]$ es 0. Una consecuencia de ello es que $P[T < 24]$ es idéntica a $P[T < 24]$. En el caso *continuo,* el añadir o eliminar un valor extremo al intervalo, no supondrá ninguna diferencia en la probabilidad final, pues acabamos de ver que dicho valor extremo en sí mismo tiene una probabilidad 0. Esto no se cumple en el caso discreto en el cuál los valores extremos tienen probabilidades positivas.

Áreas como la que tenemos representada en la Figura *5.2b* no resultan fáciles de encontrar, ya que para hacerlo tendremos que conocer previamente tanto la expresión exacta de la curva como las técnicas de cálculo empleadas. Así, aunque en este momento podamos representar $P[T < 24]$, no podemos calcular su valor numérico exacto.

Esperanza

Recuerde que el valor esperado de X corresponde a su promedio teórico. Este valor viene dado por $E[X]$ o μ. En el caso discreto vimos que μ podía calcularse a partir de la densidad mediante la Definición 4.2.3.

Excepto en los casos más fáciles, no podemos hallar el valor esperado de una variable aleatoria continua sin cálculos, aunque es posible visualizar $E[X]$ geométricamente. Tomemos la gráfica de la densidad/de X. Si hacemos un corte imaginario de la región delimitada por la gráfica de f y el eje x, como si se tratara de una pieza delgada y rígida de metal que intentara poner en equilibrio esta región sobre el filo de una cuchilla paralela al eje vertical de la gráfica, el punto en el cual dicha región se balancearía es $E[X]$. En la Figura 5.3 mostramos ejemplos típicos de estos «puntos de equilibrio». Obsérvese que en la Figura 5.2 la curva simétrica de la campana está centrada en 36 horas. Dada la simetría de la curva, 36 es el punto de equilibrio, y por ello podemos concluir que la media de tiempo transcurrido entre los temblores registrados y la erupción del volcán es de 36 horas.

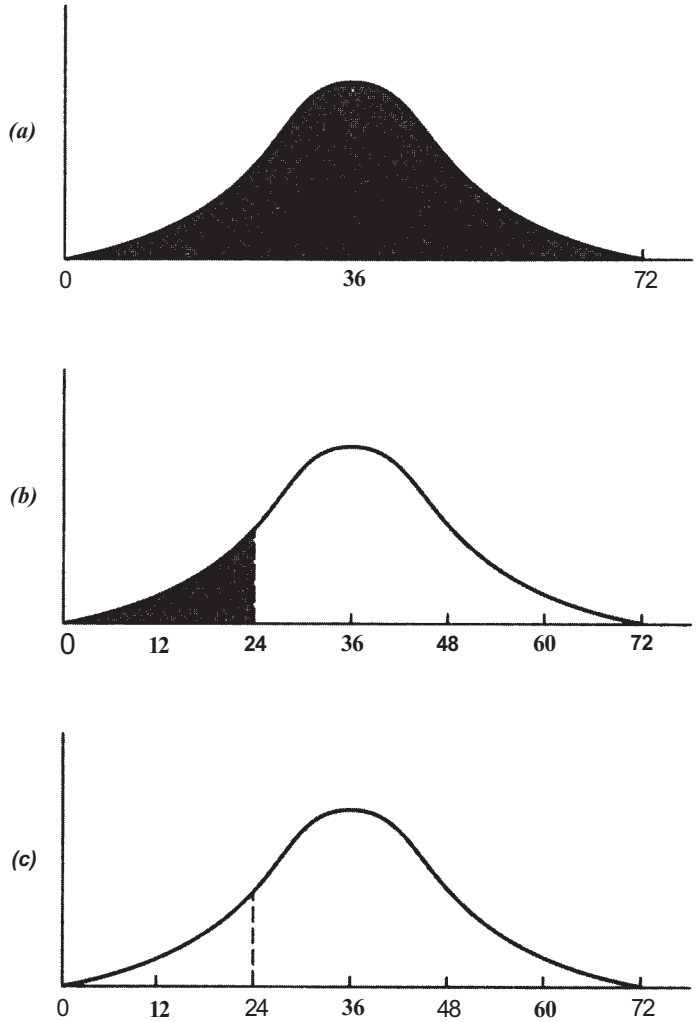


Figura 5.2. (a) Tadopta una distribución en forma de campana centrada en el valor 36. *(b)* P[*T< 24]* viene dada por el área sombreada, *(c) P[T*= 24] = 0, dado que la recta en *T*= 24 no tiene área.

EJERCICIOS 5.1

1. Supongamos que la densidad de la variable aleatoria Z, o número de centímetros cúbicos de un fármaco que han de prescribirse para el control de ataques epilépticos, es como la representada en la Figura *5.4a.*
   a) Compruebe que para que el área comprendida bajo la gráfica de *f* sea 1, *h,* la altura del triángulo, debe ser $\frac{20}{3}$. *Sugerencia:* La fórmula del área de un triángulo es $A = \frac{1}{2}$ (base)(altura).
   b) Sombree en la Figura 5.4a el área correspondiente a la probabilidad de que por lo menos 0.1 cc deban ser prescritos para controlar los ataques epilépticos.
   c) ¿Cuál es la probabilidad representada por el área sombreada de la Figura 5.4b?
   d) ¿Cuál es la probabilidad representada por el área sombreada de la Figura 5.4c?
   e*)* ¿Cuál es la probabilidad representada por el área sombreada de la Figura *5.4d?*

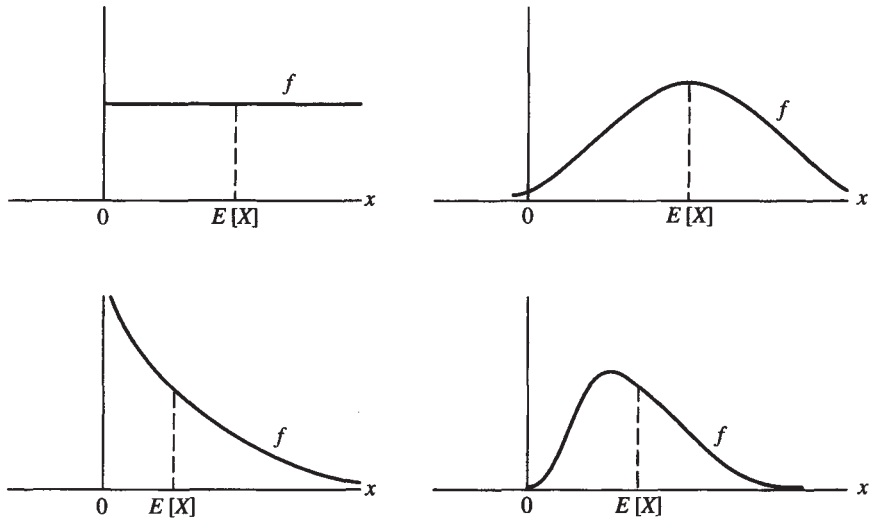


Figura 5.3. $\mu = E[X]$ = punto de equilibrio.

f) Si conociera las áreas de los apartados *c y d,* ¿cómo podría encontrar el área del apartado *el*

g) Calcule aproximadamente la dosis media requerida estimando mediante observación el punto de equilibrio.

h) La ecuación de la densidad de Z es

$$f(z) = 200z/9 \qquad 0 \leqslant z \leqslant 0.3$$

Utilice esta información para calcular el valor de/(z) cuando $z = 0.2$ y encuentre el área de la región sombreada en la Figura *5.4b* para así calcular la probabilidad de que deban prescribirse al menos 0.2 cm^3 del fármaco para controlar los ataques.

i) Utilice el método del apartado *h* para calcular la probabilidad de que deban prescribirse al menos 0.1 cm^3 de fármaco.

j) Utilice la información de los apartados *h* e *i* para calcular la probabilidad de que deban prescribirse entre 0.1 y 0.2 cm^3 de fármaco.

2. Sea *X* el porcentaje de líquido corporal perdido durante las primeras 24 horas por una persona que ha sufrido una quemadura grave. Suponiendo que *X* tiene la densidad mostrada en la Figura 5.5:
   a) ¿Cuál es la probabilidad representada por el área sombreada de la Figura 5.5?
   b) ¿Cuál es la probabilidad de que $X = 15\,\%$?
   c) Marque el área correspondiente a $P[X \geqslant 20\,\%]$.
   d) ¿Cuál es el porcentaje promedio perdido en esta situación?
3. Sea *X* el tiempo de supervivencia en años después de un diagnóstico de leucemia. La Figura 5.6 muestra la densidad de *X.*
   α) Sombree la región correspondiente a la probabilidad de que el paciente sobreviva menos de 6 meses.
   b) Si el área del apartado α tiene valor $\frac{7}{16}$, ¿cuál es la probabilidad de que un paciente sobreviva por lo menos 6 meses?
   c) ¿Cuál es la probabilidad de que un paciente sobreviva exactamente 6 meses?

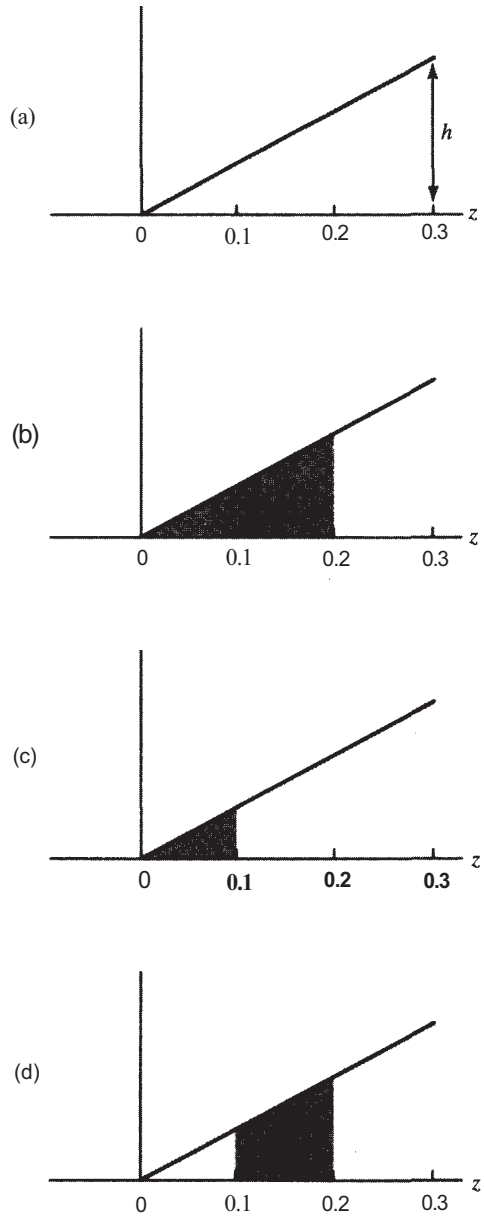


Figura 5.4. La densidad para Z, entendida como la cantidad en cm^3 de un fármaco utilizado para el control de ataques epilépticos.

4. La comunicación por vía química es una práctica corriente entre los animales. Gran parte de la comunicación entre los insectos se hace mediante hormonas liberadas al exterior, llamadas feromonas. Dichas hormonas las emplean las hormigas para marcar su paso, de modo que, arrastrando su abdomen por el suelo, pueden ir desde su nido hasta una fuente de alimento y regresar, dejando tras ellas una pista química. Cuando ya no queda alimento, las hormigas que utilizan la pista dejan de producir marcadores y ésta desaparece. La Figura 5.7 representa la densidad de X entendida como el tiempo en minutos que la pista de feromonas persiste después de la última secreción hormonal.

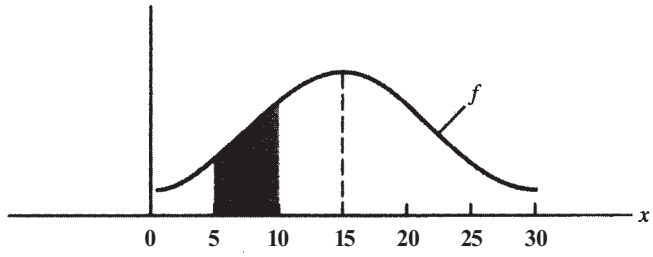


Figura 5.5. La densidad de *X,* entendida como el porcentaje de líquido corporal perdido durante las primeras 24 horas por una persona que ha sufrido una quemadura grave, es simétrica.

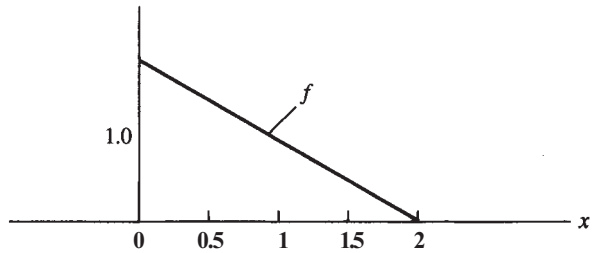


Figura 5.6. La densidad de *X, o* tiempo de supervivencia (en años) diagnosticado a pacientes con leucemia aguda.

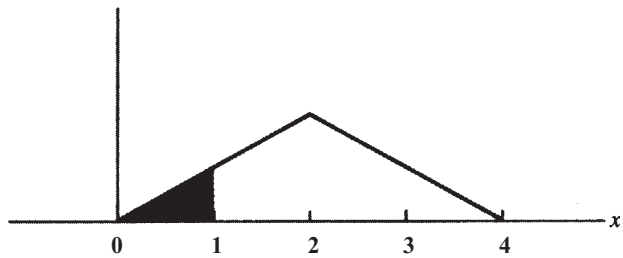


Figura 5.7. La densidad de *X,* entendida como el tiempo en minutos durante el cual persiste una pista de feromonas, es simétrica.

a) ¿Cuál es la probabilidad representada por el área sombreada?
b) Si la probabilidad del apartado *a* es $\frac{1}{8}$, ¿cuál es la probabilidad de que la pista se mantenga entre 1 y 2 minutos?
c) ¿Cuál es la probabilidad de que la pista persista durante más de 3 minutos?
d) ¿Cuál es la probabilidad de que la pista permanezca durante 2 minutos exactos?
e) ¿Cuál es el promedio de tiempo durante el cual persiste la pista?

5. *Distribución uniforme.* Una variable aleatoria cuya densidad es plana se dice que está *uniformemente* distribuida. La ecuación de la curva es $f(x) = c$, donde c es una constante. Estas densidades son de fácil manejo debido a que las áreas implicadas son siempre rectangulares. La fórmula del área de un rectángulo es *área = base* x *altura.* Supongamos que la variable aleatoria *X,* tiempo en minutos que tarda una enfermera en responder a la llamada de un paciente, está uniformemente distribuida en el intervalo de 0 a 5 minutos. La Figura 5.8 representa esta densidad.

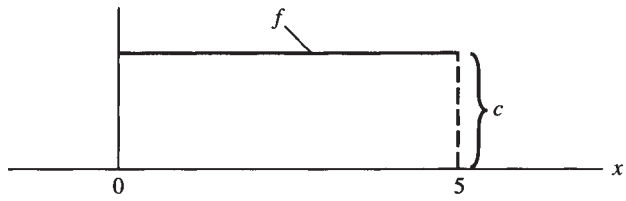


Figura 5.8. La variable aleatoria *X,* tiempo que necesita una enfermera para acudir a la llamada de un paciente, es una distribución uniforme a lo largo de un intervalo de 5 minutos.

a) Compruebe que en este caso $c = \frac{1}{5}$.
b) Sombree el área que representa la probabilidad de registrar una respuesta que exceda los 3 minutos.
c) Calcule la probabilidad representada en el apartado *b.*
d) ¿Cuál es la media en el tiempo de respuesta?

5.2. FUNCIÓN DE DISTRIBUCIÓN ACUMULADA

La función de distribución acumulada, representada por *F,* se definió en la Sección 4.3. Recuerde que esta función nos proporciona la probabilidad de que una variable aleatoria adopte un valor menor o igual al valor especificado; es decir, $F(x) = P[X \leqslant x]$. En el caso discreto, esta función se evalúa en un punto concreto x_0 sumando la densidad de todos los posibles valores de X menores o iguales a x_0; en el caso continuo, podemos hallar $F(x_0)$ calculando el área delimitada por la gráfica de densidad, a *la izquierda del punto* x_0 *e incluyendo éste.* Dado que esto implica cálculos bastante complejos, excepto en los casos más sencillos, se han creado tablas de probabilidades acumuladas para las variables aleatorias utilizadas con mayor frecuencia. Su trabajo, de aquí en adelante, será aprender a manejar estas tablas y, para conseguirlo, debe ser capaz de representar las probabilidades deseadas y expresarlas en función de *F.* Un ejemplo ilustrará esta idea.

Ejemplo 5.2.1. La variable aleatoria *X,* entendida como el límite forestal o altitud montañosa máxima (en metros) en la cual es posible el crecimiento de árboles, está influenciada por la temperatura, las condiciones del suelo y la lluvia. Supongamos que la distribución de la variable *X* adopta la forma de campana de la Figura *5.9a.* Vamos a suponer que queremos hallar la probabilidad de que el límite forestal de una montaña, seleccionada al azar, tome un valor entre los 3000 y los 3375 metros inclusive. El área que corresponde a esta probabilidad está representada en la Figura *5.9b* y podemos calcularla realizando una resta: primero hallamos el área a la izquierda de los 3375 metros, incluyendo dicha distancia (Fig. 5.9c) y después calculamos el área a la izquierda de 3000 (Fig. *5.9d),* con lo que podremos obtener el área deseada restando a la calculada en primer lugar la segunda. Podemos expresar estos pasos en función de *F* como veremos a continuación:

$$\begin{aligned} P[3000 \leqslant X \leqslant 3375] &= P[X \leqslant 3375] - P[X < 3000] \\ &= F(3375) - P[X < 3000] \end{aligned}$$

Recuerde que en el caso continuo $P[X = 3000] = 0$, por lo que $P[X \leqslant 3000] = P[X < 3000]$. Haciendo las sustituciones correspondientes tendremos que:

$$\begin{aligned} P[3000 \leqslant X \leqslant 3375] &= F(3375) - P[X \leqslant 3000] \\ &= F(3375) - F(3000) \end{aligned}$$

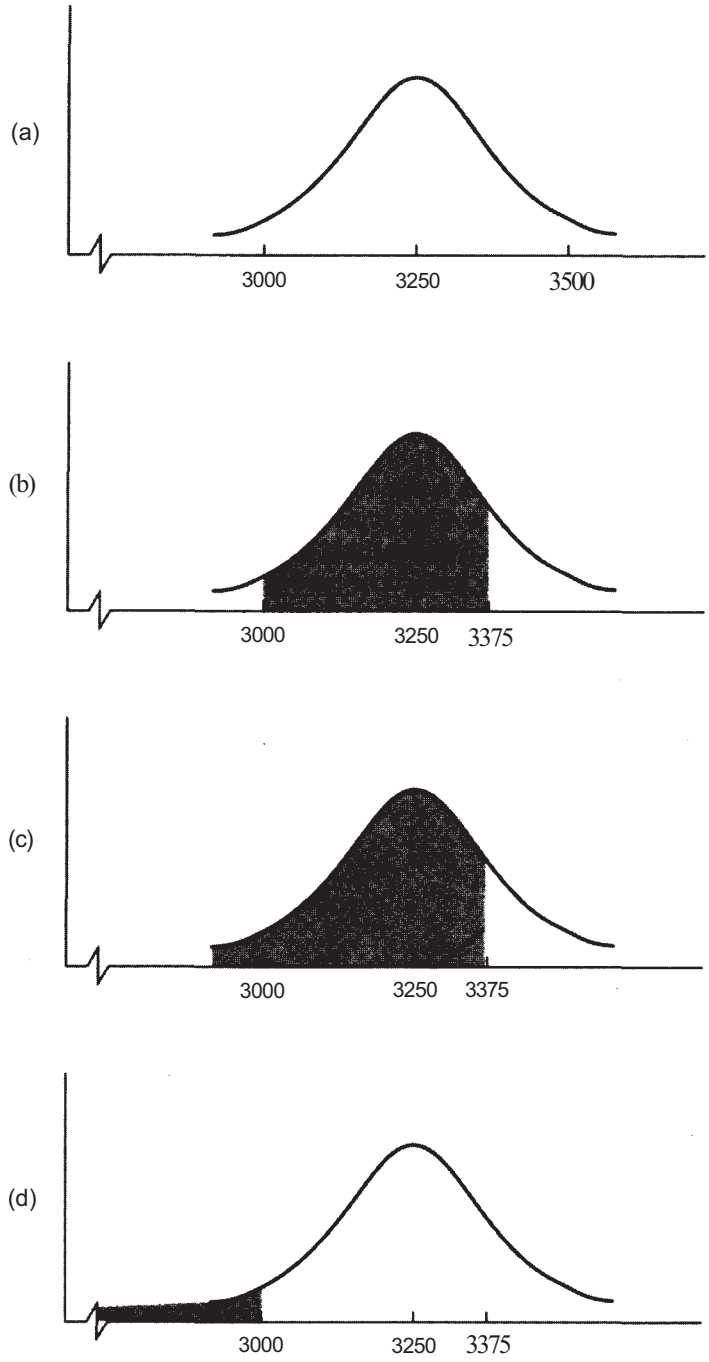


Figura 5.9. (a) Densidad supuesta de *X,* límite forestal en metros; (*b)* área = P[3000 ⩽ *X* ⩽ 3375]; *(c)* área = *P[X* ⩽ 3375] = P(3375); (d) área = *P[X*< 3000] = P[X ⩽ 3000] = F(3000).

Si dispusiéramos de una tabla de probabilidades acumuladas, podríamos obtener la probabilidad deseada localizando en la tabla los valores 3375 y 3000 y restando las probabilidades asociadas a ellos. (Basado en la información encontrada en William Keeton y Carol Hardy McFadden, *Elements of Biological Science,* 3.ª ed., W. W. Norton, New York, 1984.)

El punto más importante de esta sección es el hecho de que siempre que manejemos una variable aleatoria continua, F estará definida por el área a la izquierda del punto estudiado. Todas las tablas acumuladas de este texto se han construido de modo que estas áreas puntuales puedan ser obtenidas directamente, teniendo en cuenta que las demás deberán calcularse mediante operaciones de sustracción.

EJERCICIOS 5.2

1. Sea X la variable aleatoria número de años de funcionamiento de un marcapasos hasta que empieza a fallar. La Figura 5.10 representa la densidad de X.
   a) ¿Qué región(es) representa *F(4)?*
   b) ¿Qué probabilidad representan las regiones II y III juntas? Exprese esta probabilidad en función de F.
   c) ¿Qué probabilidad representa la región V? Exprese esta probabilidad en función de *F,*
   d) Exprese $P[X \leqslant 4]$ y $P[X < 4]$ en función de F.
2. La Figura 5.11 muestra la gráfica de densidad de una variable aleatoria X entendida como el tiempo en minutos que debe transcurrir para que un sedante haga efecto.
   a) ¿Qué región(es) del diagrama corresponde(n) a *F{2)1*
   b) ¿Qué región(es) del diagrama corresponde(n) a F(6)?
   c) Exprese la región III en función de F.
   d) Exprese la región IV en función de F.
3. Sea la variable X entendida como el tiempo de eficacia en meses para un electrodo de pH, Su densidad viene representada en la Figura 5.12.

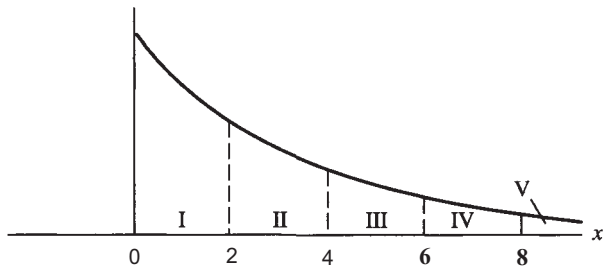


Figura 5.10. Densidad hipotética de X, tiempo transcurrido en años hasta que un marcapasos empieza a fallar.

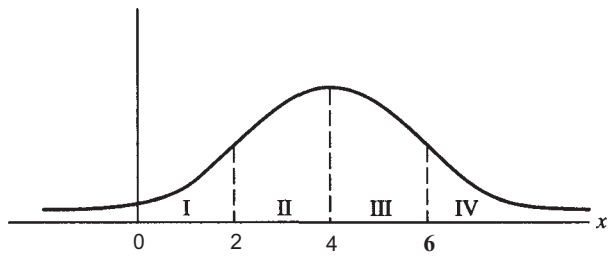


Figura 5.11. Densidad hipotética de X, tiempo que tarda un sedante en hacer efecto.

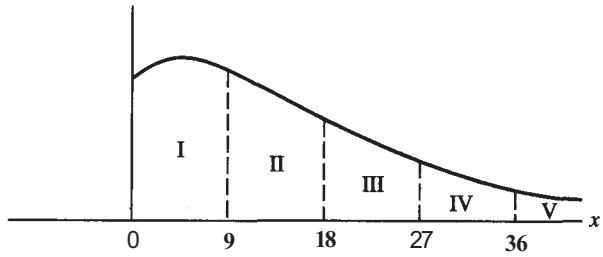


Figura 5.12. Densidad hipotética de *X,* tiempo en meses que puede durar un electrodo de pH.

a) ¿Qué regiones de la gráfica corresponden a F(27)?
b) Exprese en función de F la probabilidad de que un electrodo de pH elegido aleatoriamente funcione con eficacia durante al menos 18 meses. ¿Qué regiones corresponden a esta probabilidad?
c) Exprese en función de *F* la probabilidad de que un electrodo de pH elegido aleatoriamente funcione correctamente durante un período de 27 a 36 meses. ¿Qué regiones corresponden a esta probabilidad?

4. Cuando el refrigerante caliente de una central nuclear es vertido repentinamente en una corriente de agua, puede producirse un cambio súbito en la temperatura de ésta. Como consecuencia, a menudo se produce la muerte de los organismos que viven en ellas. Sea *X* la temperatura del río en °C a $\frac{1}{4}$ de milla de distancia corriente abajo de la central nuclear justo antes de que la sustancia refrigerante sea liberada en la corriente. Sea *Y* la temperatura registrada en el mismo punto del río 5 minutos después de que el refrigerante sea liberado. Sea $D = Y - X$ el cambio de temperatura en el río atribuido al vertido del refrigerante. La Figura 5.13 representa la gráfica de densidad *f* para *D.*
   a) ¿Qué región(es) representa(n) F(12)?
   b) ¿Qué probabilidad representa la combinación de las áreas I y II? Dado que / es simétrica en torno al valor 15, ¿cuál es el valor numérico representado por el área de las regiones I y II?
   c) ¿Qué probabilidad representa la región III?
   d) ¿Qué regiones representa F(17) - F(12)? ¿Qué probabilidad corresponde a F(17) - F(12)?
   e) ¿Qué es F(5)?
   f) ¿Cuál es la probabilidad de que el cambio de temperatura sea como máximo de 25 °C?

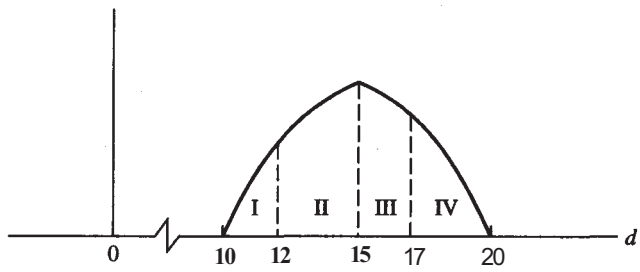


Figura 5.13. Densidad hipotética de *D,* variación de temperatura causada en un río por el vertido de un refrigerante procedente de una central nuclear.

5.3. DISTRIBUCIÓN NORMAL

La familia normal es una familia de variables aleatorias continuas. Esta distribución fue descrita por primera vez, en 1773, por Abraham De Moivre como el valor límite de la densidad binomial cuando el número de ensayos es infinito. Este descubrimiento no llamó mucho la atención y la distribución fue «redescubierta» de nuevo por Pierre-Simon Laplace y Carl Friedrich Gauss medio siglo después. Ambos se dedicaban a resolver problemas de astronomía y cada uno de ellos obtuvo la distribución normal como la que aparentemente describía el comportamiento de los errores en las medidas astronómicas.

La distribución normal es de gran importancia en el análisis y cálculo de todos los aspectos relacionados con datos experimentales en ciencias y en medicina. De hecho, la mayoría de los métodos estadísticos básicos que estudiaremos en los próximos capítulos se apoyan en la distribución normal.

Recuerde que en el caso discreto existen muchas distribuciones binomiales distintas. Cada una de ellas adopta una densidad con la siguiente forma:

$$f(x) = \frac{n!}{x!\ (n-x)!}\, p^x\, (1-p)^{n-x} \qquad \text{donde} \qquad x = 0,\ 1,2,\ ...,\ n$$

Así, para distinguir una determinada distribución binomial sólo necesitamos hallar los valores numéricos de *n y p.* En el caso de las distribuciones normales, nos encontramos con la misma situación ya que también existen muchas distribuciones diferentes. En ellas cada densidad tiene la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-1/2[(x-\mu)/\sigma]^2} \qquad (x \text{ es real})$$

donde σ es la desviación típica de la variable aleatoria y μ es su media. Para identificar una determinada variable aleatoria distribuida normalmente sólo necesitamos hallar los valores de μ *y* de σ. La ecuación de densidad descrita arriba no es sencilla aunque como trabajaremos más con las tablas de probabilidad que con ella misma, su complejidad no tiene demasiada importancia.

Podemos utilizar técnicas elementales de cálculo a fin de comprobar las propiedades que veremos a continuación.

Propiedades de la curva normal

1. La gráfica de densidad de cualquier variable aleatoria normal es una curva simétrica en forma de campana con centro en su media μ (véase Fig. 5.14). Obsérvese que, como se mencionó en el Capítulo 1, μ es un parámetro de situación en el sentido de que indica dónde está centrada o localizada la curva a lo largo del eje horizontal.
2. Los puntos de inflexión o «depresiones» en la curva se dan para valores de *X* iguales a una desviación típica a cada lado de la media ($x = \mu \pm \sigma$). La situación de estos puntos determina la forma de la curva. Cuanto mayor sea el valor de σ, más lejos de la media estarán los puntos de inflexión y más plana será la curva, de aquí que, como se indicó en el Capítulo 1, σ sea un parámetro de forma (véase Fig. 5.15).

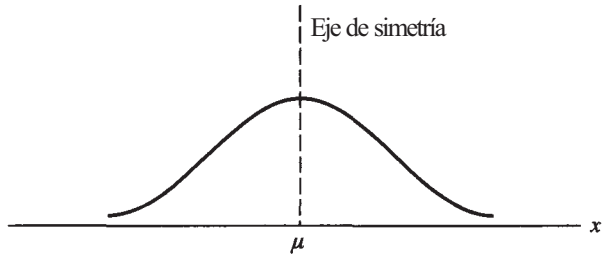


Figura 5.14. La campana queda centrada en μ, o valor medio de la variable aleatoria.

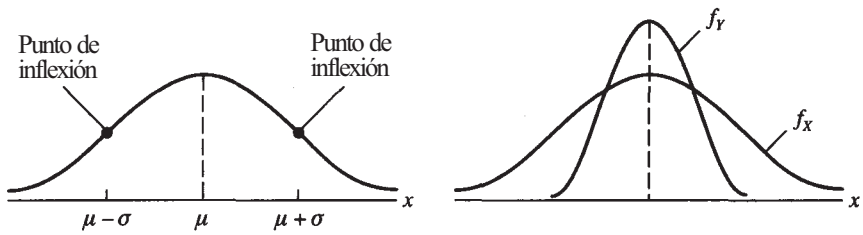


Figura 5.15. σ determina los puntos de inflexión (izquierda); $\mu_x = \mu_y$ y $\sigma_X > \sigma_Y$ (derecha).

3. Toda variable aleatoria normal es continua, por lo que pueden aplicársele todas las propiedades generales de las variables continuas desarrolladas en la Sección 5.1. En concreto, para cualquier densidad normal, f, $f(x) \geq 0$ y el área limitada por la gráfica de y el eje horizontal es 1. Las probabilidades se calculan encontrando las áreas correspondientes.

Ejemplo 5.3.1. Una de las mayores contribuciones a la contaminación atmosférica es la provocada por los hidrocarburos procedentes de los tubos de escape de los automóviles. Sea X los gramos de hidrocarburos emitidos por un automóvil por cada milla recorrida. Supongamos que X es una variable distribuida normalmente y que tiene una media de 1 g y una desviación típica de 0.25 g. La densidad de X viene dada por la siguiente ecuación:

$$f(x) = \frac{1}{0.25\sqrt{2\pi}} \cdot e^{-1/2[(x-1)/0.25]^2}$$

Su gráfica de densidad es simétrica y tiene forma de campana centrada en $\mu = 1$ con puntos de inflexión en $\mu \pm \sigma$ o $1 \pm .25$. La Figura 5.16 muestra la representación simplificada de esta densidad.

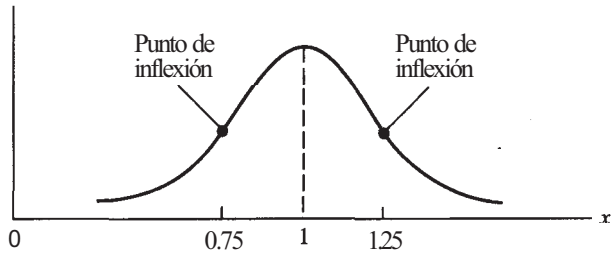


Figura 5.16. Gráfica de la densidad de X número de gramos de hidrocarburos emitidos por un automóvil en cada milla recorrida. Tenemos puntos de inflexión en $\mu \pm \sigma$ ó 1 ± 0.25.

Debemos tener en cuenta que, teóricamente hablando, una variable aleatoria normal debe poder tomar cualquier valor, lo que es claramente irreal en este caso puesto que es imposible que un automóvil produzca una cantidad negativa de hidrocarburos. Así, cuando decimos que X tiene una distribución normal, significa que a lo largo del espectro de posibles valores que puede tomar X, la curva normal nos proporciona probabilidades aceptables. Desde este punto de vista, podremos al menos obtener la probabilidad aproximada de que un automóvil escogido al azar emita entre 0.9 y 1.5 g de hidrocarburos, calculando el área bajo la gráfica de f comprendida entre el eje horizontal y las rectas $x = 0.9$ y $x = 1.5$, como muestra la Figura 5.17.

Distribución normal tipificada

Hay un número infinito de variables aleatorias normales, cada una de las cuales se caracteriza de forma exclusiva por los parámetros μ y σ^2. Para calcular las probabilidades asociadas a una curva normal específica, hay que recurrir directamente al cálculo; más concretamente, para resolver este problema se utiliza una simple transformación algebraica, conocida como *procedimiento de tipificación,* mediante la cual se puede transformar cualquier cuestión relativa a una variable aleatoria normal en otra equivalente pero referida a una variable aleatoria normal de media 0 y varianza 1. Esta variable normal particular se representa con la letra Z y se conoce como variable aleatoria *normal tipificada,* cuya función de distribución acumulada viene dada en la Tabla III del Apéndice B que proporciona $P[Z < z]$ para valores determinados de z. El Ejemplo 5.3.2. explica el uso de la Tabla III.

Ejemplo 5.3.2

a) Hallar $P[Z \leqslant 1.56] = F(1.56)$. Gráficamente, lo que buscamos es el área que aparece en la Figura 5.18a. La Tabla III del Apéndice B nos da directamente los valores de F; así, F(1.56) se calculará situando los dos primeros dígitos (1.5) en la columna encabezada por z y puesto que el tercer dígito es 6, la probabilidad deseada será la que muestre la tabla al hacer coincidir la fila rotulada 1.5 con la columna rotulada 0.06. En este caso, dicha probabilidad es de 0.9406.

b) Encontrar $P[Z \geqslant -1.29]$. El área que nos interesa aparece en la Figura 5.18b y la probabilidad se calculará por sustracción ya que:

$$\begin{aligned} P[Z \geqslant -1.29] &= 1 - P[Z < -1.29] \\ &= 1 - P[Z \leqslant -1.29] \quad \text{(ya que Z es continua)} \\ &= 1 - F(-1.29) \end{aligned}$$

La Tabla III indica que $F(-1.29) = 0.0985$, valor obtenido al hacer coincidir la fila reseñada con -1.2 con la columna reseñada por 0.09. La probabilidad deseada es, por tanto, $1 - 0.0985 = 0.9015$.

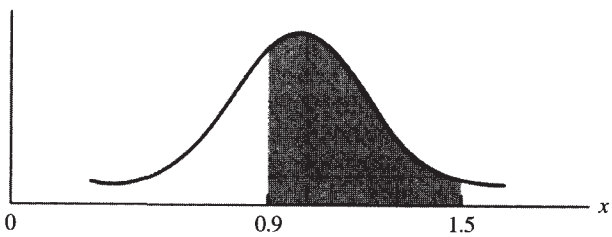


Figura 5.17. Área sombreada = $P[0.9 \leqslant X \leqslant 1.5]$.

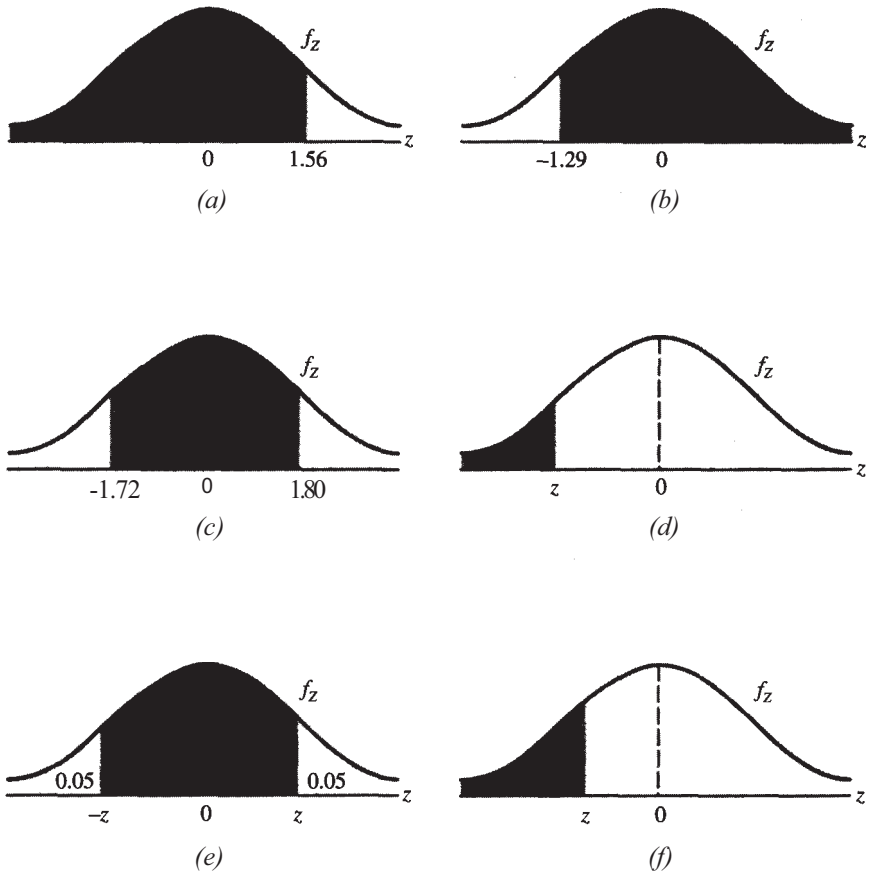


Figura 5.18. *(a)* $P[Z \leqslant 1.56]$; *ib)* $P[Z \geqslant -1.29]$; *(c)* $P[-1.72 \leqslant X \leqslant 1.80]$; *(d)* $P[Z \leqslant z] = 0.025$; *(e)* $P[-z \leqslant Z \leqslant z] = 0.90$; *(f)* $P[Z \leqslant z] = 0.10$.

c) Hallar P[-1.72 ⩽ Z ⩽ 1.80]. Esta probabilidad aparece en la Figura 5.18c. En términos de la función de distribución acumulada:

$$\begin{aligned} P[-1.72 \leqslant Z \leqslant 1.80] &= P[Z \leqslant 1.80] - P[Z < -1.72] \\ &= P[Z \leqslant 1.80] - P[Z \leqslant -1.72] \quad \text{(Z es continua)} \\ &= F(1.80) - F(-1.72) \\ &= 0.9641 - 0.0427 = 0.9214 \end{aligned}$$

d) Hallar el punto z tal que $P[Z \leqslant z] = 0.025$. La cuestión ahora es distinta a las propuestas con anterioridad. Las anteriores implicaban encontrar la probabilidad asociada a un punto dado, mientras que esta última implica encontrar el punto asociado a una probabilidad determinada. Así, en esta ocasión, nos piden hallar el punto z que aparece en la Figura 5.18d, y para hacerlo hay que leer la Tabla III a la inversa. Es decir, vamos al cuerpo de la Tabla y localizamos la probabilidad 0.025 y vemos que se encuentra en el punto de convergencia de la fila reseñada por -1.9 con la columna reseñada por 0.06. De este modo, el punto buscado es z = -1.96.

e) Hallar el punto z tal que $P[-z \leqslant Z \leqslant z] = 0.90$. Este punto se indica en la Figura 5.18e. Obsérvese que el punto z tiene la propiedad de que el área a la izquierda de z es 0.90 + 0.05 = 0.95. Así que z es un punto tal que $P[Z \leqslant z] = 0.95$. Para encontrar z,

trataremos de localizar la probabilidad 0.95 en el cuerpo de la Tabla III. Este valor exacto no aparece; sin embargo, hay dos valores, el 0.9495 que se corresponde con una $z = 1.64$ y el 0.9505 correspondiente a una $z = 1.65$, equidistantes al valor buscado 0.9500. De este modo, nos encontramos ante la siguiente situación:

Área a la izquierda	Punto
0.9495	1.64
0.9500	?
0.9505	1.65

Debido a que la probabilidad deseada es el punto medio entre las dos probabilidades extraídas de la tabla, estimaremos el punto z como el promedio entre 1.64 y 1.65; tendremos así que z es 1.645.

f) Calcular el punto z (mostrado en la Fig. 5.18/) tal que $P[Z \leqslant z] - 0.10$. Cuando buscamos 0.1000 en el cuerpo de la Tabla III vemos que no aparece en ella. Ni tampoco es el punto medio entre dos áreas como en el caso del apartado *e.* En esta situación, estimaremos z localizando el área más próxima a 0.1000 que, en este caso, es 0.1003. Así. el valor de z asociado con esta área es $z = -1.28$.

Tipificación

Para utilizar la tabla de la normal tipificada y contestar a todas las preguntas relacionadas con la variable aleatoria normal X, tenemos que volver a plantear estas cuestiones en función de Z. Este proceso recibe el nombre de *tipificación* y se lleva a cabo restando a cualquier valor de X su media y dividiendo esta diferencia por su desviación estándar (o típica). Formalmente, esta idea queda definida en el Teorema 5.3.1., mientras que el Ejemplo 5.3.3 presenta el modo en que se utiliza.

Teorema 5.3.1. Teorema de tipificación. Sea X una variable normal con media μ y varianza σ^2. La variable $(X - \mu)/\sigma$ es normal tipificada (estándar).

Ejemplo 5.3.3. El plomo, como muchos otros elementos, está presente en el medio natural. La revolución industrial y la llegada del automóvil han incrementado la cantidad de plomo en el medio hasta el punto de que, en algunos individuos, la concentración de plomo puede alcanzar niveles peligrosos. Sea X la concentración de plomo en partes por millón en la corriente sanguínea de un individuo. Supongamos que X es una variable normal con media 0.25 y desviación típica 0.11. Una concentración superior o igual a 0.6 partes por millón se considera extremadamente alta. ¿Cuál es la probabilidad de que un individuo seleccionado aleatoriamente esté incluido en esta categoría?

Para responder a esta pregunta debemos hallar $P[X \geqslant 0.6]$ lo que puede hacerse tipificando X, es decir, restando la media 0.25 y dividiendo por la desviación típica 0.11, en ambos miembros de la desigualdad. De este modo,

$$\begin{aligned} P[X \geqslant 0.6] &= P\left[\frac{X - 0.25}{0.11} \geqslant \frac{0.6 - 0.25}{0.11}\right] \\ &= P[Z \geqslant 3.18] \\ &= 1 - P[Z \leqslant 3.18] \\ &= 1 - 0.9993 = 0.0007 \end{aligned}$$

Concentraciones entre 0.4 y 0.6 partes por millón representan exposiciones al plomo debidas al desempeño de ciertas profesiones. La probabilidad de que un individuo seleccionado aleatoriamente se encuentre en este rango es:

$$P[0.4 \leqslant X \leqslant 0.6] = P\left[\frac{0.4 - 0.25}{0.11} \leqslant Z \leqslant \frac{0.6 - 0.25}{0.11}\right]$$
$$= P[1.36 \leqslant Z \leqslant 3.18]$$
$$= 0.9993 - 0.9131 = 0.0862$$

En el Ejemplo 5.3.3 dábamos un valor, 0.6, y pedíamos calcular una probabilidad. En el ejemplo que veremos a continuación trabajaremos a la inversa; es decir, a partir de una probabilidad tendremos que hallar el valor numérico asociado a ella.

Ejemplo 5.3.4. Sea X la cantidad de radiación que puede ser absorbida por un individuo antes de que le sobrevenga la muerte. Supongamos que X es normal, con una media de 500 roentgen y una desviación típica de 150 roentgen. ¿Por encima de qué nivel de dosificación sobreviviría solamente el 5 % de los expuestos?

En este caso, nos piden calcular el punto x_0 señalado en la Figura 5.19. En términos de probabilidades queremos hallar el punto x_0, tal que

$$P[X \geqslant x_0] = 0.05$$

Tipificando, se tiene

$$P[X \geqslant x_0] = P\left[\frac{X - 500}{150} \geqslant \frac{x_0 - 500}{150}\right]$$
$$= P\left[Z \geqslant \frac{x_0 - 500}{150}\right] = 0.05$$

De este modo, $(x_0 - 500)/150$ es el punto de la curva normal tipificada que deja un 5 % del área a la derecha y el 95 % restante a la izquierda. De la Tabla III en el Apéndice B, obtenemos que el valor numérico de este punto es 1.645. Igualando tendremos:

$$\frac{x_0 - 500}{150} = 1.645$$

La resolución de esta ecuación para x_0 nos da el nivel de dosificación buscado:

$$x_0 = 150(1.645) + 500 = 746.75 \text{ roentgen}$$

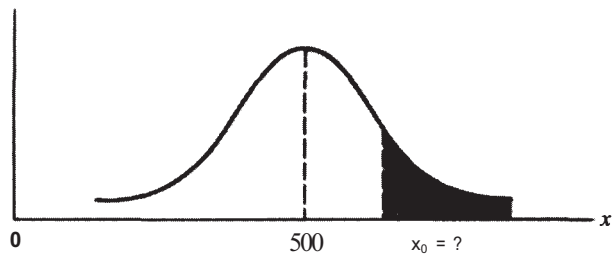


Figura 5.19. $P[X \geqslant x_0] = 0.05$.

EJERCICIOS 5.3

1. La variable X es normal, con media 5 y varianza 4.
   a) Encontrar la expresión de la densidad f de X.
   b) ¿En qué valores de X están situados los puntos de inflexión de la gráfica de f?
   c) Dibujar aproximadamente la gráfica de f.
   d) Sombrear en el dibujo anterior la región correspondiente a $P[3 < X < 7]$.
2. La almeja roja es un importante producto procedente de las costas de Virginia. Varios trabajos han sido desarrollados para estimar el tamaño de la población y para determinar las propiedades físicas y reproductoras de estas almejas. Un estudio reciente ha puesto de manifiesto que la variable aleatoria H, altura de la almeja, se distribuye aproximadamente según la curva normal con una media de 20.3 mm y una desviación típica de 1.4 mm (véase Fig. 5.20). (Basado en la información encontrada en «Population Structure of the Arkshell Clams», de Katherine McGraw y Sally Dennis, Departamento de Biología, Universidad de Radford, y Michael Castagna, Instituto de Ciencias Marinas de Virginia, College de William y Mary, 1996, artículo técnico preparado para la Administración Nacional Oceánica y Atmosférica: Servicio Nacional de Pesca Marina.)
   a) Dibujar la gráfica de la densidad de H. Identificar en ella la media y los puntos de inflexión.
   b) Sombrear el área correspondiente a la probabilidad de que la siguiente almeja que se encuentre tenga una altura que exceda los 23 mm.
   c) ¿Cuál es la probabilidad de que la siguiente almeja que se encuentre tenga exactamente una altura de 20 mm?
3. Utilizar la Tabla III del Apéndice B para calcular:
   a) $P[Z \leqslant -1.52]$.
   b) $P[Z \leqslant 1.37]$.
   c) $F(1.37)$.
   d) $P[Z \geqslant -1.42]$.
   e) $P[Z \geqslant 1.98]$.
   f) $P[-1.21 \leqslant Z \leqslant 1.73]$.
   g) $P[Z = 1.50]$.
   h) El punto z tal que $P[Z \leqslant z] \cong 0.05$.
   i) El punto z tal que $P[Z \leqslant z] \cong 0.75$.
   j) El punto z tal que $P[Z \geqslant z] \cong 0.10$.

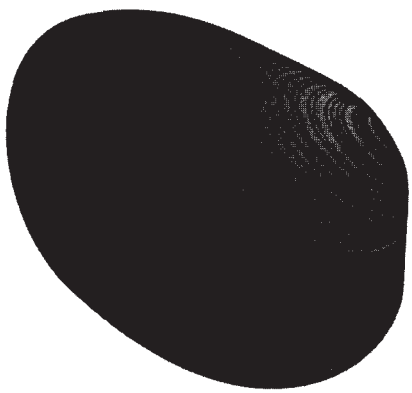


Figura 5.20. La altura de la almeja es la distancia señalada como H.

k) El punto z tal que $P[Z \geqslant z] \cong 0.80$.
l) El punto z tal que $P[-z < Z \leqslant z] \cong 0.95$.
m) El punto z tal que $P[-z \leqslant Z \leqslant z] \cong 0.99$.

4. Sea X una variable normal con media 4 y varianza 9. ¿Cuál es la distribución de la variable $(X - 4)/3$?
5. El número de Btu *(unidades térmicas británicas)* de petróleo y de derivados del petróleo consumidos por persona en Estados Unidos en 1975 está distribuido normalmente con media 153 millones y desviación típica 25 millones.
   a) Hallar $P[X \leqslant 100$ millones$]$.
   b) Hallar $P[X \geqslant 180$ millones$]$.
   c) Hallar P[100 millones $\leqslant X \leqslant 175$ millones].
   d) Hallar P[128 millones $\leqslant X \leqslant 178$ millones].
   e) Hallar el punto x_0 tal que $P[X \leqslant x_0] \cong 0.10$.
   f) Hallar el punto x_0 tal que $P[X \geqslant x_0] \cong 0.06$.
6. En 1969, se descubrió que los faisanes de Montana padecían una apreciable contaminación por mercurio que podía deberse a que habían comido semillas de plantas que fueron tratadas durante su crecimiento con metilo de mercurio. Sea X el nivel de mercurio de un pájaro en partes por millón. Supóngase que X está normalmente distribuida con media 0.25 y desviación típica 0.08. Se mata a un faisán y se determina su nivel de mercurio. Calcular $P[X \leqslant 0.3]$, $P[X \geqslant 0.17]$, P[0.2 $\leqslant X \leqslant$ 0.4] y P[0.01 $\leqslant X \leqslant$ 0.49].
7. Entre los diabéticos, a la glucemia en ayunas puede suponérsele una distribución aproximadamente normal con media 106 mg/100 mL y desviación típica 8 mg/100 mL.
   a) Calcular $P[X \leqslant 120$ mg/100 mL$]$.
   b) ¿Qué porcentaje de diabéticos tendrá niveles entre 90 y 120 mg/100 mL?
   c) Calcular $P[106 \leqslant X \leqslant 110]$.
   d) Calcular $P[X \geqslant 121$ mg/100 mL$]$.
   e) Encontrar un punto x_0 que tenga la propiedad de que el 25 % de los diabéticos tenga una glucemia en ayunas X por debajo de él.
8. En cierta población de primates, el volumen de la cavidad craneal X se distribuye aproximadamente según una curva normal con media 1200 cm^3 y desviación típica 140 cm^3.
   a) Calcular la probabilidad de que un miembro de la población seleccionado aleatoriamente tenga una cavidad craneal superior a 1400 cm^3.
   b) Hallar P[1000 $\leqslant X \leqslant$ 1050].
   c) Hallar $P[X \leqslant 1060]$.
   d) Hallar $P[X \leqslant 920]$.
   e) Encontrar un punto x_0 tal que el 20 % de los primates tenga una cavidad craneal más pequeña que él.
   f) Encontrar un punto x_0 tal que el 10 % de los primates tenga una cavidad craneal superior a él.
9. La densidad del suelo se define como la masa de materia sólida seca por unidad de volumen. Una densidad elevada implica un suelo compacto con escasos poros. Esta densidad es un factor importante para el crecimiento de las raíces, la emergencia de los brotes y la ventilación. Sea X la densidad de la tierra arcillosa Pima. Los estudios demuestran que X tiene una distribución normal con $\mu = 1.5$ y $\sigma = 0.2$ g/cm^3. *(McGraw-Hill Yearbook of Science and Technology,* 1981, pág. 361.)
   a) ¿Cuál es la densidad de X? Haga un esbozo de su función de densidad e indique sobre él la probabilidad de que X esté entre 1.1 y 1.9. Calcule esta probabilidad.
   b) Calcule la probabilidad de que una muestra de tierra arcillosa Pima tomada aleatoriamente tenga una densidad menor de 0.9 g/cm^3.

c) ¿Se sorprendería si una muestra de este tipo de suelo, seleccionada aleatoriamente, tuviese una densidad con un exceso de 2.0 g/cm^3? Explique este hecho basándose en la probabilidad de que ocurra esto.

10. La mayor parte de las galaxias adopta la forma de un disco aplanado con la mayoría de la luz proviniendo de este delgado plano fundamental. El grado de aplanamiento es distinto para cada galaxia; así, en la Vía Láctea, la mayoría de los gases está concentrada en torno al centro del plano fundamental. Sea X la distancia perpendicular desde ese centro a la masa gaseosa; esta variable tiene una distribución normal con media 0 y desviación típica 100 *parsecs*. (Un *parsec* equivale aproximadamente a 19.2 trillones de millas.) *(Enciclopedia de Ciencia y Tecnología, McGraw-Hill,* vol. 6, 1971, pág. 10.)
    a) Haga un esbozo de la gráfica de densidad de X e indique sobre ella la probabilidad de que una masa gaseosa se encuentre a una distancia dentro de 200 *parsecs* del centro del plano fundamental. Calcule esta probabilidad.
    b) ¿Qué porcentaje aproximado del conjunto de masas gaseosas se encuentra a más de 250 *parsecs* del centro del plano?
    c) ¿Qué distancia tiene la propiedad de que el 20 % de las masas gaseosas estén al menos así de lejos del centro del plano?

5.4. REGLA DE LA PROBABILIDAD NORMAL Y TABLAS MÉDICAS (OPCIONAL)

En el Capítulo 4 comentamos ya la desigualdad de Chebyshev. Esta desigualdad nos dice que «la probabilidad de que una variable aleatoria tome un valor dentro de una distancia máxima de su media de k veces la desviación típica es al menos de $1\text{-}1/k^2$». Debemos tener en cuenta que no existe ninguna restricción sobre las variables aleatorias a las que podemos aplicar esta regla. Si tenemos $k = 2$, la desigualdad nos asegura que la probabilidad de que la variable aleatoria esté a una distancia máxima de 2 veces la desviación típica de su media, es por lo menos de $1\text{-}1/2^2 = 0.75$. En otras palabras, podemos llegar a la conclusión de que para cualquier variable aleatoria X:

$$P[\mu - 2\sigma < X < \mu + 2\sigma] \cong 0.75$$

Si X tiene una distribución normal, podemos hacer una afirmación más contundente basándonos en una regla conocida como regla de la probabilidad normal, tal y como veremos a continuación.

Regla de la probabilidad normal. Sea X una variable aleatoria de distribución normal con su correspondiente μ y varianza σ^2; entonces:

a) La probabilidad de que X tome un valor a una distancia máxima de una desviación típica de su media es 0.68 ($P[\mu - \sigma < X < \mu + \sigma] \cong 0.68$).
b) La probabilidad de que X tome un valor a una distancia máxima de su media de dos veces la desviación típica es 0.95 ($P[\mu - 2\sigma < X < \mu + 2\sigma] \cong 0.95$).
c) La probabilidad de que X tome un valor a una distancia máxima de tres veces la desviación típica de su media es 0.99 ($P[\mu - 3\sigma < X < \mu + 3\sigma] \cong 0.99$).

Vemos que X tiene una distribución normal

$$P[\mu - 2\sigma < X < \mu + 2\sigma] \cong 0.95$$

en vez del 0.75 garantizado por la desigualdad de Chebyshev. La regla se deduce fácilmente.

Ejemplo 5.4.1. Para poder comprobar el apartado α de la regla de probabilidad normal, restaremos primero μ para obtener

$$\begin{aligned} P[\mu - \sigma < X < \mu + \sigma] &= P[\mu - \sigma - \mu < X - \mu < \mu + \sigma - \mu] \\ &= P[-\sigma < X - \mu < \sigma] \end{aligned}$$

Completaremos la tipificación dividiendo cada elemento de la desigualdad por σ y reemplazando $(X - \mu)/\sigma$ por Z, con lo que podremos concluir que:

$$\begin{aligned} P[\mu - \sigma < X < \mu + \sigma] &= P[-\sigma < X - \mu < \sigma] \\ &= P\left[\frac{-\sigma}{\sigma} < \frac{X - \mu}{\sigma} < \frac{\sigma}{\sigma}\right] \\ &= P[-1 < Z < 1] \end{aligned}$$

A partir de la tabla de la normal tipificada, sabemos que:

$$\begin{aligned} P[-1 < Z < 1] &= P[Z < 1] - P[Z \leqslant -1] \\ &= 0.8413 - 0.1587 \\ &= 0.6826 \end{aligned}$$

La regla de probabilidad normal está representada en la Figura 5.21, y debe ser considerada sólo como una regla rápida de tanteo que permite saber si un determinado valor de una variable aleatoria normal es inusualmente grande o pequeño o si, por el contrario, es un valor habitual, todo ello basándonos exclusivamente en el conocimiento de la media y de la desviación típica de dicha variable. Por ejemplo, supongamos que la cantidad media de lluvia caída durante el mes de junio en una región concreta es de 23 cm con una desviación típica de 5 cm. Supongamos también que estos datos indican que estamos ante una variable aleatoria con una distribución aproximadamente normal. ¿Podríamos considerar como extremadamente húmedo a un mes de junio en el que cayeran más de 33 cm de lluvia? La respuesta es afirmativa ya que, gracias a la regla de la probabilidad normal, sabemos que la probabilidad de que la lluvia máxima caída en junio se encuentre a dos veces la desviación típica de su media es 0.95; o lo que es lo mismo, que existe un 95 % de posibilidades de que la cantidad de lluvia total se encuentre dentro del intervalo $\mu \pm 2o$ o, en este caso, $23 \pm 2(5$ cm). Es decir, puede decirse que existe una alta probabilidad de que la cantidad máxima de lluvia esté entre los 13 y 33 cm, quedando sólo alrededor de un 2.5 % la posibilidad de ver un valor superior a los 33 cm. Así pues, y dado que esta probabilidad es realmente pequeña, un mes de junio en el que ocurra esto podría ser considerado como inusualmente húmedo.

Una de las aplicaciones más frecuentes de la regla de probabilidad normal surge en el contexto médico. Como usted ya sabe, cuando se toma una muestra de sangre se realizan varios análisis sobre ella; por ejemplo, suelen medirse de forma sistemática los niveles de potasio, sodio, proteínas totales, calcio y colesterol. Durante varios años se han recogido mediciones procedentes de un gran número de personas, información que ha sido utilizada para establecer con un alto grado de precisión, los niveles medios y la cantidad de variabilidad esperada en individuos sanos. Estos valores pueden ser utilizados para establecer lo que llamamos «límites 2-sigma», $\mu \pm 2o$, para cada variable estudiada, ya que gracias a la regla de la probabilidad normal sabemos que aproximadamente un 95 % de las personas sanas estarán dentro de estos límites; afortunadamente tan sólo un 5 % de la población estará fuera de ellos, de los cuales un 2.5 % presentará niveles anormalmente altos y el 2.5 % restante los presentará anormalmente bajos.

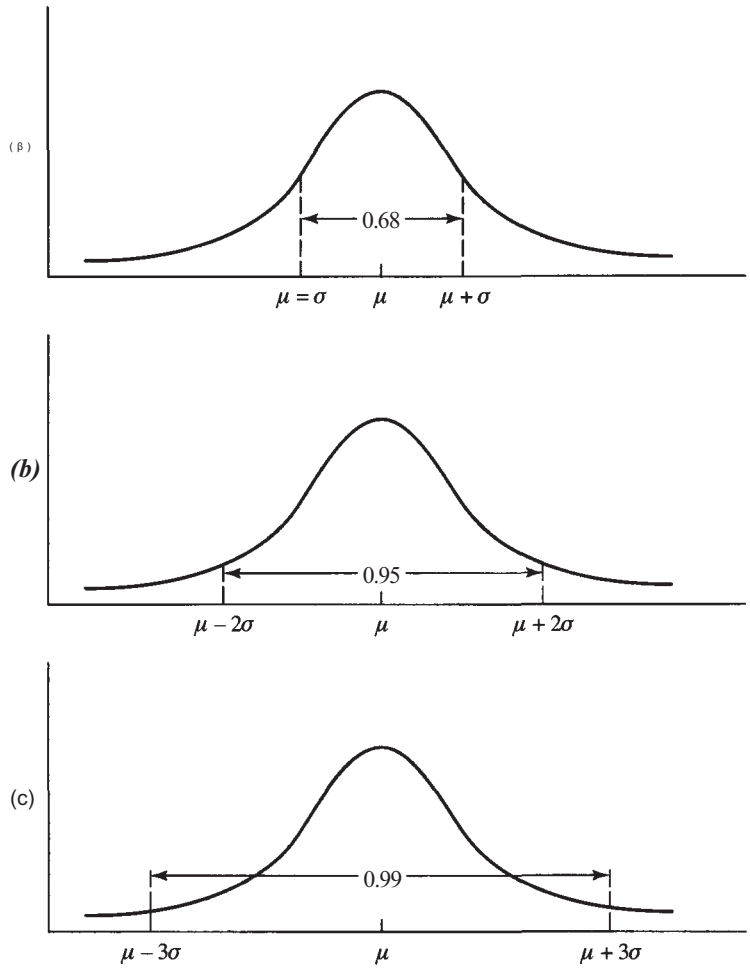


Figura 5.21. Regla de la probabilidad normal. (a) $P[\mu - \sigma < X < \mu + \sigma] \cong 0.68$; (*b*) $P[\mu - 2\sigma < X < \mu + 2\sigma] \cong 0.95$; (c) $P[\mu - 3\sigma < X < \mu + 3\sigma] \cong 0.99$.

Advirtamos que existen dos razones para observar valores más allá de los límites 2-sigma en las tablas médicas establecidas. Una persona puede estar perfectamente sana y, simplemente, ocurrir que sus niveles «normales» son inusualmente altos o bajos al compararlos con los de la población general; sin embargo, también puede darse el caso de que esta persona tenga algún problema. De este modo, la presencia de niveles inusuales debe tomarse sólo como una señal de aviso que indica la necesidad de seguir indagando.

Ejemplo 5.4.2. El nivel de potasio en una persona sana tiene una media de 4.4 ¡y una desviación típica de 0.45 mEq/L. Por la regla de la probabilidad normal sabemos que aproximadamente un 95 % de todas las personas sanas presentan lecturas comprendidas entre μ - 2σ = 3.5 y μ 2σ - 5.3. Si la lectura de un individuo en concreto está entre estos valores, el médico deducirá que no existe ningún problema con esta variable pero si la lectura está por debajo de 3.5 ó por encima de 5.3 estaremos frente a un valor inusual. Esto no significa necesariamente que exista una anomalía, dado que un 5 % de las personas sanas presentará de manera «normal» estos valores inusuales, aunque sí debería ser considerado por el médi-

co como un indicador de que podría haber algún problema. En esta situación, probablemente lo más adecuado sería someter al paciente a algún tipo de seguimiento. (Tomado de los informes químicos analíticos utilizados habitualmente en el Bapüst Medical Center, Columbia, S. C.)

Los informes que los laboratorios envían a los médicos tienen presentaciones diversas. Habitualmente, tales informes dan una lista de los límites 2-sigma para cada variable y marcan los valores considerados como anómalos. En la Figura *5.22a,* los límites 2-sigma aparecen sombreados, de tal manera que cualquier valor que se encuentre en esta región se considerará como aceptable mientras que los que se encuentren fuera serán inusuales. Por ejemplo, un nivel de 4.0 para el potasio es aceptable y el gráfico nos lo muestra. En la Figura *5.22b,* los límites 2-sigma aparecen indicados a la derecha del informe. Un valor fuera del rango de referencia será señalado escribiéndolo en la columna central sombreada del diagrama; así, por ejemplo, una lectura inusualmente alta del potasio se señalará escribiendo 5.7 | en la región sombreada tal y como muestra la figura.

Recuérdese, sin embargo, que los valores fuera de rango no son necesariamente malos. Por ejemplo, un valor de colesterol por debajo de lo normal podría no ser importante, mientras que uno más alto de lo normal no sería deseable. Este tipo de juicio es el que debe realizar el médico.

EJERCICIOS 5.4

1. Considere los datos del Ejercicio 2 de la Sección 5.3. Basándose en la regla de la probabilidad normal, ¿pensaría que resultaría inusual encontrar una almeja cuya altura estuviera entre los 18.9 y los 21.7 mm? Explíquelo.
2. Tome la variable aleatoria X del Ejemplo 5.3.4. Utilice la regla de la probabilidad normal para calcular $P[350 \leqslant X \leqslant 650]$.

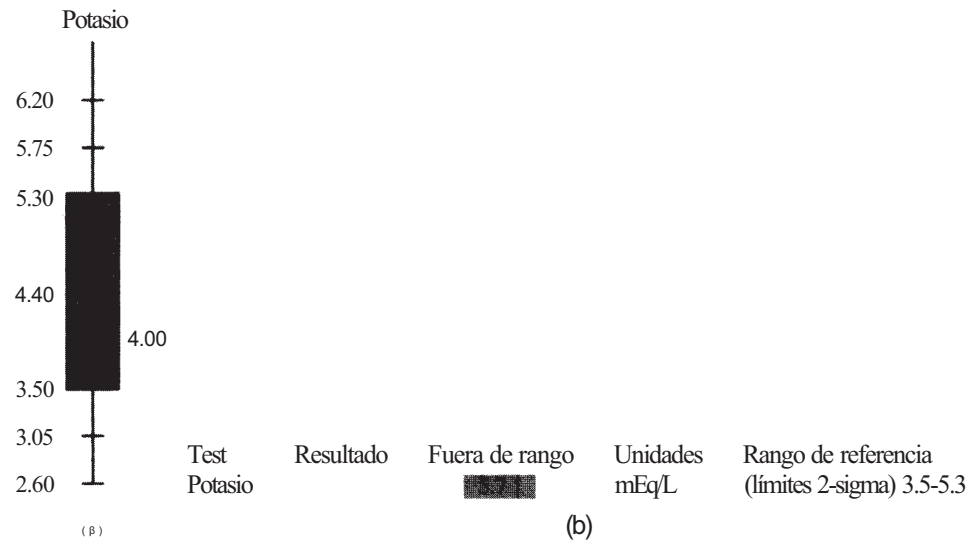


Figura 5.22. (a) Se consideran normales los valores comprendidos en el área sombreada. El valor 4.0 es aceptable, (*b)* Los valores fuera del área sombreada pueden sugerir la necesidad de una investigación más detallada. El valor 5.7 es anormalmente alto.

3. Sea X la variable aleatoria del Ejemplo 5.3.3. Sin utilizar la tabla normal, ¿cuál es la probabilidad aproximada de que la concentración de plomo, en un individuo elegido 41 azar, esté comprendida entre 0.03 y 0.47?
4. El nivel de sodio en personas sanas tiene una media de 141.5 y una desviación típica de 3.25 meq/L.
   a) ¿Cuáles son los límites 2-sigma para esta variable?
   b) Construya un gráfico similar al de la Figura 5.22a y señale en él una lectura de 149. ¿Podemos decir que este valor es inusual?
   (Tomado de los informes químicos analíticos utilizados comúnmente en el Baptist Medical Center, Columbia, S. C.)
5. El nivel de colesterol en individuos sanos depende de la edad y del sexo. En varones menores de 21 años, el valor medio es de 160 mg/dL con una desviación típica de 10 mg/dL; los varones de 21 a 29 años tienen una media de 200 y una desviación típica de 30; por encima de los 30, los varones tienen un valor medio de 220 y una desviación típica de 30.
   a) ¿Cuáles son los límites 2-sigma para cada uno de estos tres grupos de edad?
   b) Construya una tabla similar a la que teníamos en la Figura *5.22b* y marque las lecturas obtenidas para estos individuos:

Edad	Nivel de colesterol
20	125
20	200
18	165
25	200
28	160
35	200
38	210
60	270

   c) Indique y comente los datos inusuales.
   (Tomado de los informes químicos analíticos utilizados habitualmente en el Baptist Medical Center, Columbia, S. C.)
6. En el caso de mujeres, los límites 2-sigma para el colesterol en distintos grupos de edad son:

Edad	Límites 2-sigma
Menos de 21 años	140-180
De 21 a 49 años	140-280
50 o más años	180-280

   a) Calcule la media y la desviación típica aplicada en cada caso.
   b) Construya una tabla 2-sigma similar a la de la Figura 5.22a y señale las lecturas obtenidas para cada una de estas mujeres:

Edad	Nivel de colesterol
20	125
20	200
18	165
25	200
28	160
35	200
38	210
60	270

c) Señale y comente cualquier dato que le parezca inusual.
d) Compare sus resultados con los del Ejercicio 5.
(Tomado de los informes químicos analíticos utilizados comúnmente en el Baptist Medical Center, Columbia, S. C.)

HERRAMIENTAS COMPUTACIÓNALES

TI83

La calculadora TI83 está programada para calcular áreas y puntos asociados a cualquier curva normal. Los valores proporcionados por ella serán muy similares a los encontrados para los ejemplos de este libro aunque si se dan ligeras diferencias, habrá que considerar que los valores de la calculadora son más exactos.

XI. Cálculo de probabilidades Z

Para utilizar la TI83 en el cálculo de probabilidades Z, debemos especificar los límites inferior y superior. Así, para encontrar el área a la izquierda de cualquier punto, usaremos -5 como límite inferior y el propio punto como límite superior; para hallar el área a la derecha de un punto, usaremos ese punto como límite inferior y el 5 como límite superior, y para calcular el área entre dos valores dados de Z, usaremos esos mismos valores como límites. Como ejemplo, calcularemos P[Z $\leqslant$ 1.56], *P[Z* $\geqslant$ -1.29] y P[-1.72 $\leqslant$ Z $\leqslant$ 1.80]. Al margen de pequeñas diferencias, las respuestas que obtengamos serán las mismas que las del Ejemplo 5.3.2.

Tecla/Comando de la TI83	Propósito
1. 2^{ND} DISTR 2	1. Muestra en la pantalla la distribución normal acumulativa; normalcdf(.
2. (-)5	2. Introduce el valor -5 como límite inferior.
3. 1.56) ENTER	3. Introduce 1.56 como límite superior; efectúa los cálculos y muestra *P[Z* $\leqslant$ 1.56] = 0.9406197625.
4. CLEAR	4. Borra la pantalla.
5. 2^{ND} DISTR 2	5. Muestra en la pantalla la distribución acumulativa normal.

6. (-)1.29	6. Introduce -1.29 como límite inferior.
7. 5) ENTER	7. Introduce 5 como límite superior; efectúa los cálculos y muestra $P[Z \geqslant -1.29] = 0.9014743186$.
8. CLEAR	8. Borra la pantalla.
9. 2^{ND} DISTR 2	9. Muestra en la pantalla la distribución acumulativa normal.
10. (-)1.72	10. Introduce -1.72 como límite inferior.
11. 1.80) ENTER	11. Introduce 1.80 como límite superior; efectúa los cálculos y muestra $P[-1.72 \leqslant Z \leqslant 1.80] = 0.9213535499$.

XII. Cálculo de puntos Z

La calculadora TI83 puede hallar los valores z correspondientes a un área determinada a su izquierda. Para ilustrarlo, encontraremos los puntos z tal que $P[Z \leqslant z] = 0.025$, $P[Z \leqslant z] = 0.95$ y $P[Z \leqslant z] = 0.10$. Estos son los puntos cuyos valores estimamos en el Ejemplo 5.3.2 a partir de la tabla de Z. Como podrá ver, estas estimaciones son bastante buenas.

Tecla/Comando de la TI83	Propósito
1. 2^{ND} DISTR 3	1. Muestra en la pantalla la distribución normal inversa; invNorm(.
2. 0.025) ENTER	2. Introduce el valor 0.025; calcula y muestra el punto *z* que tiene un área de 0.025 a su izquierda; z =-1.959963986.
3. CLEAR	3. Borra la pantalla.
4. 2^{ND} DISTR 3	4. Muestra en la pantalla la normal inversa.
5. 0.95 ENTER	5. Introduce 0.95; calcula y muestra el punto z con área 0.95 a su izquierda; z = 1.644853626.
6. 2^{ND} DISTR 3	6. Muestra en la pantalla la normal inversa.
7. 0.10) ENTER	7. Introduce 0.10; calcula y muestra el punto z con área 0.10 a su izquierda; z = -1.281551567.

XIII. Cálculo de probabilidades normales de forma directa

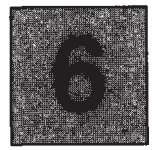
La calculadora TI83 es capaz de encontrar probabilidades sin necesidad de tipificar la variable aleatoria normal *X*. Para hacerlo, debemos especificar un límite inferior y superior de *X*, e identificar su media y su desviación típica. Para ilustrarlo, replantearemos el Ejemplo 5.3.3 en el que $\mu = 0.25$ y $\sigma = 0.11$. Calcularemos $P[X \geqslant 0.6]$ y $P[0.4 \leqslant X \leqslant 0.6]$.

Tecla/Comando de la TI83	**Propósito**
1. 2^{ND} DISTR 2	1. Muestra en la pantalla la distribución normal acumulativa; normalcdf(.
2. 0.6	2. Introduce el valor 0.6 como límite inferior.
3. 0.80	3. Introduce 0.8 como límite superior y usa 5 veces la desviación típica por encima de la media como límite superior razonable; 0.25+ 5(0.11) = 0.80.
4. 0.25	4. Introduce 0.25 como μ.
5. 0.11) ENTER	5. Introduce 0.11 como σ; efectúa los cálculos y muestra $P[X \geqslant 0.6] = 0.0007315462284$.
6. CLEAR	6. Borra la pantalla.
7. 2^{ND} DISTR 2	7. Muestra en la pantalla la distribución acumulativa normal.
8. 0.4	8. Introduce 0.4 como límite inferior.
9. 0.6	9. Introduce 0.6 como límite superior.
10. 0.25	10. Introduce 0.25 como μ.
11. 0.11) ENTER	11. Introduce 0.11 como σ; efectúa los cálculos y muestra $P[0.4 \leqslant X \leqslant 0.6] = 0.0856092456$.

XIV. Búsqueda de puntos normales

Finalmente, también puede utilizarse la calculadora TI83 para calcular el valor de X correspondiente a una determinada área a su izquierda. Para hacerlo, debemos identificar el área deseada y los valores numéricos de μ *y o.* Consideraremos el Ejemplo 5.3.4 en el que sabíamos que $\mu = 500$ y $\sigma = 150$ y queríamos calcular el punto x_0 tal que $P[X \geqslant x_0] = 0.05$ o, lo que es lo mismo, $P[X \leqslant x_0] = 0.95$. Para localizar este punto, debemos seguir estos pasos:

Tecla/Comando de la TI83	
1. 2^{ND} DISTR 3	1. Muestra en la pantalla la distribución normal inversa; invNorm(.
2. 0.95	2. Introduce el valor 0.95 como área a la izquierda deseada.
3. 500	3. Introduce 500 como μ.
4. 150) ENTER	4. Introduce 150 como σ; calcula y muestra los puntos x_0 que dejan un 95 % del área a su izquierda; $x_0 = 746.7280439$.



Inferencias sobre la media

Recordemos que el objeto de un estudio estadístico es doble. Deseamos describir la muestra que tenemos a mano y queremos sacar conclusiones o inferencias sobre la población de donde hemos extraído dicha muestra. Las técnicas que se exponen en el Capítulo 1 son suficientes para cumplir el primer objetivo. Las que se presentan en el resto del texto nos permitirán cumplir el segundo. Las decisiones tomadas respecto de la población, a partir de la información de la muestra, se basan en la probabilidad. Los conceptos relativos a las variables aleatorias discretas y continuas estudiados en los Capítulos 4 y 5 se utilizarán ampliamente en todo lo que sigue.

6.1. MUESTREO ALEATORIO Y ALEATORIZACIÓN

Tal como se ha indicado en el Capítulo 1, las inferencias sobre la población se efectúan a partir de la información obtenida de una muestra extraída de la población. La muestra es contemplada como una población en miniatura. Esperamos que el comportamiento de la variable aleatoria sobre la muestra sea una descripción precisa de su comportamiento sobre la población. Por esta razón, tomamos los valores observados de los estadísticos calculados para la muestra como aproximaciones para los parámetros de población correspondientes. Esta idea se expone en la Figura 6.1.

¿Cómo obtendremos una muestra aleatoria? No es una pregunta fácil de responder. El investigador dispone de muchos tipos de muestreo diferentes. El que se elija para un estudio concreto depende de muchas cosas. Factores como el tamaño de la población, el tipo de población, preguntas que se deben responder, tiempo y recursos disponibles, y la precisión deseada, contribuyen a la elección de la técnica de muestreo. Esto debería ser evidente. No es razonable suponer que un procedimiento diseñado para hacer un muestreo de los ficheros del archivo de un hospital, funcionará igualmente al muestrear árboles en un parque nacional o al muestrear una especie de ballenas del océano Pacífico, en peligro de extinción. En la *etapa de diseño* de un estudio, el investigador debe consultar al profesional de la estadística para que le ayude a elegir un tipo de muestreo apropiado. Deberá asesorarle sobre la forma de extraer la muestra, qué información obtener de cada objeto incluido en la muestra y cuál deberá ser el tamaño de la misma para satisfacer los objetivos del estudio.

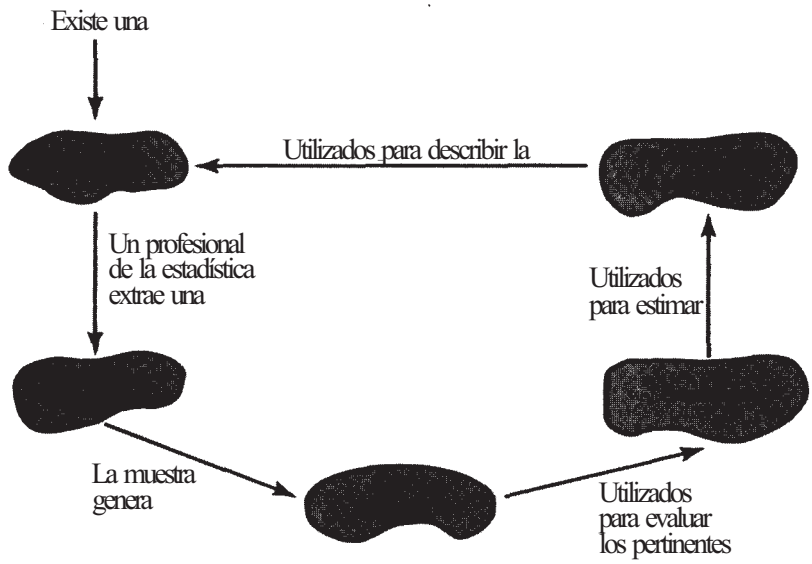


Figura 6.1. Esquema de cómo se lleva a cabo un estudio estadístico.

Muestreo aleatorio simple

Ahora consideraremos el *muestreo aleatorio simple.* Este tipo de muestreo funciona bien en muchos casos, y es fácil de utilizar y comprender; por lo general, acude a la mente cuanc o una persona no entendida oye el término *muestra.* En el muestreo aleatorio simple los objetos se seleccionan «al azar», en el sentido de que la elección de los objetos muestreados es controlada por algún mecanismo aleatorio. El investigador no incluirá ni excluirá deliberadamente un objeto particular en la muestra. Cada objeto de la población tiene las mismas probabilidades de ser elegido para el estudio que cualquier otro. El mecanismo aleatorio utilizado para lograrlo podría ser tan simple como sacar nombres de un sombrero, o tan complejo como utilizar un generador electrónico de números aleatorios.

Una de las formas más sencillas de seleccionar una muestra aleatoria simple de una población finita es utilizando una tabla de números aleatorios. Esta tabla se genera de tal forma que cada uno de los dígitos de 0 a 9 tiene la misma probabilidad de aparecer en una posición dada de la tabla que cualquier otro. De este tipo es la Tabla IV del Apéndice B. Su manejo se explica en el Ejemplo 6.1.1. Se ha elegido una muestra de pequeño tamaño, para explicar rápidamente la utilización de la tabla. La interpretación de este ejemplo no implica que las inferencias sobre grandes poblaciones se realicen de forma rutinaria utilizando muestras pequeñas. Evidentemente, éste no es el caso.

Ejemplo 6.1.1. Últimamente ha habido problemas en un hospital porque los médicos han ordenado un número excesivo de pruebas de laboratorio para sus pacientes hospitalizados. El administrador está interesado en estudiar la situación. Una pregunta a contestar es: ¿cuál fue el número medio de controles ordenados por cada consulta en este hospital el último año?

Supongamos que hay un total de 8000 expedientes de consultas de pacientes. ¿Cómo podemos seleccionar aleatoriamente cinco expedientes para el estudio? Para ello, primero observamos que los expedientes de las consultas están listados en los ficheros y pueden ser numerados del 1 al 8000. Utilizamos la Tabla IV del Apéndice B para obtener cinco números aleatorios con cuatro dígitos (0001 a 8000). Se eligen para estudio los expedientes de los

pacientes correspondientes a los números seleccionados y, por lo tanto, constituyen una muestra aleatoria simple de tamaño 5. De esta forma, no controlamos las visitas elegidas para la muestra y no podemos ser acusados de manipular los resultados del estudio.

Para empezar, se selecciona aleatoriamente un punto inicial. Una forma de hacerlo consiste en escribir los números de las filas (01 a 50) en papeles, colocar estos dentro de una caja y a continuación extraer uno al azar. El número extraído determina la fila de la Tabla IV, que contiene el punto aleatorio inicial. De la misma forma, se selecciona un número aleatorio de columna del 01 al 14. Supongamos que, una vez hecho esto, obtenemos la fila 7 y la columna 3. El número que se encuentra en esta posición de la Tabla IV es el 56420. En la Tabla 6.1 aparece la parte de la tabla que contiene este valor. Empezamos a leerla en este punto. Puede leerse de distintas formas: siguiendo la fila, columna abajo, cada dígito de la columna, o de acuerdo con cualquier otro esquema. La forma más inmediata es leer los primeros cuatro dígitos de la fila y obtener el número aleatorio 5642. Se selecciona así la consulta número 5642 como primer miembro de la muestra aleatoria simple. Siguiendo la columna hacia abajo encontramos que el siguiente número aleatorio de cuatro dígitos es 0546, que corresponde a la visita 546. La tercera y cuarta visitas seleccionadas son las números 6366 y 4334, respectivamente. Continuando columna abajo, el siguiente número aleatorio es el 8823. Puesto que sólo tenemos 8000 consultas, este número se descarta por ser demasiado grande. De este modo, el último miembro de la muestra aleatoria corresponde al siguiente número de la tabla, el 4823. Si se obtiene el mismo número aleatorio más de una vez, se descarta después de que se le haya seleccionado ya una vez. En este momento, ya tenemos una muestra aleatoria simple compuesta por el *registro de cinco consultas.* Pueden examinarse estos expedientes para obtener una muestra de 5 observaciones sobre la variable aleatoria X, número de pruebas de laboratorio encargadas por consulta. A continuación, pueden promediarse estas observaciones para hallar $\bar{x}$, la media de la muestra y la aproximación para la media de la población de la cual se ha extraído la muestra.

Como hemos visto, se han obtenido muestras aleatorias simples utilizando la tabla de números aleatorios. También pueden obtenerse por medio de números aleatorios generados por ordenador o por calculadora. En la sección de Herramientas Computacionales se explica el procedimiento con la TI83 y con un programa del SAS.

La denominación *muestra aleatoria* se utiliza de tres formas diferentes. Por ejemplo, supongamos que queremos realizar un estudio de la variable aleatoria *X,* peso al nacer de niños de madres adictas a la cocaína. Pretendemos extraer una muestra aleatoria de tamaño 10. Antes de que se haya elegido realmente algún niño para el estudio, sabemos que estamos tratando con 10 variables aleatorias $X_1, X_2, \ldots, X_{10}$, donde X_i indica el peso en el momento del nacimiento del i-ésimo niño seleccionado para el estudio. Cada una de estas variables aleatorias puede tomar supuestamente cualquier valor entre quizá 225 y 900 g. Estas 10 variables aleatorias se refieren a una muestra aleatoria de la distribución de X. Se utilizan letras mayúsculas, nuestra notación para variables aleatorias, para subrayar el hecho de que en este punto cada miembro de la muestra es una variable aleatoria. A continuación seleccionamos 10 niños para el estu-

Tabla 6.1

77921	06907	11008
99562	72905	⑤642
96301	91977	05463
89579	14342	63661
85475	36857	43342
28918	69578	88231
63553	40961	48235
09429	93969	52636
10365	61129	87529
07119	97336	71048

dio. Estos niños son una muestra aleatoria extraída de la población de niños nacidos de madres adictas a la cocaína. Aquí cada miembro de la muestra es un niño. Una vez identificadas los niños y registrados sus pesos en el momento del nacimiento se dispone de diez números, $x_1, x_2, \ldots, x_{10}$. Estos números, que representan observaciones sobre las variables aleatorias X_1 $X_2, \ldots, X_{10}$, también pertenecen a una muestra aleatoria. Ahora, cada miembro de la muestra es un número. La Figura 6.2 ilustra estos tres empleos de la denominación *muestra aleatoria.* En la práctica está claro, según el contexto, qué uso debe aplicarse.

Aleatorización

Puede utilizarse la tabla de números aleatorios para «aleatorizar» secuencias de sucesos o para asignar aleatoriamente tratamientos a unidades experimentales. Por ejemplo, supongamos que se pretende realizar un estudio para comparar los efectos de cuatro concentraciones de sulfato diferentes sobre el crecimiento de pinos. Generalmente, el investigador obtiene una colección de plántulas y las divide en cuatro grupos y cada uno de ellos recibe un tratamiento de sulfato diferente. Se supone que al principio del experimento las plántulas son idénticas.

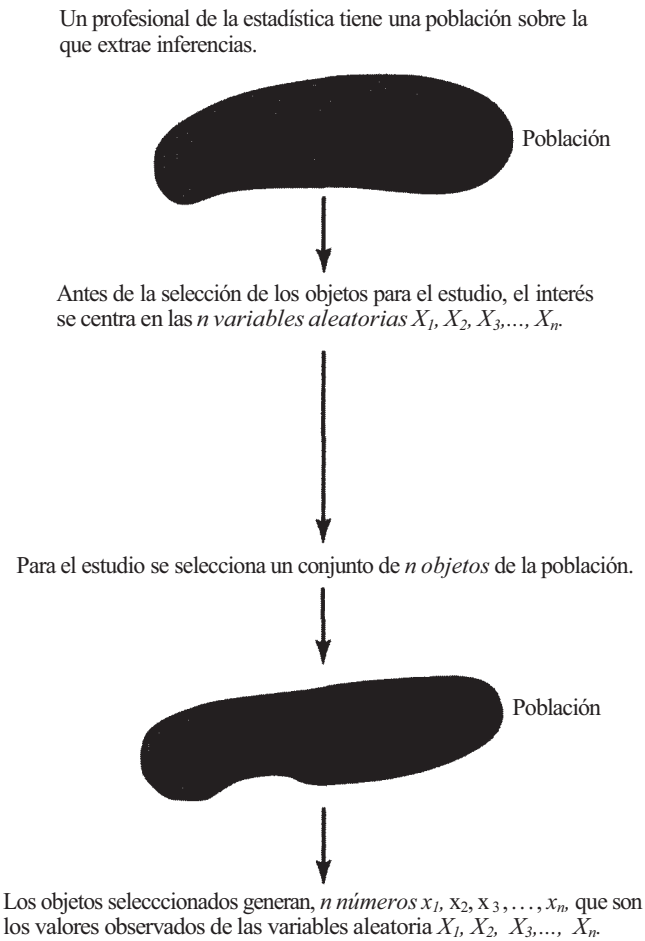


Figura 6.2. Tres formas de entender el término *muestra aleatoria.*

Durante el curso del experimento, se tratan de la misma forma en todos los aspectos, excepto en el nivel de sulfato recibido. Cualquier diferencia observada al final del experimento es atribuible a los diferentes niveles de sulfato que se hayan utilizado. Obsérvese que, aunque suponemos que las plántulas son idénticas al principio del experimento, sabemos que no es cierto. Existen ligeras diferencias de altura, peso o en el estado de salud general de la plántula. Para repartir estas diferencias en los tratamientos y protegerse contra los daños que pueda producir una influencia sistemática presente en el experimento, les asignamos tratamientos aleatoriamente. Para hacerlo, puede utilizarse la tabla de números aleatorios. Por ejemplo, supongamos que tenemos 20 plántulas y queremos asignar aleatoriamente 5 plántulas a cada uno de los cuatro tratamientos. Para ello los numeramos del 01 al 20. A continuación, elegimos aleatoriamente números de dos dígitos de la tabla de números aleatorios. Los primeros cinco números seleccionados entre 01 y 20 reciben el tratamiento *A,* los segundos cinco reciben el tratamiento *B,* los terceros cinco el tratamiento C y los restantes reciben el tratamiento *D.* El Ejemplo 6.1.2 ilustra esta idea.

Ejemplo 6.1.2. La Figura 6.3α muestra 20 plántulas de pino numeradas del 01 al 20. Supongamos que se utiliza la técnica explicada en el Ejemplo 6.1.1 para obtener el punto inicial aleatorio 27958 que se muestra en la Tabla 6.2 (fila 36, columna 3 de la Tabla IV del Apéndice B). A partir de este punto, leemos los primeros dos dígitos siguiendo la columna hacia abajo. A continuación nos desplazamos a la parte inferior de la columna 4 y la leemos hacia arriba. Los primeros cinco números hallados son 18,06,20,07 y 12. Las plántulas con estos números recibirán el tratamiento *A.* Véase la Figura *63b.* Los siguientes cinco números elegidos son 14,09,08,03 y 10. Estas plántulas reciben el tratamiento *B.* Las plántulas 16,02,17,13 y 19 reciben el tratamiento *C,* y el resto el tratamiento *D.* En la Figura 6.3c se muestra la asignación completa de los tratamientos.

Figura 6.3. (a) Las plántulas están numeradas de 01 a 20. *(b)* Las plántulas numeradas 18,06,20,07 y 12 se seleccionan primero y reciben el tratamiento *A. (c)* Se completa la asignación de tratamientos.

Tabla 6.2. El número 27958 hallado en la fila 36 y en la columna 3 nos da el punto inicial aleatorioi para seleccionar aleatoriamente números de dos dígitos del 01 al 20

Fila/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	43342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953

Tabla 6.2. El número 27958 hallado en la fila 36 y en la columna 3 nos da el punto inicial aleatorioi para seleccionar aleatoriamente números de dos dígitos del 01 al 20

Fila/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	43342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953

Esta técnica puede utilizarse en otras aplicaciones. Por ejemplo, para asignar aleatoriamente protocolos experimentales a pacientes en un entorno médico, para asignar aleatoriamente tratamientos experimentales a ratones de laboratorio o para formar aleatoriamente grupos, de manera que se les puedan asignar tareas específicas como miembros de un equipo de investigación.

EJERCICIOS 6.1

1. Utilice la técnica explicada en el Ejemplo 6.1.1 para hallar un punto aleatorio inicial en la Tabla IV del Apéndice B. Utilice este punto inicial para asignar aleatoriamente tratamientos a las 20 plántulas de pino del Ejemplo 6.1.2. ¿A cuántas plántulas se les ha asignado el mismo tratamiento que el asignado por la aleatorización del Ejemplo 6.1.2?
2. *Diversidad de especies.* El índice de diversidad de especies es un índice comparativo que se utiliza para medir el efecto de una perturbación, como la contaminación del agua, en los organismos vivos. Puede determinarse la diversidad de la población antes y después de la perturbación y efectuarse una comparación. En general, un índice pequeño tras ja perturbación es una indicación de que la perturbación ha tenido un efecto negativo. El siguiente ejemplo ilustra la técnica utilizada para determinar el índice en una muestia particular.

Ejemplo. Supongamos que se examina una muestra de agua con el microscopio y se descubre que contiene tres miembros de la especie A, cuatro de la especie *B* y siete de la especie *C,* para un total de 14 microorganismos. Marcamos las posiciones tal como se muestra a continuación:

_ _ _ _ _ _ _ _ _ _ _ _ _ _

De la tabla de números aleatorios seleccionamos 14 números aleatorios de dos dígitos, del 0l al 14. Los tres primeros elegidos determinan las posiciones asignadas a la especie *A,* los siguientes cuatro se asignan a la especie *B* y los últimos siete a la especie C. Supongamos que nuestro punto aleatorio inicial está en la fila 11, columna 5. Leyendo hacia abajo a partir de este punto en la Tabla 6.2, vemos que los primeros tres números elegidos son 09, 13 y 04. Estas posiciones están asignadas a la especie *A,* tal como se muestra a continuación.

_ _ _ *A* _ _ _ _ *A* _ _ _ *A* _

Los siguientes cuatro números elegidos son 02,01,14 y l0 y estas posiciones están asignadas a la especie *B.*

B B _ *A* _ _ _ _ *A B* _ _ *A B*

Las posiciones restantes están asignadas a la especie *C* para obtener la secuencia aleatoria

B B C A C C C C A B C C A B

Ahora contamos el número de series en la secuencia, donde una serie es una secuencia de los mismos microorganismos. A continuación se muestran las series de la secuencia anterior.

B B C A C C C C A B C C A B

La secuencia contiene nueve series. El *índice de comparación secuencial* (SCI, *sequential comparison index)* se define

$$\text{SCI} = \frac{\text{número de series}}{\text{número de especímenes}}$$

En este caso,

$$\text{SCI} = \frac{9}{14} = 0.64$$

Se puede afinar la medida de la diversidad con una segunda medida, el *índice de diversidad* (DI, *diversity index),* que se obtiene multiplicando el SCI por el número de especies diferentes encontradas en la muestra. Así,

$$\text{DI} = (\text{SCI})\ (\text{número de especies})$$

En nuestro ejemplo

$$\text{DI} = 0.64(3) = 1.92$$

Se repite el procedimiento de aleatorización para obtener un segundo SCI y un segundo DI. El índice de diversidad para la muestra se toma como el promedio de los dos valores del DI. Experiencias realizadas han puesto de manifiesto que, cuando este procedimiento se aplica en el estudio de arroyos, un índice de diversidad mayor de 12 indica ausencia de contaminación; valores de este índice iguales o menores de 8 indican una fuerte contaminación. (Información de James Brower, Jerrold Zar y Cari Von Ende, *Field and Laboratory Methodsfor General Ecology,* Editores W. C. Brown, Publishers, Dubuque, Iowa, 1990, págs. 51-52.)

a) Obtener una segunda aleatorización para el ejemplo anterior y calcular su SCI y su DI. Promediar los dos DI disponibles para obtener el índice de diversidad de la muestra.

b) En un estudio sobre el efecto de una planta de tratamiento de aguas residuales sobre los microorganismos que viven en un río, se tomaron muestras de agua de la planta, aguas arriba y aguas abajo. Se obtuvieron los siguientes datos:

Microorganismo	Arriba	Abajo
Diatomeas	12	5
Espiroquetas	2	1
Protistas	4	1

Obtener dos lecturas del SCI y del DI de cada muestra. Hallar el índice de diversidad de cada muestra promediando las dos lecturas del DI. ¿Existe alguna indicación de que la planta de tratamiento de aguas residuales pueda estar afectando la diversidad de microorganismos en el río? (Basado en un estudio realizado por Joseph Hutton, Departamento de Biología y Servicio de Consultoría Estadística, Universidad de Radford, Radford, Virginia, 1990.)

3. Un guarda forestal desea estudiar una muestra de *Pinus taeda* de una gran región arbolada de forma que pueda estimarse el diámetro medio de los árboles a la altura del pecho (DMAP). Para ello, obtiene un mapa topográfico del área. En el mapa hay marcada una cuadrícula de forma que quedan definidos 200 cuadrados de tamaño 10 x 10 m. Se selec-

ciona una muestra aleatoria de 20 cuadros y se obtiene el diámetro de cada pino dentro de cada cuadro.

a) Utilizar un generador de números aleatorios para obtener 20 números aleatorios de tres dígitos, entre 001 y 200.

b) La Tabla V del Apéndice B proporciona información sobre los pinos en los 200 cuadros. Para la muestra, calcular x, el diámetro medio de los árboles de la muestra a la altura del pecho. ¿Es este valor el DMAP de toda la zona? Explicarlo. Comparar el promedio obtenido con el de alguno de sus compañeros de clase.

c) Para la muestra, calcular $\tilde{n}$, número medio de árboles por cuadro. Multiplicar $\tilde{n}$ por 200 para estimar el número total de árboles de la zona. ¿Se aproxima su estimación a 600, el número real de árboles expuestos en la Tabla V? (Datos basados en Harold Burkhart et al., *Yields of Old-Field Loblolly Pine Plantations,* División Forestal y de Recursos Naturales, Pub. FWS-3-72, VPI y SU, Blacksburg, Va.)

4. La Tabla XIII del Apéndice B proporciona el sexo y la presión arterial sistólica y diastólica de 120 pacientes de una clínica particular. Utilizar la tabla de números aleatorios para obtener una muestra aleatoria simple, de tamaño 20, de dicha población. Anotar el sexo y las presiones arteriales de los individuos seleccionados.
5. Un biólogo está investigando el efecto del pH sobre el cultivo del guisante. Van a emplearse 30 macetas que contienen dos plantas cada una. Se utilizarán tres niveles de pH con 10 macetas que recibirán el tratamiento *A,* 10 el tratamiento *B* y el resto el tratamiento *C.* Las 30 macetas se dispondrán en el invernadero sobre una mesa en seis filas, tal como se muestra a continuación:

a) ¿Por qué podría ser arriesgado, desde un punto de vista práctico, tratar a todas las macetas en las dos primeras filas con el primer nivel de pH, todas las macetas de las filas 3 y 4 con el segundo nivel de pH y el resto con el tercer nivel de pH?

b) Utilizar la tabla o un generador de números aleatorios para determinar la asignación de los niveles de pH a las 30 macetas.

6. Se pretende comparar dos fármacos, A y B, con un placebo P. Se realizará una prueba con treinta personas, formando tres grupos de 10 personas cada uno y asignándole aleatof riamente a cada grupo uno de los fármacos. Emplee la técnica explicada en el Ejempío 6.1.1 para asignar aleatoriamente los fármacos a los sujetos. Compare su asignación con la de su compañero de clase. Suponga que es el sujeto número 12, ¿qué tratamiento recibirá a través de su asignación? ¿Y con la de su compañero? ¿En cuántos casos las dos aleatorizaciones asignaron los mismos tratamientos a los sujetos?

6.2. ESTIMACIÓN PUNTUAL DE LA MEDIA E INTRODUCCIÓN A LA ESTIMACIÓN POR INTERVALO: TEOREMA CENTRAL DEL LÍMITE

Ya hemos visto que la media muestral observada se utiliza para describir la posición de la muestra y para aproximarnos al valor medio de la población de la que procede dicha muestra. Está claro que los valores numéricos reales obtenidos para la media muestral variarán de una

muestra a otra. Este estadístico es, por tanto, una variable aleatoria. Para subrayar este punto utilizaremos una letra $\bar{X}$ mayúscula para designar la media muestral en un ámbito general. Una vez extraída la muestra y obtenido el valor numérico para la media muestral, cambiaremos a la letra minúscula como hemos hecho anteriormente. El Ejemplo 6.2.1 ilustra la notación.

Ejemplo 6.2.1. Los investigadores de la Environmental Protection Agency (EPA) se interesan por la calidad del aire. Uno de los indicadores de la calidad del aire es el número medio de microgramos de partículas en suspensión por metro cúbico de aire. Es decir, el interés se centra en μ, la media de la variable aleatoria X, número de microgramos de partículas en suspensión por metro cúbico de aire. Para controlar la situación se hace una lectura cada seis días, extrayendo un metro cúbico de aire a través de un filtro y determinando el número de microgramos de partículas en suspensión concentradas en él. Después de un período de treinta días, se ha generado una muestra aleatoria X_1, X_2, X_3, X_4, X_5, de tamaño 5. Supóngase que los valores observados de estas variables, para el período dado de 30 días son

$x_1 = 58 \quad x_3 = 57 \quad x_5 = 59$

$x_2 = 70 \quad x_4 = 61$

El valor observado del estadístico $\bar{X}$ se halla promediando estos cinco valores. En este caso

$$\bar{x} = \frac{58 + 70 + 57 + 61 + 59}{5} = 61$$

Obsérvese que, después de evaluado el estadístico $\bar{X}$ para esta muestra concreta, cambiamos a letra minúscula.

En este momento, deben introducirse algunos términos nuevos. Un estadístico utilizado para aproximar un parámetro de población se denomina *estimador* del parámetro. El número obtenido cuando se evalúa el estimador para una muestra en particular, es una *estimación* del parámetro. En el Ejemplo 6.2.1, $\bar{X}$ es un estimador de μ; el número 61 es la estimación de μ basada en la muestra dada. El estadístico $\bar{X}$ se denomina *estimador puntual de* μ porque al evaluarlo para una muestra en concreto da un solo número o punto.

El sentido común señala $\bar{X}$ como el estimador más lógico para μ. Se puede probar que este estimador tiene también algunas buenas propiedades matemáticas. En particular, que en un muestreo repetido de una población con media μ los valores de $\bar{X}$ fluctuarán alrededor de μ. También se puede demostrar que para muestras de tamaño grande, los valores de $\bar{X}$ varían muy poco de una muestra a otra. Así, los valores de $\bar{X}$ están centrados en μ, valor que se pretende estimar a través de este estadístico, y para muestras grandes, se espera que la mayoría de los valores observados caigan cerca de μ. Esto significa que si se dispone de una muestra de tamaño moderado y se estima μ por medio de $\bar{x}$, es probable que esta estimación sea bastante precisa. Estas cuestiones se prueban en el Apéndice A y se ilustran en el Ejemplo 6.2.3.

Estimación por intervalos

Hemos visto que la media muestral es un buen estimador puntual de la media poblacional. El inconveniente principal es que un único valor observado de $\bar{X}$ generalmente no es exactamente igual a μ; habrá cierta diferencia entre $\bar{x}$ y μ. Sería conveniente poder tener idea de lo cerca que está nuestra estimación del verdadero valor de la media poblacional. También sería bueno poder dar información de lo seguros o confiados que estamos de la precisión de la estimación.

Para tener una idea, no sólo del valor de la media, sino también de la precisión de la estimación, los investigadores optan por el método de estimación por intervalo o *intervalos*

de confianza. Un intervalo estimador es lo que su propio nombre indica, un intervalo aleatorio, cuyos puntos extremos L_1 y L_2 son estadísticos. Esto se utiliza para determinar un intervalo numérico a partir de una muestra. Se espera que éste contenga el parámetro de la población que está siendo estimado. Al ampliar la estimación de un punto a un intervalo, ganamos un pequeño margen de error, lo cual nos permite, con base en la teoría de la probabilidad, dictaminar sobre la confianza que tenemos en el estimador.

Un intervalo de confianza de μ es un intervalo $[L_1,L_2]$ que incluye a la media con un grado de certidumbre establecido. Por ejemplo, un intervalo de confianza del 95 % es tal que $P[L_X \leqslant \mu \leqslant L_2] \cong 0.95$; un intervalo de confianza del 99% satisface la condición de que $P[L_1 \leqslant \mu \leqslant L_2] \cong 0.99$. Decir que un intervalo es un intervalo de confianza del 95 % de μ significa que, cuando se utiliza en un muestreo repetido de la población, el 95 % de los intervalos resultantes deberá contener a μ; debido al azar, el 5 % no incluirá la verdadera media de la población. El grado de confianza deseado es controlado por el investigador. La Figura 6.4 ilustra esta idea.

Para construir un intervalo de confianza de μ primero hallaremos una variable aleatoria cuya expresión contenga a μ y cuya distribución se conozca, al menos aproximadamente. Para ello debemos considerar de nuevo la distribución de $\bar{X}$. Recordemos que, puesto que $\bar{X}$ es una variable aleatoria, tiene una distribución de probabilidad. ¿Cuál es la forma de la distribución? ¿Cuál es su media?¿Y su varianza? El Teorema 6.2.1 contesta a todas estas cuestiones, en el caso de que las muestras se extraigan de una distribución normal. Este teorema se prueba en el Apéndice A. Para el caso de que las muestras provengan de una distribución que no es normal, se utiliza el Teorema 6.2.2.

Teorema 6.2.1. Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño n de una distribución que es normal con media μ y varianza σ^2. Entonces $\bar{X}$ es normal con media μ y varianza σ^2/n. Además, la variable aleatoria

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

es *normal tipificada.*

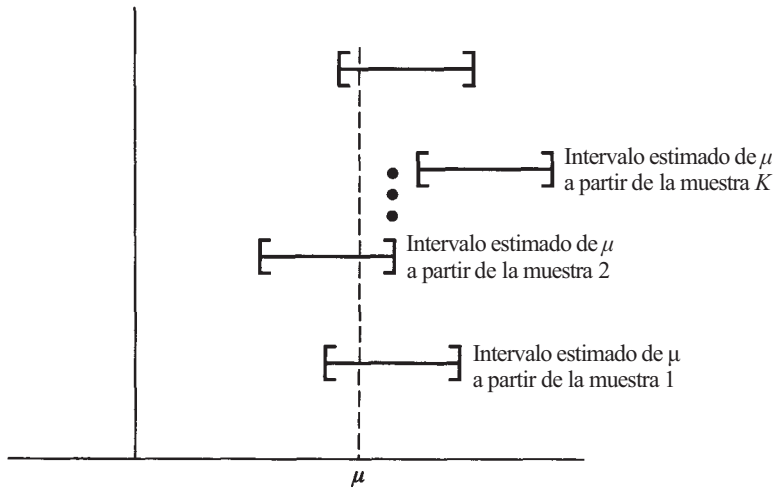


Figura 6.4. De los intervalos construidos utilizando $[L_1, L_2]$ se espera que el 95% contenga a μ, la verdadera pero desconocida media poblacional.

Obsérvese que la variable aleatoria $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ contiene al parámetro μ, y su distribución se sabe que es normal tipificada. Esta variable puede utilizarse para determinar la fórmula general para un intervalo de confianza de μ. Ilustramos el método considerando primero la construcción de un intervalo de confianza del 95 %. La técnica usada puede generalizarse fácilmente para cualquier grado de confianza pretendido.

Ejemplo 6.2.2. Hallemos un intervalo de confianza, del 95 %, de μ, número medio de microgramos de partículas en suspensión por metro cúbico de aire, sobre la base de la muestra aleatoria de tamaño 5 dada en el Ejemplo 6.2.1. De dicho ejemplo se sabe que una estimación puntual de $\mu_{\bar{x}} = 61$. Supóngase que por experiencias anteriores *se sabe que X,* número de microgramos de partículas en suspensión por metro cúbico de aire, está normalmente distribuida, con varianza $\sigma^2 = 9$. Queremos extender la estimación puntual a un intervalo de los números reales, de tal forma que podamos tener una confianza del 95 % de que el intervalo obtenido contenga al verdadero valor de μ. Es decir, queremos determinar L_1 y L_2 de forma que $P[L_1 \leqslant \mu \leqslant L_2] = 0.95$ (véase Fig. 6.5).

Para hacerlo así, consideremos la partición de la curva normal tipificada dibujada en la Figura 6.6. Puede verse que

$$P[-1.96 \leqslant Z \leqslant 1.96] = 0.95$$

En este caso, $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ y, por tanto, podemos concluir que

$$P\left[-1.96 \leqslant \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqslant 1.96\right] = 0.95$$

Para encontrar los puntos extremos de un intervalo de confianza de μ del 95 %, aislamos algebraicamente μ en el centro de la desigualdad precedente:

$$P\left[-1.96 \leqslant \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqslant 1.96\right] = 0.95$$

$$P\left[\frac{-1.96\sigma}{\sqrt{n}} \leqslant \bar{X} - \mu \leqslant \frac{1.96\sigma}{\sqrt{n}}\right] = 0.95$$

$$P\left[-\bar{X} - \frac{1.96\sigma}{\sqrt{n}} \leqslant -\mu \leqslant -\bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right] = 0.95$$

$$P\left[\bar{X} + \frac{1.96\sigma}{\sqrt{n}} \geqslant \mu \geqslant \bar{X} - \frac{1.96\sigma}{\sqrt{n}}\right] = 0.95$$

$$P\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right] = 0.95$$

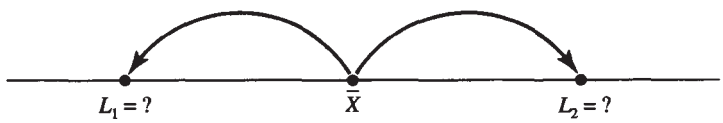


Figura 6.5. L_1 y L_2 son estadísticos tal que $P[L_1 \leqslant \mu \leqslant L_2] \cong 0.95$.

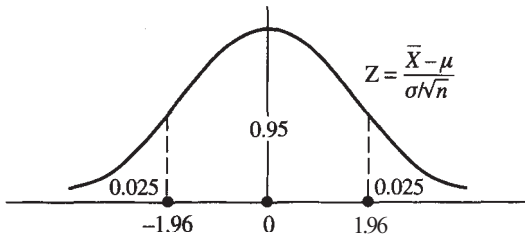


Figura 6.6. Partición de Z para obtener un intervalo de confianza de μ del 95% $P[-1.96 \leqslant Z \leqslant 1.96] = 0.95$

Vemos que los límites superior e inferior del intervalo de confianza del 95 % son

$$L_1 = \bar{X} - \frac{1.96\sigma}{\sqrt{n}} \quad L_2 = \bar{X} + \frac{1.96\sigma}{\sqrt{n}}$$

Puesto que se supone que σ^2 es 9, L_1 y L_2 son estadísticos. Sus valores observados por la muestra son

$$\bar{x} - 1.96\left(\frac{3}{\sqrt{5}}\right) = 61 - 2.63 = 58.37$$

$$\bar{x} + 1.96\left(\frac{3}{\sqrt{5}}\right) = 61 + 2.63 = 63.63$$

(Véase Fig. 6.7.) Puesto que este intervalo se obtuvo usando un procedimiento que, en muestreos repetidos, contendrá a la media en un 95 % de las veces, podemos tener un 95 % de confianza de que μ esté verdaderamente entre 58.37 y 63.63.

Con el fin de generalizar este procedimiento para cualquier grado de confianza deseado, sólo deberemos sustituir el valor de 1.96 por un punto apropiado de la tabla de Z. Por ejemplo, para hallar un intervalo de confianza del 99 % para la media poblacional, comenzamos con la partición de la curva de Z de la Figura 6.8. Los límites para este intervalo son

$$\bar{x} - 2.575\left(\frac{3}{\sqrt{5}}\right) = 61 - 3.45 = 57.55$$

$$\bar{x} + 2.575\left(\frac{3}{\sqrt{5}}\right) = 61 + 3.45 = 64.45$$

Con un 99 % de confianza la media μ estará comprendida entre 57.55 y 64.45. En general, cualquier intervalo de confianza para μ con σ^2 conocida toma la forma

$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

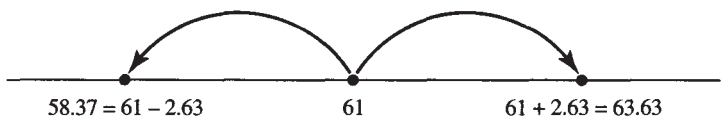


Figura 6.7. Intervalo de confianza de μ del 95%.

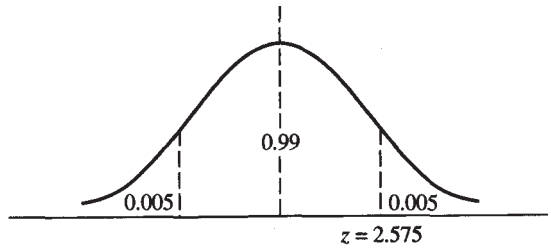


Figura 6.8. Partición de Z necesaria para obtener un intervalo de confianza de μ del 99%. El punto necesario es z = 2.575.

Dos observaciones son evidentes a partir de esta fórmula. La primera es que cada intervalo de confianza está centrado en $\bar{x}$. La segunda es que la amplitud del intervalo depende de tres factores: 1) la confianza deseada, 2) la desviación típica, 3) el tamaño muestral. En los Ejercicios 2 a 4 se le pedirá que investigue sobre el efecto de estos factores en la amplitud del intervalo.

Teorema central del límite

Hay una nueva puntualización que hacer. Los límites $\bar{X} \pm z(\sigma/\sqrt{n})$ se han deducido suponiendo que la variable X es normal. Si no se satisface esta condición, pueden emplearse los límites de confianza dados, mientras que la muestra no sea demasiado pequeña. Estudios experimentales han demostrado que, para muestras tan pequeñas como 25, los límites anteriores son generalmente satisfactorios, a pesar del hecho de ser aproximados. Ello se debe a un importante teorema, formulado por primera vez al principio del siglo XIX por Laplace y Gauss. Este teorema, conocido como teorema central del límite, se enuncia a continuación.

> **Teorema 6.2.2. Teorema central del límite.** Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n, de una distribución con media μ y varianza σ^2. Entonces, para n grande, $\bar{X}$ es aproximadamente normal con media μ y varianza σ^2/n. Además, para n grande, la variable aleatoria $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ es aproximadamente normal tipificada.

La demostración de este teorema está fuera de los objetivos de este texto. En esencia establece que, aunque la muestra proceda de una distribución no normal, $\bar{X}$ será aproximadamente normal, siempre y cuando el tamaño de la muestra no sea demasiado pequeño. En el Ejemplo 6.2.3. se ilustra el teorema.

> **Ejemplo 6.2.3.** Considérese la variable aleatoria X, número obtenido en el lanzamiento de un dado. La densidad para X está representada en la Tabla 6.3. Obsérvese que X es discreta y uniformemente distribuida. Esta distribución se aleja bastante de ser una distribución normal, la cual es continua y con forma de campana. Por consiguiente, si queremos construir intervalos de confianza para μ basado en los límites
>
> $$\bar{X} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$$
>
> debemos aplicar el teorema central del límite. Esto significa que el tamaño de la muestra utilizada no debe ser muy pequeño. Considérese un experimento en el que el dado se lanza $n = 25$ veces y se obtiene la media $\bar{x}$.

Tabla 6.3. Densidad para *X,* número obtenido en el lanzamiento de un solo dado

x	1	2	3	4	5	6
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

De acuerdo con el teorema, la variable aleatoria $\bar{X}$ tiene aproximadamente una distribución normal. Esto quiere decir que si repetimos el experimento un gran número de veces y representamos los valores $\bar{x}$ observados, el histograma obtenido debe tener forma aproximada de campana. Como $E[\bar{X}] = \mu$, valor medio de los valores de $\bar{x}$ observados debe estar cerca de la media verdadera de X que es 3.5, según se demostró en la Sección 4.2. De esta forma, se espera que los valores observados fluctúen alrededor del valor 3.5. El teorema central del límite supone que Var $\bar{X} = \sigma^2/n$, donde σ^2 es la verdadera varianza de X. Los métodos de la Sección 4.2 pueden ser utilizados para demostrar que Var $\bar{X} = \sigma^2 = 2.916$. Esto supone que $\sigma = \sqrt{2.916} \cong 1.708$. Así, en este caso, Var $\bar{X}$ =2.916/25 = 0.116.

Se utilizó el SAS para simular el experimento del lanzamiento de un dado 50 veces. Los resultados de esta simulación están representados en la Figura 6.9. En esta simulación, se permitió que el SAS eligiera sus propias clases. Obsérvese que el histograma no tiene forma perfecta de campana, pero sugiere considerablemente dicha forma. La media de los 50 valores de $\bar{x}$ es 3.51; este promedio se aproxima al valor teórico de 3.5. La varianza de los 50 valores de $\bar{x}$ es 0.143; esta varianza es, en cierto modo, mayor que el valor teórico de 0.116.

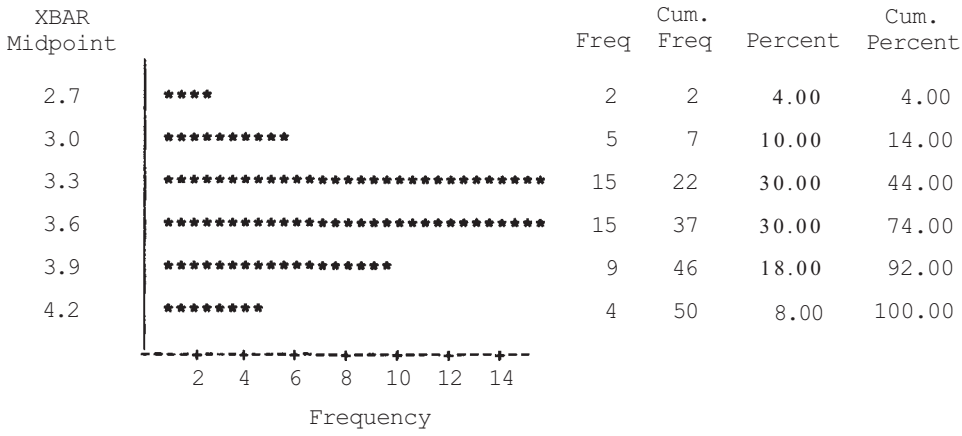
La Figura 6.10 representa algunas simulaciones adicionales. En la Figura 6.10a, los 50 valores de $\bar{x}$ están basados, cada uno, en muestras de tamaño $n = 5$; los de la Figura 6.10b, en muestras de tamaño $n = 10$ y los de la Figura 6.10c, en muestras de tamaño $n = 40$. Obsérvese que, a medida que n aumenta, los límites de las clases y el número de clases cambian, ya que estas características son función de los datos en sí mismos; la forma de campana se hace en cierta manera más pronunciada; la media de los valores $\bar{x}$ se aproxima al verdadero valor 3.5 y la varianza de los valores de $\bar{x}$ disminuye. Todas estas cosas son anticipadas por el teorema central del límite.

La simulación por ordenador puede ayudarnos a tener un mejor conocimiento de la noción de intervalo de confianza. Los 50 valores de $\bar{x}$ obtenidos en la simulación del Ejemplo 6.2.3, donde $n = 25$, están representados en la Tabla 6.4. Cada uno de los valores $\bar{x}$ ha sido sustituido en la fórmula

$$\bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{o en} \quad \bar{x} \pm 1.96\left(\frac{1.708}{\sqrt{25}}\right)$$

para formar un intervalo de confianza para μ del 95 %. Los límites numéricos de esos intervalos vienen dados en la tabla por L_1 y L_2. La teoría postula que alrededor de un 95 % de los intervalos construidos contendrá el verdadero valor de la media, 3.5; debido al azar, cerca de un 5 % no lo contendrá. En nuestro caso, es de esperar que un 5 % de los 50 intervalos (alrededor de 2.5) no incluyan a la media poblacional. Obsérvese que, de hecho, 2 de los 50 intervalos obtenidos la excluyen. Este resultado concuerda bastante con lo que predice la teoría.

Hay que tener en cuenta que en un estudio real solamente se calcula *un* intervalo de confianza. Este intervalo puede contener o no el verdadero valor de μ. Siempre se espera que el intervalo construido para nuestro estudio sea uno de los intervalos que incluya entre sus límites a μ, en vez de ser uno de los pocos que, por la aleatoriedad de la muestra, lo deja fuera. También hay que tener en cuenta que para utilizar la fórmula dada en esta sección, σ, el verdadero valor de la desviación estándar de X, *debe ser conocido.* Si no se conoce y debe estimarse de los datos, entonces la fórmula no es apropiada.



XBAR Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
2.7	****	2	2	4.00	4.00
3.0	**********	5	7	10.00	14.00
3.3	**********	15	22	30.00	44.00
3.6	**********	15	37	30.00	74.00
3.9	**********	9	46	18.00	92.00
4.2	********	4	50	8.00	100.00

2 4 6 8 10 12 14

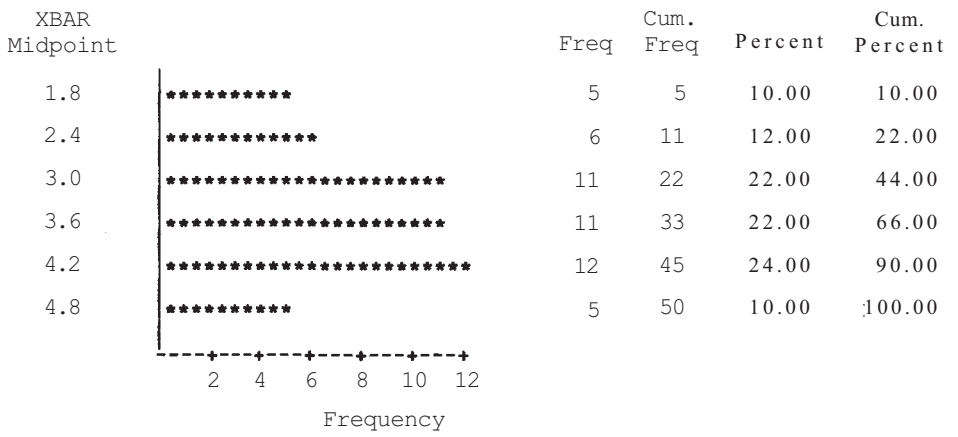
Frequency

DISTRIBUTION OF XBAR

Analysis Variable : XBAR

Mean	Variance	Std Dev	Std Error	N
3.510	0.143	0.378	0.053	**50**

Figura 6.9. Simulación basada en 50 muestras, cada una de tamaño 25.



XBAR Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
1.8	**********	5	5	10.00	10.00
2.4	**********	6	11	12.00	22.00
3.0	**********	11	22	22.00	44.00
3.6	**********	11	33	22.00	66.00
4.2	**********	12	45	24.00	90.00
4.8	**********	5	50	10.00	100.00

2 4 6 8 10 12

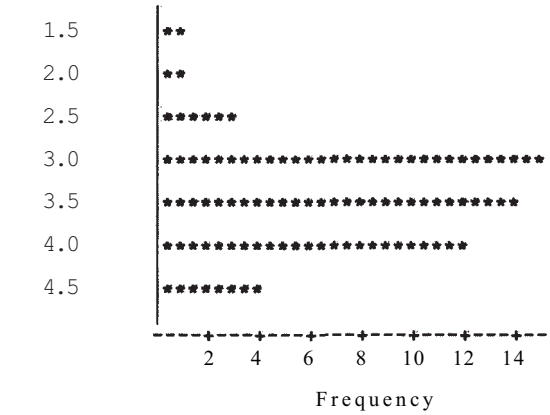
Frequency

DISTRIBUTION OF XBAR

Analysis Variable : XBAR

Mean	Variance	Std Dev	Std Error	N
3.412	0.693	0.832	0.118	50

Figura 6.10a. Simulación basada en 50 muestras, cada una de tamaño (a) $n = 5$, *(b)* $\pi = 10$, y (c) $n = 40$.



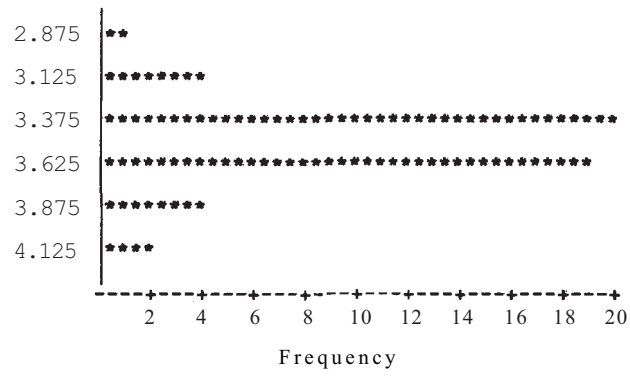
XBAR Midpoint	Freq	Cum. Freq	Percent	Cum. Percent
1.5	1	1	2.00	2.00
2.0	1	2	2.00	4.00
2.5	3	5	6.00	10.00
3.0	15	20	30.00	40.00
3.5	14	34	28.00	68.00
4.0	12	46	24.00	92.00
4.5	4	50	8.00	100.00

DISTRIBUTION OF XBAR

Analysis Variable : XBAR

Mean	Variance	Std Dev	Std Error	N
3.428	0.374	0.612	0.087	50

Figura 6.106



XBAR Midpoint	Freq	Cum. Freq	Percent	Cum. Percent
2.875	1	1	2.00	2.00
3.125	4	5	8.00	10.00
3.375	20	25	40.00	50.00
3.625	19	44	38.00	88.00
3.875	4	48	8.00	96.00
4.125	2	50	4.00	100.00

DISTRIBUTION OF XBAR

Analysis Variable : XBAR

Mean	Variance	Std Dev	Std Error	N
3.482	0.064	0.253	0.036	50

Figura 6.10c

Tabla 6.4. Intervalos de confianza del 95% de la media de X, puntuación obtenida en el lanzamiento de un dado

OBS	$\bar{X}$	L1	L2	Situación	OBS	$\bar{X}$	L1	L2	Situación
1	3.64	2.97046	4.30954	detenido	26	3.48	2.81046	4.14954	detenido
2	3.24	2.57046	3.90954	detenido	27	3.32	2.65046	3.98954	detenido
3	3.52	2.85046	4.18954	detenido	28	3.40	2.73046	4.06954	detenido
4	3.88	3.21046	4.54954	detenido	29	3.40	2.73046	4.06954	detenido
5	3.60	2.93046	4.26954	detenido	30	4.00	3.33046	4.66954	detenido
6	2.88	2.21046	3.54954	detenido	31	3.16	2.49046	3.82954	detenido
7	2.96	2.29046	3.62954	detenido	32	3.72	3.05046	4.38954	detenido
8	4.16	3.49046	4.82954	detenido	33	4.12	3.45046	4.78954	detenido
9	3.32	2.65046	3.98954	detenido	34	3.60	2.93046	4.26954	detenido
10	3.48	2.81046	4.14954	detenido	35	3.84	3.17046	4.50954	detenido
11	2.88	2.21046	3.54954	detenido	36	3.64	2.97046	4.30954	detenido
12	3.20	2.53046	3.86954	detenido	37	2.60	1.93046	3.26954	excluido
13	2.68	2.01046	3.34954	excluido	38	4.12	3.45046	4.78954	detenido
14	3.96	3.29046	4.62954	detenido	39	3.36	2.69046	4.02954	detenido
15	4.12	3.45046	4.78954	detenido	40	3.72	3.05046	4.38954	detenido
16	3.44	2.77046	4.10954	detenido	41	3.92	3.25046	4.58954	detenido
17	3.72	3.05046	4.38954	detenido	42	3.80	3.13046	4.46954	detenido
18	3.44	2.77046	4.10954	detenido	43	3.80	3.13046	4.46954	detenido
19	3.44	2.77046	4.10954	detenido	44	3.52	2.85046	4.18954	detenido
20	3.68	3.01046	4.34954	detenido	45	3.52	2.85046	4.18954	detenido
21	3.64	2.97046	4.30954	detenido	46	3.40	2.73046	4.06954	detenido
22	2.92	2.25046	3.58954	detenido	47	3.44	2.77046	4.10954	detenido
23	3.24	2.57046	3.90954	detenido	48	3.24	2.57046	3.90954	detenido
24	3.80	3.13046	4.46954	detenido	49	3.68	3.01046	4.34954	detenido
25	3.96	3.29046	4.62954	detenido	50	2.92	2.25046	3.58954	detenido

EJERCICIOS 6.2

1. Las observaciones siguientes corresponden a una muestra aleatoria de tamaño 9 de la variable aleatoria *X,* consumo de carbón por servicios eléctricos en millones de toneladas, en un año dado:

   406 395 400 450 390
   410 415 401 408

   Hallar una estimación puntual para μ, consumo medio de carbón para servicios eléctricos. ¿Es el valor que ha obtenido igual al consumo medio de carbón para electricidad en el año en cuestión? Explicarlo.
2. Hallar un intervalo de confianza del 90 % del número medio de microgramos de partículas de aire por metro cúbico, basándose en los datos del Ejemplo 6.2.1. ¿Es más grande o más pequeño este intervalo que el hallado anteriormente?
3. En general, ¿esperaría que un intervalo de confianza de μ del 90 % fuera más grande o más pequeño que el intervalo de confianza del 95 %, basado en la misma muestra?
4. El tamaño de la muestra desempeña un papel en la determinación de la longitud de un intervalo de confianza. Considerar dos intervalos de confianza del 95 % de μ, basándose en muestras de tamaño n_1 y n_2 extraídas de la misma población. Si $n_1 > n_2$, ¿qué intervalo de confianza será mayor?

5. Realizar diez veces el experimento descrito en el Ejemplo 6.2.3. ¿Varían los valores $\vec{x}$ alrededor del valor 3.5, tal como se esperaba? Promediar los diez valores de $\bar{x}$. ¿Se aproxima esta media bastante al valor 3.5? Calcular la varianza de los diez valores $\bar{x}$. ¿Ha obtenido un valor cercano a 0.116, tal como se esperaba? Calcule cada uno de sus diez valores $\bar{x}$ para formar un intervalo de confianza del 95 % para μ. ¿Alguno de sus intervalos no contienen la media real de 3.5?
6. *Error estándar o típico de la media.* Puesto que $\bar{X}$ es una variable aleatoria, tiene una media y una varianza. Sabemos que Var $\bar{X} = \sigma^2/n$. La desviación típica de $\bar{X}$ recibe el nombre de «error estándar de la media».
   a) ¿Cuál es la fórmula del error estándar de la media?
   b) ¿Qué influencia tiene el error estándar de la media en la forma del intervalo de confianza de μ?
7. *a)* Extraer cinco muestras aleatorias simples de los datos de DMAP de la Tabla V del Apéndice B, cada una de tamaño 10. Hallar x para cada muestra. La media para la población de la Tabla V es 6.254, con una varianza de 0.4829. ¿Sus valores x oscilan alrededor del valor 6.254 tal como se esperaba? ¿Se aproxima mucho la media de sus cinco valores de $\bar{x}$ a 6.254?
   b) Calcular Var $\bar{X}$. Hallar la varianza de sus cinco valores de $\bar{x}$. ¿Se aproxima su varianza al valor esperado?
   c) Calcular el error estándar teórico de la media.
   d) Utilizar el menor de los valores de $\bar{x}$ para construir un intervalo de confianza del 99 % para μ. ¿Contiene el intervalo numérico calculado el valor de μ, como se espera? Si no, ¿contendrá un intervalo de confianza del 90 % el valor de la media?
   e) Repetir el apartado *d,* utilizando ahora el mayor valor de $\bar{x}$.
8. La mayor parte de las especies de coniferas tiene piñas de polen y piñas de semilla. El polen desprendido por la piña macho es transportado por el viento hasta la piña hembra donde se fertilizan los huevos. Considerar la variable X, tiempo transcurrido entre la polinización y la fertilización. Supóngase que para los pinos, X está normalmente distribuida con una media de seis meses y una desviación típica de dos meses. Considerar el estadístico $\bar{X}$, basado en una muestra aleatoria de 25 piñas hembras. ¿Cuánto vale $E[\bar{X}]$? ¿Y Var $\bar{X}$? ¿Y el error estándar de la media?
9. La leucemia mieloblástica aguda es uno de los cánceres más mortales. Considérese la variable X, tiempo en meses que sobrevive un paciente después del diagnóstico inicial de la enfermedad. Suponga que X está normalmente distribuida, con una desviación típica de tres meses. Los estudios indican que $\mu = 13$ meses. Considerar la media muestral $\bar{X}$, basada en una muestra aleatoria de tamaño 16. Si la información anterior es correcta, ¿cuáles son los valores numéricos de $E[\bar{X}]$, Var $\bar{X}$ y el error estándar de la media?
10. Considerar 200 muestras de tamaño 25 extraídas de una población con media μ desconocida. Suponiendo que las 200 medias muestrales obtenidas se utilizan para construir 200 intervalos de confianza del 90 % para μ. ¿Aproximadamente cuántos de estos intervalos esperaría que no contuvieran a μ?
11. Se ha realizado un experimento lanzando 30 veces una moneda. Sea X_i, $i = 1, 2, 3, \ldots, 30$, definida por

$$X_i = \begin{cases} 1 & \text{si sale cara} \\ 0 & \text{otro caso} \end{cases}$$

   a) Utilizar los métodos de la Sección 4.2 para verificar que la media de $X_i = \frac{1}{2}$ y que su varianza es $\frac{1}{4}$. Así, $X_1, X_2, \ldots, X_{30}$ es una muestra aleatoria de una distribución con $\mu = \frac{1}{2}$ y varianza = $\frac{1}{4}$.

b) Considerar la variable aleatoria $\bar{X} = \sum X_i/30$. Argumentar que $\bar{X}$ da la proporción de lanzamientos que han salido cara.
c) Por el teorema central del límite, ¿cuál es la distribución de $\bar{X}$?
d) ¿Cuál es la distribución de la siguiente variable aleatoria?

$$\frac{\bar{X} - 0.5}{\sqrt{0.5(0.5)/30}}$$

6.3. INTERVALO DE CONFIANZA PARA LA MEDIA POBLACIONAL Y LA DISTRIBUCIÓN DE *T*

Obsérvese que para obtener una estimación puntual para la media poblacional μ, no es necesario conocer la varianza de la población; la media muestral $\bar{X}$ proporciona una estimación bastante buena de μ, independientemente del valor de σ^2. En todo caso, los límites del intervalo de confianza de μ dados en la Sección 6.2 son $\bar{X} \pm z\sigma/\sqrt{n}$. Suponer que, siendo desconocida la media de la población, se va a conocer la varianza es, en términos prácticos, una hipótesis nada realista. En la mayor parte de los casos, el estudio estadístico que interesa se hace por primera vez, por lo que no hay forma de conocer previamente cuál es la media o la varianza de la población en cuestión. Consideramos en esta sección un problema más real, hacer inferencias sobre una media poblacional cuando se considera que la varianza de la población es *desconocida.*

Para obtener una fórmula general para el intervalo de confianza de μ, bajo estas circunstancias, es natural empezar por considerar la variable aleatoria utilizada anteriormente, es decir,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Es necesario en tal caso abordar dos problemas:

1. El valor de σ no es conocido y debe ser estimado.
2. La distribución de la variable obtenida, reemplazando σ por un estimador no es conocida.

El primer problema se ha resuelto en la Sección 1.5. Recuérdese que primero hemos utilizado el estadístico

$$\frac{\sum (X_i - \bar{X})^2}{n}$$

como un estimador para σ^2. Se ha rechazado este estimador porque, en promedio, tiende a subestimar σ^2. Para obtener un estimador cuyos valores observados estén centrados en σ^2, dividimos por n - 1 y definimos la varianza muestral por

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

La desviación típica muestral viene dada por $S = \sqrt{S^2}$. Para resolver el segundo problema, consideremos de nuevo la variable aleatoria $(\bar{X} - \mu)/(\sigma/\sqrt{n})$, que es aproximadamente nor-

mal tipificada. Reemplacemos la desviación típica poblacional σ, que ahora consideramos desconocida, por su estimador 5, para obtener la variable aleatoria $(\bar{X} - \mu)/(S/\sqrt{n})$. Para utilizar esta variable aleatoria a fin de obtener una fórmula general de un intervalo de confianza de μ, cuando no se conoce σ^2, debe responderse a una pregunta, ¿cuál es su distribución? Se puede demostrar que la distribución no está lejos de ser normal tipificada. De hecho, si la variable base X es normal, entonces la variable aleatoria $(\bar{X} - \mu)/(S/\sqrt{n})$ sigue lo que se llama una *distribución de T con n* - 1 grados de libertad. Nos detendremos brevemente a considerar las características generales de las variables aleatorias T.

Propiedades de las variables aleatorias *T*

1. Hay un número infinito de variables aleatorias T, cada una identificada por un parámetro γ, llamados *grados de libertad.* El parámetro γ es siempre un entero positivo. La notación T_γ, designa una variable aleatoria T con γ grados de libertad.
2. Cada variable aleatoria T es continua.
3. La gráfica de la densidad de cada variable aleatoria T es una curva simétrica con forma de campana centrada en cero.
4. Decimos que γ es un parámetro de forma en el sentido de que cuando γ crece, la varianza de la variable aleatoria T decrece. De este modo, cuanto más grande es el número de grados de libertad, más apuntada se vuelve la curva en forma de campana asociada con la variable.
5. Cuando el número de grados de libertad crece, la curva T se aproxima a la curva normal típica (véase Fig. 6.11).

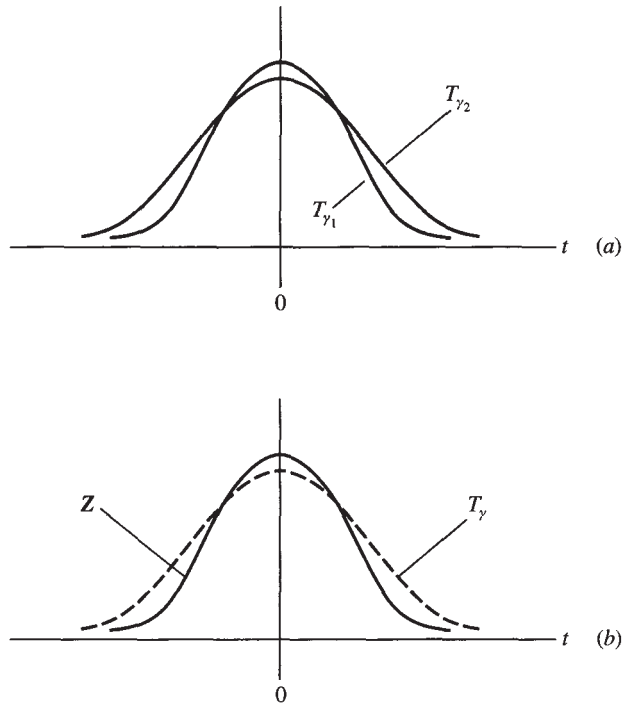


Figura 6.11. (a) Típica relación entre dos curvas Tcon $\gamma_1 > \gamma_2$. *(b)* Típica relación entre una curva T y la curva normal tipificada.

En la Tabla VI del Apéndice B se da un compendio parcial de valores de la función de distribución acumulada F para variables T. La Tabla VI está construida de manera que los grados de libertad aparecen en la cabecera de las filas, las probabilidades pertinentes aparecen en la cabecera de las columnas y los puntos asociados con estas probabilidades están situados en el cuerpo de la tabla.

A medida que aumentan los grados de libertad, disminuyen los cambios en los valores de la lista. Por esta razón, las últimas filas de la tabla nos permiten hallar la aproximación de los puntos para una serie de valores y, más que para un solo valor individual. La última fila, rotulada con el signo oo, se utiliza cuando y es mayor que 100. El Ejemplo 6.3.1 ilustra el uso de esta tabla.

Ejemplo 6.3.1. Considérese la variable aleatoria T_{10}.

a) De la Tabla VI del Apéndice B, $P[T_{10} \leqslant 1.372] = F(1.372) = 0.90$ (véase Fig. 6.12).

b) Por la simetría de la distribución T, el área a la izquierda de -1.372 es igual al área a la derecha de 1.372. De la Figura 6.12, $P[T_{10} \leqslant -1.372] = 0.10$.

c) Hallar el punto t tal que $P[-t \leqslant T_{10} \leqslant t] = 0.95$.. Puesto que queremos que el 95 % del área esté entre $-t$ y t, el 5 % del área estará por debajo de $-t$ o por encima de í. Este 5 % queda dividido en dos áreas del 2.5 %, cada una. Para hallar t, observe en la Figura 6.13 que el área a la izquierda de t es 0.95 + 0.025 = 0.975. El valor en la fila 10 y la columna 0.975 de la tabla T es 2.228. El punto í tal que $P[-t \leqslant T_{10} \leqslant t] = 0.95$ es $t = 2.228$ (véase Fig. 6.13).

La última fila de la Tabla VI del Apéndice B está rotulada con el signo ∞. Los puntos que aparecen en esa fila son puntos realmente asociados con la curva normal tipificada. Obsérvese que cuando y crece, los valores de cada columna de la Tabla VI se aproximan a los valores que aparecen en la última fila.

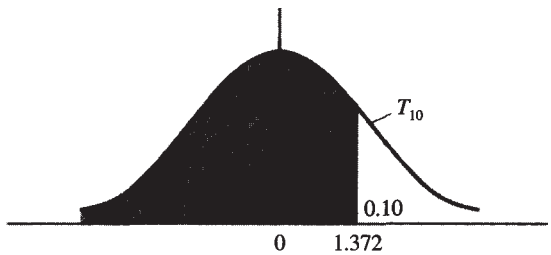


Figura 6.12. El área a la izquierda de 1.372 es 0.90; por lo tanto. $F(1.372) = P[T_{10} \leqslant 1.372] = 0.90$.

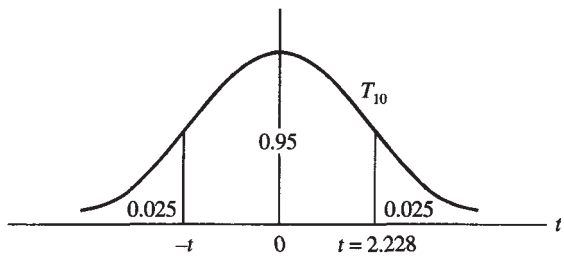


Figura 6.13. El área a la izquierda de 2.228 es 0.975; por ello, $P[-2.228 \leqslant T_{10} \leqslant 2.228] = 0.95$.

Ahora es fácil determinar la forma general para un intervalo de confianza de μ cuando no se conoce σ^2. Únicamente nos falta advertir que las dos variables aleatorias

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{y} \quad T_\nu = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tienen la misma estructura algebraica. Por lo tanto, el procedimiento algebraico presentado en el Ejemplo 6.2.2 se realizará exactamente como se indicó, reemplazando σ por S *y z* por jr. Estas sustituciones dan lugar al Teorema 6.3.1.

Teorema 6.3.1. Intervalo de confianza de μ cuando se ha estimado σ^2. Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria simple de tamaño n, de una distribución normal de media μ y varianza σ^2. Entonces un intervalo de confianza de μ viene dado por

$$\bar{X} \pm t \frac{S}{\sqrt{n}}$$

donde el punto t está basado en la distribución T_{n-1}.

Ejemplo 6.3.2. Las manadas de lobos son territoriales, con territorios de 130 km^2 o más. Se piensa que los aullidos de los lobos, que comunican información tanto de la situación como de la composición de la manada, están relacionados con la territorialidad. Se obtuvieron los siguientes valores observados para *X,* duración en minutos de una sesión de aullidos de una determinada manada sometida a estudio. Supongamos que *X* está normalmente distribuida.

1.0	**1.8**	**1.6**	**1.5**	**2.0**	**1.8**
1.2	**1.9**	**1.7**	**1.6**	**1.6**	
1.7	**1.5**	**1.4**	**1.4**	**1.4**	

Una estimación puntual para la duración media de una sesión de aullidos en esta manada es $\bar{x}$ = 1.57 minutos. La varianza muestral para estos datos es $s^2 = 0.066$. Una estimación para σ es

$$s = \sqrt{0.066} = 0.26 \text{ minutos}$$

Se puede hallar un intervalo de confianza de μ del 95 % considerando la partición representada en la Figura 6.14 de la curva T_{15}, obtenida de la Tabla VI del Apéndice B. Los límites pata un intervalo de confianza de μ del 95 % son:

$$\bar{x} \pm t \frac{s}{\sqrt{n}} = 1.57 \pm 2.131 \frac{0.26}{\sqrt{16}}$$
$$= 1.57 \pm 0.14$$

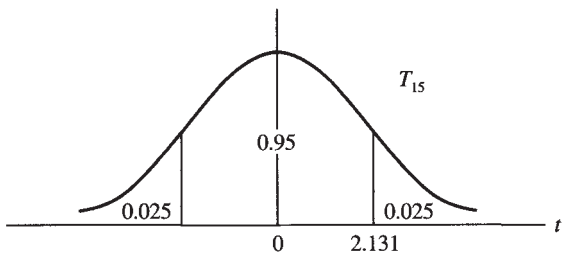


Figura 6.14. Partición de T_{15} para obtener un intervalo de confianza de μ del 95%.

Tenemos un 95 % de confianza de que la duración media de una sesión de aullidos para esa manada particular está entre 1.43 y 1.71 minutos.

Debemos llamar la atención sobre varios puntos. Primero, el número de grados de libertad implicados en la búsqueda de un intervalo de confianza de μ cuando no se conoce σ^2 es n - 1, el tamaño de la muestra menos 1. Para muestras grandes, este valor puede no aparecer en la Tabla VI del Apéndice B. En este caso, se utilizará la última línea de la tabla (∞) para hallar los puntos que nos interesan. Segundo, una vez más hemos partido del supuesto de normalidad. La variable aleatoria $(\overline{X} - \mu)/(s/\sqrt{n})$ sigue una distribución normal si la variable X está normalmente distribuida. La validez de esta suposición puede ser contrastada visualmente construyendo un histograma o un diagrama de tallos y hojas. Un método analítico para probar la normalidad se presenta en el Capítulo 13. Si la forma de X es aproximadamente acampanada, los métodos basados en la distribución T son generalmente buenos. Esto es cierto cuando X es, de hecho, discreta. Sin embargo, si hay motivos para sospechar que la variable en estudio tiene una distribución muy alejada de la normal, no deben utilizarse los procedimientos estadísticos basados en la distribución T. Se utilizarán, mejor, algunas técnicas de distribución libre que se tratan en el Capítulo 13.

EJERCICIOS 6.3

1. Los datos siguientes son las alturas, en metros, de veinte pinos blancos del este, *Pinus strobus.*

17.16	22.00	10.08	15.00
07.02	10.67	11.16	10.92
11.10	04.05	15.93	07.22
08.19	16.45	07.38	10.00
14.10	10.26	11.96	10.00

   a) Dibujar un diagrama de tallos y hojas adosado para estos datos. Use tallos de 0, 0, 1, 1 y 2, 2. Utilice el segundo dígito de cada número como hoja. ¿Tiene forma aproximada de campana?
   b) Estimar μ, σ y σ^2.
   (Basado en los datos coleccionados por Sabrina Norton. Departamento de Biología. Universidad de Radford, 1994.)
2. A continuación, se recoge una muestra aleatoria de 16 observaciones sobre una variable aleatoria X, número de libras de carne de vaca consumidas el pasado año, por persona, en los Estados Unidos:

118	110	117	120	119
115	112	112	113	122
125	130	115	118	123

   Utilizar estos datos para estimar μ, σ y σ^2.
3. Sea T_{15} una variable aleatoria T con 15 grados de libertad. Utilizar la Tabla VI del Apéndice B para hallar:
   a) El valor de t tal que $P[T \leqslant t] = 0.95$.
   b) El valor de t tal que $P[T \geqslant t] = 0.025$.
   c) El valor de t tal que $P[T \leqslant t] = 0.05$.

d) El valor de t tal que $P[T \geqslant t] = 0.975$.
e) $P[T_{15} \geqslant 2.602]$.
f) $P[T_{15} \leqslant -1.341]$.
g) $P[-1.753 \leqslant T_{15} \leqslant 1.753]$.
h) El valor de t tal que $P[-t \leqslant T_{15} \leqslant t] = 0.95$.
i) El valor de t tal que $P[-t \leqslant T_{15} \leqslant t] = 0.99$.

4. Los investigadores que estudian el fotoperíodo utilizan como planta experimental el cardillo. La variable observada fue *X,* número de horas de oscuridad ininterrumpida, por día, necesarias para producir floración. Se obtuvieron los siguientes datos:

15.0	13.0	15.1	16.0	13.5
15.5	13.2	14.9	14.7	

Estimar μ, σ y σ^2.

5. Se obtiene una muestra aleatoria de 1000 adultos aparentemente sanos con el fin de establecer un patrón con respecto al que se considerará una lectura «normal» de calcio. Se extrae una muestra de sangre de cada adulto. La variable estudiada es *X,* número de miligramos de calcio por decilitro de sangre. Se han obtenido una media muestral de 9.5 y una desviación típica muestral de 0.5. Supóngase que X presenta una distribución aproximadamente normal. Hallar un intervalo de confianza de μ del 95 %.

6. Se ha realizado un estudio del efecto del calor en la tasa de movilidad de los caracoles terrestres grandes. Los datos siguientes se han obtenido de *X,* distancia en centímetros recorrida por una muestra de 20 caracoles sometidos a una temperatura de 11 °C por encima de la temperatura ambiente (temperatura ambiente = 18 °C).

$$\bar{x} = 4.855 \qquad s = 0.7178$$

Construir un intervalo de confianza para la distancia media recorrida por los caracoles cuando la temperatura es de 24 °C. ¿Qué suposición debe hacerse acerca de la distribución de X? Si la distancia media recorrida a la temperatura ambiente es de 2.885 centímetros, ¿hay evidencia de que el calor tiende a aumentar la distancia media recorrida por los caracoles? Explicarlo. (Basado en un experimento realizado por Joseph Christian, Departamento de Biología, Universidad de Radford, 1996.)

7. Los datos siguientes han sido obtenidos de una muestra aleatoria simple de tamaño 30, de la distribución *X,* porcentaje de aumento del contenido de alcohol en la sangre de una persona, después de ingerir cuatro cervezas.

$$\bar{x} = 41.2 \qquad 5 = 2.1$$

Calcular un intervalo de confianza del 90 % para el porcentaje medio de alcohol en la sangre de una persona, después de tomar cuatro cervezas. Si se calcula un intervalo de confianza del 95 % para μ, ¿será de mayor o de menor amplitud que el anterior, del 90 %? ¿Creería la afirmación de que el incremento medio es menor del 35 %? Explicarlo.

8. En un estudio sobre utilización de agua en una ciudad pequeña, se extrae una muestra aleatoria de 25 casas. La variable de interés es *X,* número de galones de agua utilizados por día. Uno de los días de la semana, aleatoriamente elegido, se obtuvieron los siguientes valores. Supóngase que X es normal.

175	185	186	168	158
150	190	178	137	175
180	200	189	200	180
172	145	192	191	181
183	169	172	178	210

a) Dibujar un diagrama de tallos y hojas para estos datos. ¿Tiene forma aproximada de campana?

b) Estimar μ, σ y σ^2.

c) Hallar un intervalo de confianza para μ del 90%. El depósito de la ciudad es lo suficientemente grande para satisfacer una media de consumo de 160 galones por día. ¿Podría haber problema de escasez en la ciudad? Explicar la respuesta con base en el intervalo de confianza obtenido.

9. Las granjas de patos, alineadas en las orillas del Great South Bay, han contaminado seriamente el agua. Uno de dichos contaminantes es nitrógeno en forma de ácido úrico. La siguiente es una muestra aleatoria de nueve observaciones de *X,* número de libras de nitrógeno producidas por granja y día:

4.9	5.8	5.9	6.5	5.5
5.0	5.6	6.0	5.7	

Suponiendo que *X* es normal, construir un intervalo de confianza del 99 % para μ.

10. Se está poniendo a prueba un proceso que en fotobiología se denomina abscisión, con la esperanza de aumentar la cosecha de fruta (porcentaje de fruta mantenida en los árboles) en los naranjos de Florida. El proceso implica exponer los árboles a luz coloreada durante quince minutos cada noche. Se recolectó fruta de 10 árboles experimentales bajo condiciones normales primero, y después tras el nuevo tratamiento. Resultaron las siguientes observaciones para *X,* porcentaje en que se incrementó la recolección de fruta de un año al siguiente:

29	37	32	34	39
30	36	35	27	40

Considerando que *X* es normal, construir un intervalo de confianza del 95 % del incremento medio del porcentaje de fruta cosechada. El promotor del nuevo proceso pretende que éste incremente la recolección en un promedio del 40 %. ¿Cree usted en esta afirmación? Explicar la respuesta con base en el intervalo de confianza hallado.

11. Utilice los datos del Ejercicio 1 para calcular un intervalo de confianza del 90 % de la altura media del pino blanco del este, de la zona donde se ha obtenido la muestra.

12. El calibre de un árbol es el diámetro medido 6 pulgadas por encima del suelo. Se ha obtenido una muestra de 16 árboles entre 12 y 14 pies de altura cultivados en un vivero particular y se ha determinado el calibre de cada uno de ellos. Se obtuvieron los siguientes datos (en pulgadas):

2.3	1.9	1.7	2.1	1.5	1.8	1.8	1.1
2.1	1.5	2.0	1.6	,1.3	1.6	1.5	1.3

Hallar un intervalo de confianza del 95 % del calibre medio de los árboles cultivados en el vivero. Para estar seguros de que el tamaño medio de los árboles es proporcional a la resistencia del tronco, el calibre medio para árboles de este tamaño debería ser de 2

pulgadas. ¿Cumplen esta característica los árboles cultivados aquí? (Basado en datos encontrados en Gary Moll, «The Best Way to Plant Trees», *American Forests,* abril de 1990, págs. 61-64.)

6.4. INTRODUCCIÓN A LOS CONTRASTES DE HIPÓTESIS

En un problema de contraste de hipótesis, existe una teoría preconcebida relativa a la característica de la población sometida a estudio. Esto implica que en cualquier estudio estadístico haya dos teorías o hipótesis implícitas: la hipótesis que propone el experimentador y la negación de esta hipótesis. La primera, denotada por H_1, se llama *hipótesis alternativa* o *hipótesis de investigación,* mientras que la última, que se denota por H_0, se llama *hipótesis nula.* El propósito del experimento es decidir si la prueba tiende a apoyar o a refutar la hipótesis nula. Cuando formulamos H_0 y H_1 debemos tener en cuenta tres afirmaciones generales:

1. La hipótesis nula es la hipótesis de la «no diferencia». En términos prácticos esto quedaría recogido en la afirmación de que la igualdad forma parte de H_0.
2. Se ha de hacer todo lo que sea posible por detectar o fundamentar la hipótesis alternativa. Es decir, llamar H_1, a su teoría de investigación preconcebida.
3. Las hipótesis estadísticas se formulan siempre con la esperanza de que sea posible rechazar H_0 y, por lo tanto, aceptar H_1.

Ejemplo 6.4.1. Se está estudiando un nuevo fármaco para utilizarlo en el tratamiento del cáncer de piel. Se espera que sea eficaz en la mayoría de los pacientes sobre los que se aplica. La compañía que produce el fármaco quiere obtener alguna prueba estadística que apoye tal afirmación. Sea p la proporción de pacientes para los cuales el fármaco será eficaz. Puesto que nosotros hacemos lo posible para apoyar o descubrir la hipótesis alternativa, ésta, en este caso, es que $p > 0.5$. Ello implica automáticamente que la hipótesis nula es la negación de H_1, es decir, que $p < 0.5$. De modo que las dos hipótesis en juego son:

$$\underline{H_0}: p < 0.5$$
$$H_1: p > 0.5$$

Obsérvese que la afirmación de igualdad forma parte de la hipótesis nula. Obsérvese también que, desde el punto de vista del fabricante, se espera que H_0 sea rechazada, propiciando así que H_1 sea aceptada.

Una vez que se ha seleccionado una muestra y se han recogido los datos, debe tomarse una decisión. Ésta será rechazar H_0 o dejar de hacerlo. La decisión se toma observando el valor de algún estadístico cuya distribución de probabilidad, bajo la presunción de que H_0, es cierta, se conoce. A tal estadístico se le denomina *estadístico del contraste.* Si el valor del estadístico cuando H_0 es cierta difiere de lo esperado, rechazaremos la hipótesis nula en favor de la hipótesis alternativa; en caso contrario, no rechazaremos la hipótesis nula. Esto significa que al final de cualquier estudio de contraste de hipótesis nos veremos forzosamente en una de las situaciones siguientes:

1. Habremos rechazado H_0 siendo cierta; por tanto, habremos cometido lo que se conoce como un *error de tipo I.*
2. Habremos tomado la decisión correcta de rechazar H_0, siendo la alternativa H_1 cierta
3. Habremos dejado de rechazar H_0, siendo cierta la alternativa H_1; por tanto, habremos cometido lo que se conoce como un *error de tipo II.*
4. Habremos tomado la decisión correcta de dejar de rechazar H_0, siendo H_0 cierta.

Estas posibilidades se resumen en la Tabla 6.5

Tabla 6.5. Posibles formas de proceder en el contraste de hipótesis

	Estado real	
Decisión tomada	**H_0 cierta**	**H_1 cierta**
Rechazar H_0	Error de tipo I	Decisión correcta
Dejar de rechazar H_0	Decisión correcta	Error de tipo II

Ejemplo 6.4.2. Consideremos el problema de poner a prueba el fármaco del Ejemplo 6.4.1. Las hipótesis a contrastar son

$$H_0: p \leqslant 0.5$$
$$H_1: p > 0.5 \quad \text{(el fármaco es eficaz en la mayoría de los pacientes)}$$

Si se comete un error de tipo I, habremos rechazado H_0 siendo cierta. En términos prácticos, habremos concluido que el fármaco es eficaz para una mayoría de usuarios cuando no lo es. Este error puede conducir a la comercialización de un fármaco que es ineficaz para la mayoría de los pacientes. Se cometerá un error de tipo II si dejamos de rechazar H_0 cuando H_1 es cierta. En tal caso se concluirá que la tasa de eficacia del fármaco es del 50 % o menos cuando, de hecho, es efizaz para una mayoría de los pacientes sobre los que se apüca. Este error puede conducir a que no se comercialice un fármaco útil. Ambos errores son muy importantes. El error de tipo I es el que generalmente se considera más grave, ya que daría como resultado un retraso en el tratamiento apropiado de la enfermedad.

Obsérvese que es posible incurrir en error, con independencia de la decisión que se adopte. Cada vez que se rechaza H_0, puede producirse un error de tipo I; cada vez que H_0 no es rechazada, puede producirse un error de tipo II. No hay forma de evitar este dilema. El trabajo del profesional de la estadística es diseñar métodos para contrastar hipótesis que mantengan a un nivel razonablemente bajo las probabilidades de cometer cualquiera de los dos tipos de error. Hay dos formas de distinguir entre H_0 y H_1. Más adelante las mostraremos.

EJERCICIOS 6.4

1. En 1969 se calculó que, en Estados Unidos, un 8 % del contenido de las basuras caseras era metal. Debido al incremento de los procesos de reciclaje se espera que esta cifra se haya reducido. Se realiza un experimento para verificar esta suposición.
   a) Construir las hipótesis nula y alternativa apropiadas para el experimento. Obsérvese que el parámetro de interés es μ, la cantidad media de desperdicios caseros.
   b) Explicar en términos prácticos qué ocurre si se comete un error de tipo I.
   c) Explicar en términos prácticos qué ocurre si se comete un error de tipo II.
2. En 1974, el 38 % de las mujeres de Estados Unidos de edades comprendidas entre los 17 y los 24 años había fumado o aún lo hacía. Se teme que esta cifra haya aumentado. Se realiza un experimento para conseguir pruebas que apoyen este argumento.
   a) Construir las hipótesis nula y alternativa apropiadas para el experimento.
   b) Explicar en términos prácticos qué ocurre si se comete un error de tipo I.
   c) Explicar en términos prácticos qué ocurre si se comete un error de tipo **II.**

3. Antiguos estudios muestran que el germicida DDT puede acumularse en el cuerpo. En 1965 la concentración media de DDT en las partes grasas del cuerpo de las personas en Estados Unidos fue de 9 ppm. Se espera que, como resultado de estrictos controles, esta concentración haya decrecido.
   a) Construir las hipótesis nula y alternativa para documentar esta afirmación.
   b) Explicar en términos prácticos las consecuencias de cometer un error de tipo I y un error de tipo II.
4. El nivel medio de radiación latente en Estados Unidos es de 0.3 rem por año. Se teme que como resultado del aumento en el uso de materiales radiactivos esta cifra haya aumentado.
   a) Construir las hipótesis nula y alternativa apropiadas para documentar esta afirmación.
   b) Explicar en términos prácticos las consecuencias de cometer un error de tipo I y un error de tipo II.
5. En la ejecución de una prueba para detectar la presencia del virus del SIDA, se realizó el siguiente contraste:

   H_0: no se encuentra el virus
   H_1: existe el virus

   Explicar lo que sucedería si se introdujera un error de tipo I. ¿Qué sucederá si se produce un error de tipo II? ¿Cuál de estos dos errores cree que es más serio?
6. Se piensa que la mayoría de los fumadores comienza a fumar a los 18 años de edad. Se realiza un estudio con el fin de corroborar esta teoría y obtener fondos para una campana antitabaco. Sea p la proporción de fumadores que comienzan a fumar a los 18 años. Estamos contrastando

   H_0: $p \leqslant 0.5$
   H_1: $p > 0.5$ (la mayoría de fumadores comienza a los 18 años)

   Explicar las consecuencias económicas de cometer error de tipo I. Explicar las consecuencias que tiene para la salud cometer error de tipo II. ¿Cuál de estos dos errores piensa que es más importante?
7. Se han realizado muchos estudios para investigar el efecto de la «lluvia ácida» en el crecimiento de las plantas. En una zona particular bajo condiciones normales un esqueje de cerezo silvestre *(cornus florida)* crecerá una media de 8 pulgadas en el primer año. Se piensa que la lluvia ácida impedirá su crecimiento. ¿Cuáles son H_1 y H_0? ¿Qué ocurrirá si se comete error de tipo I?

6.5. CONTRASTES DE HIPÓTESIS DE LA MEDIA POBLACIONAL: CONTRASTE T

Consideremos ahora el problema de contrastar hipótesis concernientes a la media de una población. Esto implica que antes de llevar a cabo el experimento, uno tiene en mente un valor para μ. El propósito del experimento es obtener pruebas que contribuyan a defender o a refutar el valor en hipótesis. Obsérvese que, en cada uno de los tres ejemplos siguiente, nuestra teoría, la situación que deseamos detectar o defender, se llama H_1.

Ejemplo 6.5.1. El Departamento de Salud de Estados Unidos ha fijado en 70 el número medio de bacterias por centímetro cúbico de agua, que constituyen un nivel máximo aceptable para las aguas en que se practica la recogida de almejas. Un valor medio superior a 70 parece ser peligroso porque comer almejas recogidas en tales aguas puede causar la hepatitis. A fin de establecer un patrón gubernamental para las aguas, interesa contrastar

$$H_0: \mu \leqslant 70$$
$$H_1: \mu > 70$$

Ejemplo 6.5.2. Un estudio reciente del ecosistema en un bosque de hoja caduca indica que, en el bosque natural, el promedio neto de transformaciones del nitrógeno en nitrato presenta un incremento de 2 kg por hectárea y año. Los ingenieros de montes creen que una desfoliación de la maleza del bosque conduciría a un descenso de este valor. El contraste de hipótesis que interesa es

$$H_0: \mu \geqslant 2$$
$$H_1: \mu < 2$$

Ejemplo 6.5.3. El promedio total de proteínas en sangre en un adulto sano es de 7.25 g/dL. En un análisis de sangre, el técnico está contrastando

$$H_0: \mu = 7.25$$
$$H_1: \mu \neq 7.25$$

Como puede verse en los ejemplos, una hipótesis sobre μ puede adoptar una de tres formas diferentes. Sea μ_0, denominado *valor nulo,* el valor hipotético de la media poblacional. Las tres formas generales son

I	II	III
$H_0: \mu \leqslant \mu_0$	$H_0: \mu \geqslant \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu > \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu \neq \mu_0$

El estadístico utilizado para contrastar cada hipótesis es

$$T_{n-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Obsérvese que $\bar{X}$ estima la verdadera media poblacional. Si H_0 es cierta, $\bar{X}$ está estimando μ_0, por lo que la diferencia entre $\bar{X}$ y μ_0, debería ser pequeña. En cada caso, un valor pequeño del estadístico es una indicación de que *no* debería rechazarse H_0. En el caso I, la hipótesis de investigación es que $\mu > \mu_0$. Si esto es cierto, $\bar{X}$ está realmente estimando un valor medio mayor que el valor nulo. Deberíamos esperar que $\bar{X}$ fuera superior a μ_0, forzando a la diferencia $\bar{X} - \mu_0$ a ser positiva. Así, en el caso I, rechazamos H_0, en favor de H_1 para valores positivos grandes del estadístico. Dado que estos valores caen en el lado derecho de la recta real, el contraste descrito en el caso I se denomina *contraste con cola a la derecha.* Un argumento parecido nos conduce a la conclusión de que en el caso II rechazamos H_0 en favor de H_1 para valores negativos grandes del estadístico. El contraste se denomina *contraste con cola a la izquierda.* En el caso III, el contraste se denomina *contraste con dos colas.* Rechazamos la hipótesis nula para los valores inusualmente grandes o pequeños del estadístico. ¿Qué

queremos decir con valores «inusualmente» grandes o pequeños? Estos son valores del estadístico que se consideran raros. Sería sorprendente observar estos valores si H_0 fuera verdad. Sabemos que, si el valor nulo μ_0 es correcto, este estadístico sigue una distribución T con n - 1 grados de libertad. Este hecho puede utilizarse para comprobar si nuestro experimento ha producido o no un resultado inusual. Esto se hace calculando el valor P o valor de probabilidad del contraste donde por valor P entendemos lo siguiente:

> **Definición 6.5.1**. El *valor P* de un contraste es la probabilidad de que el estadístico asuma un valor tan extremo o más que el que observamos cuando suponemos que la hipótesis nula es cierta.

Hodges and Lehmann *(Basic Concepts of Probability and Statistics,* Holden-Day, San Francisco, 1970) describen el valor P «como el que da, en un solo número adecuado, una medición del grado de sorpresa que el experimento causaría en un partidario de la hipótesis nula». Para un contraste con cola a la derecha, el valor P es el área bajo la curva T_{n_1} hacia la derecha del valor observado del estadístico; para un contraste con cola a la izquierda, es el área de la izquierda. Los valores P para el contraste de dos colas se explicarán más adelante. Rechazamos H_0 si creemos que el valor P es demasiado pequeño para haberse producido razonablemente al azar. Los tres ejemplos siguientes ilustran el cálculo de los valores P.

> **Ejemplo 6.5.4.** Considérese el Ejemplo 6.5.1. Los estadísticos del Departamento de Salud de Estados Unidos se ocupan de vigilar las aguas en las que se realiza la pesca. El trabajo consiste en detectar cuándo el recuento medio de bacterias asciende por encima del nivel máximo de seguridad, cuyo valor se fijó en 70. Puesto que lo que interesa detectar se toma como hipótesis alternativa, estamos contrastando
>
> H_0: $\mu \leqslant 70$ (las aguas son seguras)
> H_1: $\mu > 70$ (las aguas no son seguras)
>
> Se extrae una muestra aleatoria de tamaño 9, y se determina el recuento X de bacterias para cada caso. El estadístico del contraste es
>
> $$T_{n-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$
>
> o
>
> $$T_8 = \frac{\bar{X} - 70}{S/3}$$
>
> Dado que $\bar{X}$ es un estimador razonable para la media, esperamos que el valor observado de X esté próximo a 70, si H_0 es cierta. Esto fuerza al numerador del estadístico $\bar{X}$ - 70 a ser pequeño, ya que es la causa de que el estadístico del contraste sea pequeño. Sin embargo, si H_1, es cierta, esperamos que $\bar{X}$ sea mayor que 70, forzando a $\bar{X}$ - 70 a ser grande y *positivo,* lo que produce un resultado grande y positivo para el estadístico. De aquí que, lógicamente, rechazaremos H_0 en favor de H_1, siempre que el valor observado del estadístico del contraste sea positivo y demasiado grande para que su aparición se deba al azar.
>
> Cuando se realizó el experimento se obtuvieron los siguientes valores:
>
> **69 74 75 70 72**
> **73 71 73 68**
>
> Para este conjunto de datos
>
> $$\bar{x} = 71.7 \qquad s = 2.3$$

El valor observado del estadístico del contraste es

$$\frac{\bar{x} - 70}{s/3} = \frac{71.7 - 70}{2.3/3} = 2.22$$

¿Es este valor inusualmente grande? Para responder a esta pregunta, calculamos el valor *P* del contraste. Por definición, el valor *P* es la probabilidad de observar un valor tan extremo o más extremo que aquel realmente obtenido. Para un contraste con cola a la derecha «más extremo» significa a la derecha del valor obtenido. Por lo que en este caso

$$P = P[T_8 > 2.22]$$

Esta probabilidad se ilustra en la Figura 6.15a. Para aproximar el valor *P* buscamos el número 2.22 en la fila 8 de la Tabla VI del Apéndice B.

Este valor cae entre los números 1.860 y 2.306. Como $P[T_8 \geq 1.860] = 0.05$ (Fig.6.15b)y $P[T_8 \geq 2.306] = 0.025$ (Fig. 6.15c), el valor *P* de nuestro contraste está entre 0.025 y 0.05. Esto se expresa escribiendo $0.025 < P < 0.05$. Ahora hagamos un juicio de valor. ¿Es pequeña esta probabilidad? Puesto que la mayoría puede considerar que una probabilidad de esta magnitud lo es, rechazamos H_0 y concluimos que las aguas no son buenas para la pesca.

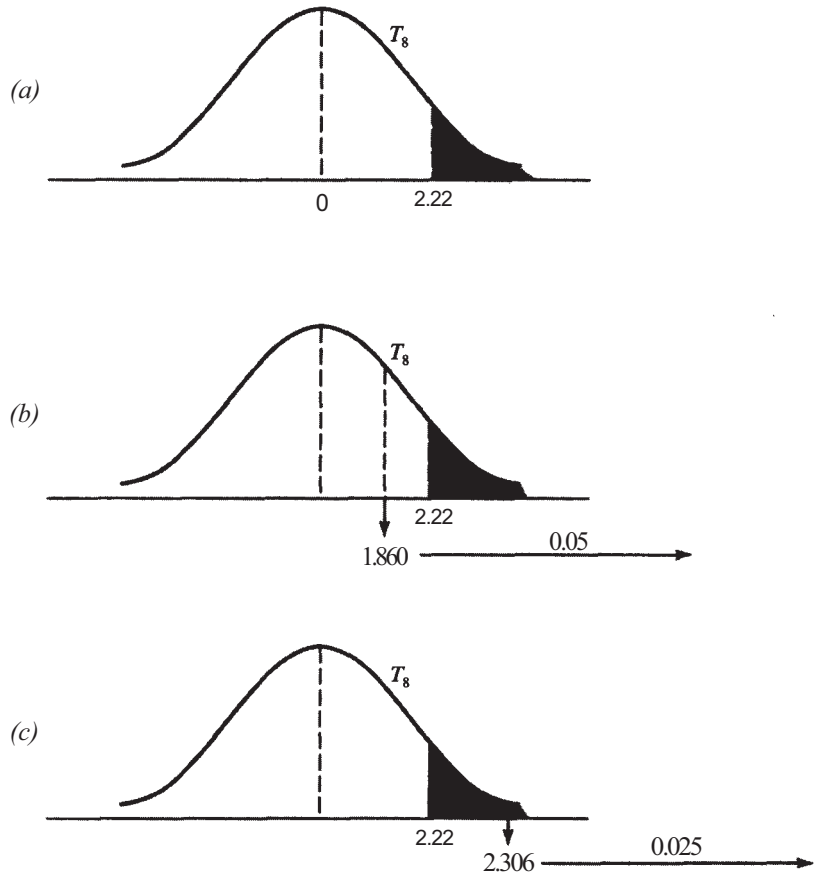


Figura 6.15. (a) $P = P[T_8 \geq 2.22]$, *(b)* $P[T_8 \geq 1.86] = 0.05$, (c)$P[T_8 \geq 2.306] = 0.025$.

No existen normas estrictas sobre cuán pequeño debe ser un valor P para que se rechace H_0. Recuérdese que siempre que rechacemos H_0 corremos el riesgo de cometer un error de tipo I. El valor P es una medida del grado de riesgo que corremos cuando hacemos nuestra hipótesis de investigación. Si las consecuencias de caer en dicho error son muy graves, entonces P debe ser muy pequeña antes de que nos decidamos a rechazar H_0. Si caer en un error de tipo I sólo produce un inconveniente, entonces H_0 puede rechazarse para valores P grandes. Una regla empírica aproximada es que H_0 no debería rechazarse para valores P que excedan de 0.10.

El siguiente ejemplo ilustra el cálculo del valor P para un contraste con cola a la izquierda.

Ejemplo 6.5.5. Se sometieron a contraste las hipótesis

$$H_0: \mu \geqslant 2 \qquad H_1: \mu < 2$$

del Ejemplo 6.5.2, arrancando la maleza en un área de 15 hectáreas de un bosque experimental. Se limpió el área para impedir el crecimiento. Después de un año, se determinó el cambio del nitrógeno a nitrato, por hectárea, analizando el agua de lluvia en 15 puntos dentro del bosque. Se obtuvieron los siguientes resultados.

$\bar{x}$ = -3 (pérdida *media* de 3 kg por hectárea)
s = 7.5 kg por hectárea

¿Es esto una prueba de que arrancar la maleza del bosque provoca un descenso en el cambio medio neto de nitrógeno a nitrato, por hectárea y año?

El valor del estadístico de contraste, de acuerdo con los datos, es

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{-3 - 2}{7.5/\sqrt{15}} = -2.58$$

Para un contraste con cola a la izquierda, un valor «más extremo» que el obtenido es un valor a la izquierda de -2.58. Así, el valor P viene dado por $P = P[T_{14} \leqslant -2.58]$. En la Figura 6.16a, se muestra el valor P para el contraste. En la Tabla VI, fila 14, vemos que el número 2.58 está situado entre 2.145 y 2.624. Puesto que la distribución T es simétrica, $P[T_{14} \geqslant 2.145] = P[T_{14} \leqslant -2.145] = 0.025$. Según la Figura 6.16b, es evidente que el valor P, área a la izquierda de -2.58, es menor que 0.025. Puesto que $P[T_{14} \leqslant -2.624] = 0.01$ y el área a la izquierda de -2.624 es menor que la representada por el valor P (véase Fig. 6.16c), sabemos que $P > 0.01$. Combinando estos dos resultados, vemos que $0.01 < P < 0.025$. Estas probabilidades son pequeñas. Rechazamos H_0 y concluimos que la retirada de la maleza del bosque dio como resultado una disminución de la concentración media del nitrógeno en forma de nitratos.

Se ha producido controversia sobre la forma correcta de calcular un valor P con dos colas. Si la distribución del estadístico, suponiendo que H_0 es cierta, es simétrica, entonces es razonable multiplicar por dos el valor P con una cola. Puesto que la distribución T es simétrica, éste es el procedimiento natural a utilizar en la realización de un contraste T. Si la distribución del estadístico de H_0 no es simétrica, el tema es más complejo. Los profesionales de la estadística han ofrecido diferentes sugerencias, pero todavía no se ha alcanzando ningún consenso. Sin embargo, el procedimiento más común consiste en considerar un valor P de dos colas como dos veces el valor P de una cola.

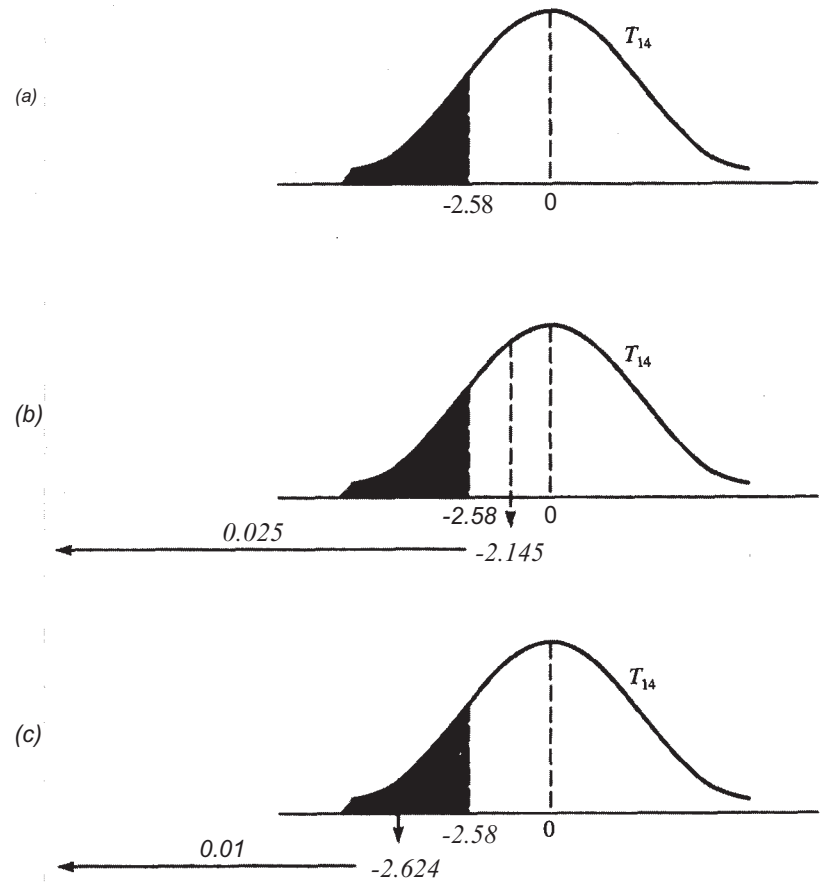


Figura 6.16. (a) $P = [T_{14} \leqslant -2.58]$, (b) $P < P[T_{14} \leqslant -2.145] = 0.025$, (c) $P > P[T_{14} \leqslant -2.624] = 0.01$.

Ejemplo 6.5.6. Se realiza una serie de ocho análisis de sangre sobre un determinado paciente, a lo largo de varios días. La variable considerada es el nivel total de proteínas. Puesto que el nivel de proteínas en sangre no puede ser ni demasiado grande ni demasiado pequeño, se desea detectar cualquiera de los dos hechos. De este modo, estamos contrastando

$$H_0: \mu = 7.25 \quad \textit{(normal para un adulto sano)}$$
$$H_1: \mu \neq 7.25$$

basada en una muestra de tamaño 8. El contraste tiene dos colas.

¿Qué conclusión puede extraerse a partir de las siguientes observaciones?

7.23	7.25	7.28	7.29
7.32	7.26	7.27	7.24

Para estos datos,

$$\bar{x} = 7.268 \qquad s = 0.029$$

El valor observado del estadístico es

$$\frac{\bar{x} - 7.25}{s/\sqrt{n}} = \frac{7.268 - 7.25}{0.029/\sqrt{8}} = 1.756$$

Puesto que el número 1.756 está entre los valores 1.415 y 1.895 en la fila 7 de la Tabla VI del Apéndice B, el valor *P* para un contraste con cola a la derecha estaría entre 0.05 y 0.10. Como estamos realizando un contraste con dos colas, se duplican estos valores para obtener $0.10 < P < 0.20$. Basándonos en este hecho, no podemos rechazar H_0. No tenemos pruebas suficientes para pretender que $\mu \neq 7.25$.

Valores alfa prefijados

Algunos profesionales de la estadística prefieren adoptar un método ligeramente diferente para contrastar hipótesis. En particular, consideran cuidadosamente las consecuencias de cometer un error de tipo I. Entonces, determinan el riesgo que estarían corriendo. Esta decisión se toma *antes* de reunir y analizar los datos. De esta forma, la probabilidad de cometer un error de tipo I se fija antes de realizar el experimento. Esta probabilidad prefijada de error se denomina *nivel de significación* o *tamaño del contraste.* Generalmente, se indica mediante la letra alfa (a). Para realizar este tipo de contraste, procederemos tal como se ha explicado en los ejemplos anteriores. Si el valor *P* hallado es menor o igual al nivel a prefijado, rechazamos H_0; de lo contrario, no rechazamos H_0.

Ejemplo 6.5.7. Cada especie de luciérnaga tiene un modo peculiar de centelleo. Para una determinada especie, consiste en un destello corto de luz seguido por un período de reposo que se piensa que tiene una duración media de menos de cuatro segundos. Deseamos contrastar

$$H_0: \mu \geqslant 4$$
$$H_1: \mu < 4$$

Si cometemos un error de tipo I, juzgaremos mal el período medio de reposo y probablemente escribiremos un resultado engañoso sobre la investigación. El error puede ser descubierto eventualmente por otros investigadores y acarrear problemas. Sin embargo, las consecuencias de nuestro error no son vitales. Podemos tolerar una probabilidad bastante grande de cometer un error de tipo I. Por esta razón, prefijemos a en el nivel 0.10. Se obtuvieron los siguientes datos acerca del período de reposo entre centelleos, para una muestra de 16 luciérnagas de esta especie

3.9	4.1	3.6	3.7	4.0	4.3
3.8	3.2	3.7	4.2	4.0	
3.5	3.5	3.8	3.4	3.6	

¿Apoya la evidencia el tiempo de reposo medio propuesto, de menos de 4 s?

Para este conjunto de datos,

$$\bar{x} = 3.77 \qquad \frac{\bar{x} - 4}{s/\sqrt{n}} = \frac{3.77 - 4}{0.30/4} = -3.06$$

$$s = 0.30$$

De la Tabla VI del Apéndice B, el valor *P* para el contraste, $P[T_{15} \leqslant -3.06]$, es menor que 0.005. Puesto que este valor *P* está por debajo del nivel *a.* prefijado de 0.10, rechazamos H_0 y concluimos que el tiempo medio de reposo es inferior a 4 s.

EJERCICIOS 6.5

1. Si contrastamos

$$H_0: \mu \leqslant 10$$
$$H_1: \mu > 10$$

basándonos en una muestra aleatoria de tamaño 20, el valor observado del estadístico T es 3. ¿Cuál es el valor P del contraste? ¿Cree que debería rechazarse H_0?

2. Si contrastamos

$$H_0: \mu \geqslant 5$$
$$H_1: \mu < 5$$

basándonos en una muestra aleatoria de tamaño 24, el valor observado para el estadístico T es -2.00. ¿Cuál es el valor P del contraste? ¿Cree que debería rechazarse H_0?

3. Al contrastar

$$H_0: \mu = 2$$
$$H_1: \mu \neq 2$$

basándonos en una muestra aleatoria de tamaño 16, el valor observado del estadístico T es 1.5. ¿Cuál es el valor P del contraste? ¿Cree que debería rechazarse H_0?

4. Uno de los efectos del DDT sobre los pájaros es la inhibición en la producción de la enzima anhidrasa carbónica. Esta enzima controla el metabolismo del calcio. Se cree que, como resultado final, las cáscaras de los huevos son mucho más finas y débiles de lo normal. Para comprobar esta teoría, se realizó un estudio alimentando a los gavilanes con una mezcla de 3 partes por millón (ppm) de dieldrina y 15 ppm de DDT. Se comparó el espesor de las cáscaras con el espesor medio conocido para pájaros no afectados por el DDT y se anotó el porcentaje de disminución de espesor en las cáscaras. Una muestra aleatoria de tamaño 16 dio lugar a una media muestral del porcentaje de disminución del 8 %, con una desviación típica muestral del 5 %. Utilizar esta información para contrastar

$$H_0: \mu \leqslant 0$$
$$H_1: \mu < 0 \text{ (el espesor de la cáscara disminuye)}$$

¿Cuál es el valor P aproximado para el contraste? ¿Cree el lector que la teoría ha sido estadísticamente corroborada? Explicar la respuesta basándose en el valor P.

5. La concentración media de dióxido de carbono en el aire es del 0.035 %. Se piensa que inmediatamente por encima de la superficie del suelo dicha concentración es mayor.
   d) Construir las hipótesis nula y alternativa que se requieren para conseguir un apoyo estadístico para este argumento.
   b) Se analizaron 144 muestras de aire seleccionadas aleatoriamente y tomadas a la distancia de un pie del suelo. Resultó una media muestral del 0.09 % y una desviación típica muestral del 0.25 % ¿Cuál es el valor P del contraste? ¿Piensa el lector que se ha comprobado estadísticamente el argumento establecido?
6. Con frecuencia es difícil cultivar plantas cerca de los lugares donde hay nogales negros porque las raíces de estos árboles lixivian una toxina llamada juglona en el suelo de los

alrededores. Se cree que el pH promedio del suelo próximo a las raíces es 0.7 puntos superior al del situado más lejos de las raíces. Para corroborar esta teoría, se ha seleccionado una muestra de 25 de estos árboles y se ha determinado el pH cerca y lejos de sus raíces. La variable aleatoria X es la diferencia de pH obtenida por sustracción en el orden: pH cerca de las raíces-pH lejos de las raíces. Ha resultado una media muestral $\bar{x} = 0.8$, con una desviación típica muestral $s = 0.3$. ¿Apoyan estos datos la teoría de la investigación de que $\mu > 0.7$? Explicarlo. Si se hace la afirmación de la investigación, ¿cuál es la probabil - dad de error? (Basado en información Diane Relf, *Plants at War,* Virginia Cooperative Extension Service, NA: 1-NA, 1983.)

7. Se ha realizado un experimento para estudiar el efecto del ejercicio físico en la reducción del nivel de colesterol en pacientes ligeramente obesos, con riesgo de infarto de miocardio. Ochenta pacientes son sometidos a un régimen específico de ejercicios mientras mantienen una dieta normal. Transcurridas cuatro semanas, se anotará la variación del nivel de colesterol. Se piensa que el programa reducirá la media del nivel de colesterol en más de 25 puntos. Al final del estudio, los datos obtenidos son:

$$\bar{x} = 27 \qquad s = 18$$

¿Corroboran estos datos la teoría de la investigación? Si se hace la afirmación de la investigación, ¿cuál es la probabilidad de error?

8. Los murciélagos al volar localizan un objeto sólido emitiendo agudos chillidos y escachando el eco. Se piensa que el alcance medio efectivo máximo para este sistema de localización por eco es de más de 6 metros. Para confirmar esta hipótesis, se seleccionó una muestra aleatoria de 16 murciélagos. Cada murciélago fue soltado en un área grande y cercada, que contenía un obstáculo sólo. Se anotó la distancia del objeto a la que se observó que viraba el murciélago. Se repitió el experimento varias veces para cada murciélago, y se determinó para cada uno la distancia media del viraje. Se obtuvieron las siguientes observaciones:

6.2	6.8	6.1	5.7	6.1	6.3	5.8	6.3
5.9	6.3	6.4	6.0	6.3	6.2	5.9	6.1

Hallar el valor P para este conjunto de datos. ¿Qué conclusión práctica puede extraerse de ellos? ¿Qué tipo de error se puede estar cometiendo?

9. Considérense las hipótesis dadas en los Ejercicios 1 a 3. En cada caso, ¿puede ser rechazada H_0, al nivel $\alpha = 0.05$? Explicarlo.

10. El nivel máximo aceptable de exposición a radiación de microondas en Estados Unidos se ha establecido en una media de 10 microvatios por centímetro cuadrado. Se teme que un gran transmisor de televisión pueda contaminar el aire del entorno inmediato, elevando el nivel de radiación de microondas por encima del límite de seguridad.
    a) Construir las hipótesis nula y alternativa necesarias para obtener pruebas que confirmen este supuesto.
    b) La siguiente es una muestra aleatoria de nueve observaciones sobre X, número de microvatios por centímetro cuadrado, tomadas en lugares próximos al transmisor:

9	11	14	10	10
12	13	8	12	

Hallar el valor P del contraste.

¿Puede rechazarse H_0 a un nivel a = 0.1? ¿Qué conclusión práctica puede extraerse? ¿Qué tipo de error puede cometerse?

11. Normalmente las hojas de *Mimosa pudica* son horizontales. Si se toca ligeramente una de ellas, las hojas se pliegan. Se afirma que el tiempo medio desde el contacto hasta el cierre completo es 2.5 s. Se realiza un experimento para comprobar este valor.
    a) Construir la hipótesis apropiada con dos colas.
    b) Se obtuvieron las siguientes observaciones sobre una variable *X,* tiempo transcurrido entre el contacto y el cierre completo:

    3.0 2.9 2.8 2.7 2.6
    2.4 2.5 2.4 2.6 2.7

    Hallar el valor *P* del contraste.
    ¿Puede rechazarse H_0 con un nivel $\alpha = 0.10$? ¿A qué tipo de error nos arriesgamos?
12. El número medio de días de clínica requeridos por pacientes de edad, antes de que pudieran disfrutar de los cuidados del hogar, era de 17 días. Se espera que un nuevo programa reduzca esta cifra. ¿Prueban los datos siguientes la hipótesis de investigación al nivel $\alpha = 0.05$? Explicarlo basándose en el valor *P* del contraste.

    3 5 12 7 22 6 2
    18 9 8 20 15 3 36
    38 43

    (Basado en información hallada en: Julianne Oktay y Patricia Volland, «Post-Hospital Support Program for the Frail Elderly and Their Caregivers», *American Journal of Public Health,* enero 1990, págs. 29-45.)
13. El número medio de ataques de angina de pecho por semana entre los pacientes es de 1.3. Se está probando un nuevo medicamento y se espera que reduzca esta cifra. Los datos se obtienen mediante la observación de una muestra de 20 pacientes, que están utilizando el nuevo medicamento.

    1 3 0 1 1 1 0 2 2 0
    0 1 0 0 0 1 1 1 1 0

    ¿Puede rechazarse la hipótesis de investigación al nivel 0.01? Explicarlo, basándose en el valor *P* del contraste. (Basado en la información hallada en un anuncio en el *American Journal of Nursing,* septiembre de 1990, pág. 13.)

6.6. TAMAÑO MUESTRAL: INTERVALOS DE CONFIANZA Y POTENCIA (OPCIONAL)

En el diseño de un experimento para estudiar la media, una de las primeras preguntas a la que debe responderse es ¿cuál debe ser el tamaño mínimo de la muestra para cumplir los objetivos del estudio? Esta pregunta es bastante fácil de responder si el objetivo es la estimación. Sin embargo, si la intención del estudio es contrastar hipótesis, la situación es más compleja. En esta sección, consideramos el problema de determinar un tamaño de muestra apropiado en cada uno de estos casos.

Tamaño de la muestra: estimación

Para que un intervalo de confianza sea útil, debe ser lo suficientemente pequeño como para indicar el valor de μ razonablemente bien con un grado de confianza alto. Si un experimento

no está planificado o la planificación es deficiente, existen diversas posibilidades de que el intervalo de confianza resultante sea demasiado grande para ser útil al investigador. Por ejemplo, un experimento que nos permite concluir que «tenemos una confianza del 95 % de que el verdadero peso medio de un niño, en el momento del nacimiento, está entre 2 y 20 libras (900 y 9000 g)» no tiene mucho valor. ¿Qué factores afectan a la longitud de un intervalo de confianza? Consideremos el intervalo de confianza de la media representado en la Figura 6.17. La longitud del intervalo es

$$2t\frac{s}{\sqrt{n}}$$

Si se utiliza $\bar{x}$ como estimación de μ, ambos valores diferirán, a lo sumo, en la mitad de la longitud del intervalo. Esto es, la distancia máxima entre $\bar{x}$ y μ deberá ser, como mucho, *d,* donde

$$d = ts/\sqrt{n}$$

La longitud se ve afectada por tres factores. Estos son:

1. La confianza deseada, que controla el valor de *t.*
2. La variabilidad de la muestra, que se mide por *s.*
3. El tamaño de la muestra.

Para garantizar que el intervalo es lo suficientemente pequeño como para ser informativo, debemos especificar la longitud y la confianza deseadas. A continuación, elegimos el tamaño de la muestra de forma que puedan cumplirse estas especificaciones. Un ejemplo ilustrará el modo de hacerlo.

Ejemplo 6.6.1. Supongamos que debe realizarse un estudio para estimar el peso medio en el momento del nacimiento de niños cuyas madres son adictas a la cocaína. ¿Qué tamaño debe tener la muestra para estimar esta media con un margen de $\frac{1}{2}$ libra, con una confianza del 95 %? La frase «con un margen de $\frac{1}{2}$ libra» significa que la diferencia entre la media estimada de $\bar{x}$ y la media verdadera de μ es, como máximo, de $\frac{1}{2}$ libra. Esta situación se ilustra en la Figura 6.18. Dado que μ debería pertenecer al intervalo de confianza, estaría dentro de un margen de $\frac{1}{2}$ le libra, si el intervalo es de una longitud de 1 libra. Consideremos la ecuación

$$d = t\frac{s}{\sqrt{n}}$$

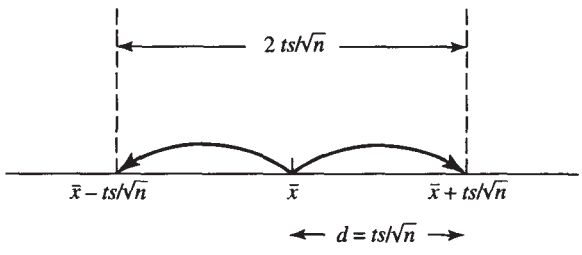


Figura 6.17. Un intervalo de confianza de μ viene dado por $\bar{x} \pm ts/\sqrt{n}$. La longitud del intervalo es $2ts/\sqrt{n}$. El valor de d es la mitad de la longitud, $ts/\sqrt{n}$.

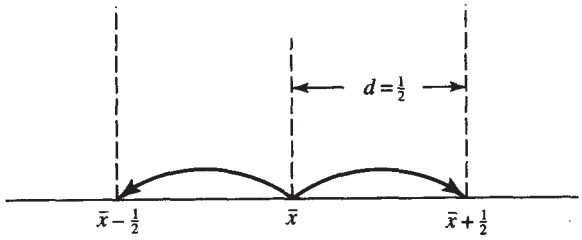


Figura 6.18. Puesto que μ debe pertenecer ai intervalo, estará dentro de \ libra de $\bar{x}$ si el intervalo tiene una longitud de 1.

Resolviéndola para *n,* obtenemos

$$n = \frac{t^2 s^2}{d^2}$$

En este caso, donde $d = \frac{1}{2}$,

$$n = \frac{t^2 s^2}{(\frac{1}{2})^2}$$

¿Qué punto *t* deberá utilizarse? Recuérdese que el punto utilizado en la construcción de un intervalo de confianza depende de la confianza y del tamaño de la muestra. Aquí conocemos la confianza deseada, el 95 %, pero estamos intentando hallar el tamaño de la muestra. Para resolver este problema, recordemos que, para muestras grandes, los puntos *t* pueden aproximarse mediante puntos z. Puesto que el valor *t* correcto no puede determinarse hasta que se conozca el tamaño de la muestra, aproximemos su valor utilizando el punto *z* asociado al intervalo de confianza del 95 *%.* Este punto, 1.96, se muestra en la Figura 6.19. En este punto sabemos que

$$n = \frac{(1.96)^2 s^2}{(\frac{1}{2})^2}$$

Tenemos todavía un problema por resolver. ¿Cuánto vale s^2? ¿Qué valor deberá utilizarse para estimar la varianza poblacional? Hay varias formas de responder a esta pregunta.

1. Utilizar el estimador de un estudio anterior.
2. Efectuar un pequeño estudio preliminar o piloto y utilizar el valor de s^2 hallado para ayudar a planificar el experimento mayor.
3. Recordar que la ley de probabilidad normal garantiza que *X* estará el 95 % de las veces dentro de 2 veces la desviación típica de su media. Por lo tanto, el rango de *X* es aproximadamente de 4 veces la desviación típica. Podemos utilizar el rango dividido por 4 para aproximar *s.*

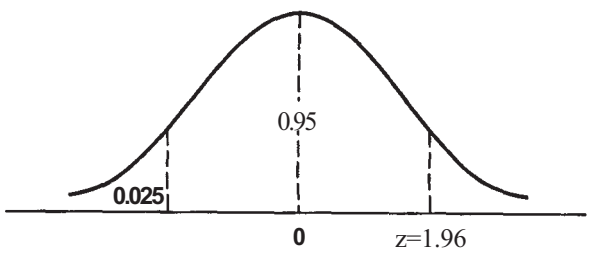


Figura 6.19. Partición de la distribución Z necesaria para construir un intervalo de confianza del 95%.

Aquí podemos suponer con seguridad que el peso en el momento del nacimiento de estos niños probablemente estará entre 2 y 12 libras. Por lo tanto, el rango es 10, y s es aproximadamente $\frac{10}{4} = 2.5$. El tamaño muestral requerido es

$$n = \frac{(1.96)^2(2.5)^2}{(\frac{1}{2})^2} = 96.04$$

Puesto que no podemos muestrear 0.04 partes de un niño, tomaremos una muestra de tamaño 97.

En general, el tamaño de muestra requerido para estimar μ dentro de un grado establecido de precisión con una determinada confianza es

$$n = \frac{Z^2 s^2}{d^2}$$

donde z es un punto de la distribución Z cuyo valor depende de la confianza deseada; s^2 es un estimador para la varianza poblacional y d es la mitad de la longitud del intervalo final deseado.

Ejemplo 6.6.2. ¿Qué tamaño debe tener una muestra para estimar el diámetro medio a la altura del pecho de una plantación de pinos maduros, dentro de un margen de 6 pulgadas, con una confianza del 90 %? Aquí, la longitud del intervalo deseado es de 1 pie y $d = \frac{1}{2}$. En la Figura 6.20, vemos que el punto z asociado con un intervalo de confianza del 90% es 1.645. Los guardabosques saben por experiencia que el diámetro a la altura del pecho está entre 3 y 8 pies. Por lo tanto, el rango de X es 5 y un estimador para la desviación típica de la población es $\frac{5}{4} = 1.25$. El tamaño de la muestra requerido es

$$n = \frac{(1.645)^2(1.25)^2}{(\frac{1}{2})^2} = 16.9$$

Tomaremos 17 árboles para la muestra.

Tamaño de la muestra: contrastes de hipótesis

El problema de determinar el tamaño adecuado de una muestra para contrastar una hipótesis sobre el valor de la media poblacional requiere especial atención. El Ejemplo 6.6.3 presenta un caso típico en el que aparece este problema.

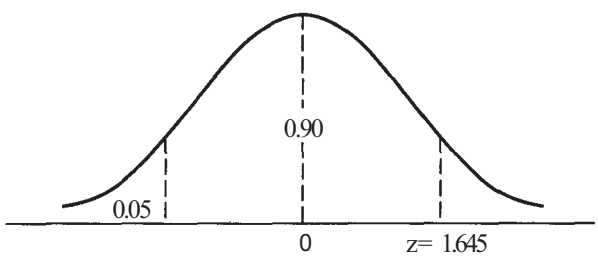


Figura 6.20. Partición de la distribución Z necesaria para obtener un intervalo de confianza del 90%.

Ejemplo 6.6.3. Un investigador está estudiando un fármaco que se utilizará para reducir el nivel de colesterol en varones adultos de 30 años o más. La variable aleatoria considerada es el cambio en el nivel de colesterol antes y después de utilizar el fármaco restando según el orden de lectura, haciendo «antes» menos «después». Si el fármaco es eficaz, la diferencia media de las lecturas deberá ser positiva. La hipótesis a comprobar es

H_0: $\mu \leqslant 0$ (el fármaco no es efectivo)
H_1: $\mu > 0$ (el fármaco es efectivo en la reducción del nivel de colesterol)

¿Qué tamaño deberá tener la muestra a utilizar para realizar el contraste?

Recuérdese que en el contraste de hipótesis son posibles dos errores. Podríamos rechazar H_0 cuando H_0 es verdadera, cometiendo así un error de tipo I. En nuestro ejemplo, comercializaríamos un fármaco no eficaz. La probabilidad de cometer este error puede controlarse acordando prefijar a en algún valor aceptablemente bajo, tal como se ha explicado en la Sección 6.5. Se produce un error de tipo II cuando no rechazamos H_0 aunque sepamos que la hipótesis del investigador es verdadera. En nuestro ejemplo, un error de tipo II no nos permite reconocer la eficacia del fármaco que se está estudiando. La probabilidad de cometer un error de tipo II se denomina beta *(ß)*. Es evidente que en cualquier contraste de hipótesis, idealmente, *ß* debería ser pequeño. Si nuestra teoría de investigación es correcta, queremos saberlo. La probabilidad de que seamos capaces de detectar una teoría de investigación verdadera se denomina *potencia.* Para el cálculo, potencia $= 1 - \beta$. Estas probabilidades se visualizan en la Tabla 6.6. Puesto que la potencia es la probabilidad de que podamos comprobar nuestra investigación cuando tenemos razón, los experimentos se diseñarán de forma que la potencia sea alta.

No es posible obtener algo por nada, incluso en un diseño estadístico. Si a está prefijada en un valor extremadamente pequeño para minimizar la probabilidad de cometer un error de tipo I, se paga un precio. Alfa y beta están relacionadas entre sí, de forma que cuando α disminuye, β aumenta. Esto, a su vez, implica que disminuye la potencia. Para mejorar la potencia, debe ajustarse el tamaño de la muestra elegida de acuerdo con el nivel α elegido.

Supongamos que la hipótesis nula es incorrecta y $\mu \neq \mu_0$. El sentido común indica que es más fácil detectar una gran diferencia entre μ_0 y μ, que una muy pequeña. En nuestro ejemplo es más fácil detectar un cambio medio en el nivel de colesterol de 20 mg/dL que detectar un cambio de 1 mg/dL. Un experimento diseñado para detectar grandes diferencias no requerirá una muestra tan grande como otro diseñado para detectar pequeñas diferencias.

La variabilidad también desempeña un papel importante en la elección del tamaño de la muestra. Es más fácil extraer conclusiones precisas en relación con las poblaciones estables,

Tabla 6.6. En cualquier caso de contraste de hipótesis pueden surgir cuatro situaciones. Un estudio ideal es aquel en el que a y ß son pequeños y la potencia es alta

	Estado real	
Decisión	H_0 cierta	H_1 cierta
Rechazar $\mathbf{H_0}$	Error de tipo I (probabilidad = α)	Decisión correcta (probabilidad = potencia)
Dejar de rechazar $\mathbf{H_0}$	Decisión correcta	Error de tipo II (probabilidad = β)

que hacerlo con poblaciones que presentan un alto grado de variabilidad. Los tamaños de las muestras pueden ser más pequeños en el primer caso que en el segundo.

Para elegir un tamaño de muestra apropiado el investigador debe considerar tres cosas.

1. El tamaño deseado del contraste (α).
2. La magnitud que se considera importante de la diferencia entre μ_0 y μ.
3. La varianza de la población muestreada.

Se han construido tablas para determinar el tamaño de muestra adecuado para elegir entre μ_0 y algún valor μ_1 alternativo. La Tabla VII del Apéndice B es una de estas tablas. Para utilizarla, el sujeto experto en la materia, el investigador, debe especificar μ_1 el valor alternativo de μ que reviste importancia práctica. También debe especificar α y β, fijando así la potencia deseada. Además, debe obtenerse un estimador de σ, la desviación típica de la población. El Ejemplo 6.6.4 ilustra el uso de la Tabla VIL

Ejemplo 6.6.4. Contrastemos H_0: $\mu > 10$ contra H_1: $\mu > 10$. Supongamos que es importante detectar una media de 12. Queremos diseñar un experimento de forma que la potencia para detectar una media de 12 sea del 90 %. Supongamos que decidimos prefijar α en 0.05 y que un estimador para la varianza de la población basada en un estudio piloto es $s^2 = 16$. Para determinar el tamaño adecuado de la muestra introducimos la Tabla VII con α, β y Δ, donde $\Delta = |\mu_0 - \mu_1| / \hat{\sigma}$. El símbolo $\hat{\sigma}$ representa una estimación de la desviación típica de la población. En este caso, $\alpha = 0.05$, $\beta = 0.10$ y $\Delta = 110 - 121/4 = 0.5$. Estamos realizando un contraste con cola a la derecha (un solo lado). La Tabla 6.7 muestra una parte de la Tabla VII El tamaño de la muestra, 36, se halla localizando en la tabla los valores deseados de α, β y Δ.

Ejemplo 6.6.5. Considérese el estudio descrito en el Ejemplo 6.6.3. Puesto que no queremos arriesgarnos a comercializar un fármaco ineficaz, fijemos α en 0.01. Desde un punto de vista práctico, se decide que no merece la pena comercializar el fármaco a menos que el descenso de la media en el nivel de colesterol sea al menos de 20 mg/dL. Dado que un cambio de esta magnitud sería una contribución importante al tratamiento del colesterol alto, desearíamos ser capaces de detectar una diferencia de este tamaño. Diseñemos el estudio de forma que nuestra potencia sea de 0.95 ($\beta = 0.05$). Se sabe que la desviación típica de las lecturas de este grupo de edad es 30. (Véase el Ejercicio 5 de la Sección 5.4.) Para determinar el tamaño de la muestra correcto buscamos en la Tabla VII con $\alpha = 0.01$, $\beta = 0.05$ y $\Delta = 10 - 201/30 = 0.67$. Obsérvese que $\Delta = 0.67$ no aparece en la lista. Puesto que es más seguro tener una muestra bastante grande en lugar de tener una demasiado pequeña, adoptamos un criterio conservador utilizando $n = 41$ ($\Delta = 0.65$) en lugar de $n = 35$ ($\Delta = 0.70$).

El mensaje de esta subsección es simple. Los experimentos no planificados a menudo son experimentos mal realizados. Los tamaños muestrales deben elegirse cuidadosamente y, en general, las muestras pequeñas producen una potencia pequeña. Desde el principio, los experimentos basados en muestras pequeñas están generalmente condenados al fracaso, a menos que la diferencia entre μ_0 y μ sea muy grande. Las muestras pequeñas simplemente no pueden detectar las reducidas, pero a veces importantes, diferencias entre μ_0 y μ.

EJERCICIOS 6.6

1. Considérense los datos del Ejemplo 6.3.2 como estudio piloto. ¿Cuál debe ser el tamaño de una muestra para estimar la duración media de una sesión de aullidos de una manada de lobos, con un margen de 0.1 minutos y una confianza del 99 %?
2. ¿Cuál debe ser el tamaño de una muestra para estimar el nivel medio del total de proteínas entre los adultos con un margen de 0.5 g/dL y una confianza del 95 %, si se sabe que estos valores tienen un rango de aproximadamente 2.5 g/dL?

Tabla 6.7. El contraste es de un solo lado con $\alpha = 0.05$, $\beta = 0.10$, potencia = 0.90 y $\Delta = 0.50$, por lo que $n = 36$

Contraste de un solo lado / Contraste de dos lados	Nivel del contraste t																			
	α = 0.005 / α = 0.01					**α = 0.01 / α = 0.02**					**α = 0.025 / α = 0.05**					**α = 0.05 / α = 0.1**				
$\beta =$	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	**0.1**	**0.2**	0.5	
0.05																					0.05
0.10																					0.10
0.15																				122	0.15
0.20										139					99					70	0.20
0.25					100					90				128	64			139	101	45	0.25
0.30				134	78				115	63			119	90	45		122	97	71	32	0.30
0.35			125	99	58			109	85	47		109	88	67	34		90	72	52	24	0.35
0.40		115	97	77	45		101	85	66	37	117	84	168	151	26	101	70	55	40	19	0.40
0.45		92	77	62	37	110	81	68	53	30	93	67	54	41	21	80	55	44	33	15	0.45
0.50	100	75	63	51	30	90	66	55	43	25	76	54	44	34	18	65	45	[36]	27	13	0.50
Valor de $\Delta = \frac{\mu - \mu_0}{\sigma}$																					
0.55	83	63	53	42	26	75	55	46	36	21	63	45	37	28	15	54	38	30	22	11	0.55
0.60	71	53	45	36	22	63	47	39	31	18	53	38	32	24	13	46	32	26	19	9	0.60
0.65	61	46	39	31	20	55	41	34	27	16	46	33	27	21	12	39	28	22	17	8	0.65
0.70	53	40	34	28	17	47	35	30	24	14	40	29	24	19	10	34	24	19	15	8	0.70
0.75	47	36	30	25	16	42	31	27	21	13	35	26	21	16	9	30	21	17	13	7	0.75
0.80	41	32	27	22	14	37	28	24	19	12	31	22	19	15	9	27	19	15	12	6	0.80
0.85	37	29	24	20	13	33	25	21	17	11	28	21	17	13	8	24	17	14	11	6	0.85
0.90	34	26	22	18	12	29	23	19	16	10	25	19	16	12	7	21	15	13	10	5	0.90
0.95	31	24	20	17	11	27	21	18	14	9	23	17	14	11	7	19	14	11	9	5	0.95
1.00	28	22	19	16	10	25	19	16	13	9	21	16	13	10	6	18	13	11	8	5	1.00

3. Un investigador quiere estimar la pérdida de peso media alcanzada por pacientes en una clínica de adelgazamiento durante la primera semana, en régimen de dieta controlada y ejercicios. ¿Cuál debe ser el tamaño de una muestra para estimar esta media con un margen de 0.5 libras y una confianza del 95 %? Supongamos que se dispone de los siguientes datos a partir de las observaciones preliminares de cinco individuos:

   3.0 2.7 4.0 5.0 1.2

4. Un anuncio aparecido en abril de 1990 en *American Forests* afirmaba que un nuevo producto químico utilizado para preparar un terreno para plántulas de pinos produciría pinos de un año de edad con una altura media más de 1 pie mayor que la de los árboles cultivados en terrenos preparados utilizando el siguiente mejor tratamiento químico. Se sabe que *X,* crecimiento en el primer año, generalmente está entre 0 y 36 pulgadas. ¿Cuál será el tamaño mínimo de una muestra para contrastar lo pretendido con un nivel $\alpha = 0.05$ y con potencia = 0.90?
5. Considérese el experimento del Ejemplo 6.5.4. Si es crucial que se detecte una media tan alta como 7.2, ¿qué tamaño debe tener una muestra para contrastar la hipótesis dada al nivel $\alpha = 0.05$ de tal forma que la potencia del contraste sea de 0.95? (Utilizar los datos del Ejemplo 6.5.4 como estudio piloto.)
6. Se ha informado de que el consumo medio diario de nutrientes en mujeres jóvenes sanas es de 2300 kcal. La desviación típica registrada es de 237 kcal. El estudio está basado en una muestra pequeña y, por lo tanto, interesa repetirlo. ¿Qué tamaño debe tener una muestra para detectar una media que difiera de la registrada en 100 kcal, en más o en menos, con $\alpha = 0.10$ y potencia = 0.9? *(Sugerencia:* El contraste es de dos colas.) (Basado en la información hallada en Reinhold Tuschl et al., «Energy Expenditure and Everyday Eating Behavior in Healthy Young Women», *American Journal of Clinical Nutrition* julio de 1990, págs. 81-86.)

HERRAMIENTAS COMPUTACIONALES

TI83

XV. Generador de números aleatorios

La calculadora TI83 puede generar números enteros aleatorios por lo que desempeña las mismas funciones que la tabla de números aleatorios citada en la Sección 6.1. El uso del generador requiere que se especifique el mayor y el menor número entero deseados. Lo ilustraremos mediante la selección de 10 números enteros aleatorios entre 0 y 100.

Tecla/Comando de la TI83	**Propósito**
1. MATH ◁	1. Accede a la pantalla de probabilidad de la calculadora.
2. 5	2. Accede al generador de números aleatorios; rand Int (.
3. 0 100)	3. Indica que se seleccionaran números entre 0 y 100.
4. ENTER	4. Selecciona y presenta en pantalla el primer número aleatorio.

5. ENTER	5. Selecciona y presenta en pantalla el segundo número aleatorio.
6. Continuar pulsando ENTER hasta que sean seleccionados 10 números aleatorios.	6. Selecciona y presenta en pantalla los restantes números aleatorios.

XVI. Intervalo de confianza *T* para μ

Se pueden construir intervalos de confianza de la media con la calculadora TI83, si se dispone de los datos sin procesar de la muestra o si se conocen los valores de $\bar{x}$ y de *s*. Lo mostraremos calculando el intervalo de confianza del Ejemplo 6.3.2.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 1.0 ENTER 1.2 ENTER 1.8 ENTER	2. Introduce el conjunto de datos muestrales sin procesar.
3. STAT ◁ 8	3. Selecciona el programa utilizado para construir el intervalo de confianza *T* de μ.
4. cursor en DATA	4. Indica el conjunto de datos muestrales sin procesar que se está utilizando.
5. ▽ ▽ ▽ ▽ ENTER	5. Construye un intervalo de confianza, del 95 %, de μ.

Obsérvese que la calculadora incorpora por defecto el 0.95 % como nivel de confianza. Puede ser modificado sin más que sustituirlo por el nuevo valor deseado.

Para construir un intervalo de confianza cuando los datos de que se dispone son $\bar{x}$ y *s* (Ejemplo 6.3.2.), los pasos a seguir son:

Tecla/Comando de la TI83	**Propósito**
1. STAT ◁ 8	1. Selecciona el programa para construir el intervalo de confianza *T* de μ.
2. cursor a stats ENTER	*2.* Indica que se utilizarán los valores de $\bar{x}$ y *s*.
3. ▽ 1.57	3. Introduce para $\bar{x}$ el valor 1.57.
4. ▽ 0.26	4. Introduce para *s* el valor 0.26.

5. ▽ 16	5. Introduce como tamaño de la muestra el valor 16.
6. ▽ 0.95	6. Introduce 0.95 para el nivel de confianza.
7. ▽ ENTER	7. Construye el intervalo de confianza.

XVII. Contraste *T*

La TI83 está programada para hacer cualquiera de los tres tipos de contrastes *T* de la media. Puede realizarse el contraste tanto si se dispone del conjunto de datos muestrales, como si sólo se conocen los valores de *x y s*. Lo haremos con los datos del contraste de dos colas del Ejemplo 6.5.6. En este caso veremos la forma de calcular el valor *P y* mostrarlo en pantalla.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 7.23 ENTER 7.32 ENTER ⋮ 7.24 ENTER	2. Introduce los datos.
3. STAT ◁ 2	3. Accede a la pantalla del contraste *T* para una muestra.
4. cursor en DATA ENTER ▽	4. Indica el conjunto de datos muestrales que se está utilizando.
5. 7.25 ▽ ▽ ▽	5. Introduce el valor 7.25 de la hipótesis nula.
6. ▽	6. Indica que el contraste es de dos colas.
7. ▷ ENTER	7. Evalúa el estadístico de contraste; halla y muestra el valor *P* del contraste. (Nótese que habrá cierto error de redondeo en el valor calculado a mano. El resultado que proporciona la calculadora es el más preciso de los dos.)

Paquete estadístico SAS

VII. Contraste *T*

El SAS permite realizar el contraste *T* para una muestra mediante PROC MEANS. Puede utilizarse también este procedimiento para calcular $\bar{x}$, s y $s/\sqrt{n}$, que son necesarios para construir el intervalo de confianza de la media. Este procedimiento puede aplicarse únicamente en el caso de que el valor asignado en la hipótesis nula sea 0, para cualquier otro valor

hay que hacer una transformación. En particular, si $\mu_0 \neq 0$, se forma una nueva variable restándole el valor de μ_0, a cada dato. Si la media de la variable original X es μ_0, entonces la media de $X - \mu_0$ es 0. Así, para contrastar H_0: $\mu_0 = \mu_0$ vía PROC MEANS, contrastaremos H_0: $\mu_x - \mu_0 = 0$. Utilizaremos el Ejemplo 6.5.6, en el que $\mu_0 = 7.25$.

Código SAS	**Propósito**
OPTTONS LS = 80 PS = 60 NODATE;	Fija las especificaciones de impresión.
DATA BLOOD;	Designa al conjunto de datos.
INPUT X @@;	Nombra a la variable; puede aparecer más de una observación por línea.
TRANSFRM = X - 7.25;	Forma una nueva variable; si la media de X es 7.25, la media de TRANSFRM es 0.
LINES;	Indica que los datos siguen a continuación.
7.23 7.32 7.25 7.26 7.28 7.27 7.29 7.24	Líneas de datos.
	Señala el final de los datos.
PROC MEANS MEAN STD TPRT;	Solicita que se indique lo que se desea que calcule entre $\bar{x}$, s y $s/\sqrt{n}$, para todas las variables; contrasta H_0: $\mu_0 = 0$ para todas las variables y muestra el valor P del contraste.
TITLE 'TTEST ON THE AVERAGE PROTEIN LEVEL';	Titula la salida.

La salida del programa se presenta más abajo. Los estadísticos de la variable X se imprimen en la forma:

① $\bar{x}$, media de X.
② s, desviación típica de X.
③ $s/\sqrt{n}$, error estándar de la media.

Obsérvese que estos valores coinciden, diferencias de redondeo aparte, con los obtenidos anteriormente. Los estadísticos para la variable TRANSFRM = X - 7.25 aparecen en la forma siguiente:

④ media de TRANSFRM; obsérvese que esta media es, como se esperaba, $\bar{x} - 7.25$.
⑤ desviación típica de TRANSFRM; es igual a s, porque si se resta 7.25 a cada uno de los valores de X, no cambia la variabilidad de la dispersión.
⑥ error estándar de TRANSFRM; es igual a $s/\sqrt{n}$.

El contraste T de interés aparece en ⑦ y el valor P correspondiente, en ⑧. El valor observado del estadístico T difiere algo del hallado en el texto (valor del texto = 1.756) debido a que el SAS utiliza siete decimales en sus cálculos. En este contraste, y en todo lo que sigue, el SAS realiza automáticamente el *contraste de dos colas*. El valor $P = 0.1334$ mostrado en ⑧ es para la hipótesis alternativa de dos colas H_1: $\mu \neq 7.25$. Esto es coherente con la afirmación del texto de que $0.10 < P < 0.20$.

El SAS no construye el intervalo de confianza tipo T automáticamente. Sin embargo, como puede ver, calcula fácilmente $\bar{x}$ y $s/\sqrt{n}$. Rápidamente puede obtenerse con ellos el intervalo de confianza.

Nota: Para obtener el valor *P* correcto de un contraste *T* de una cola, debe dividir por dos el valor *P* que proporciona el SAS. El programa no pensará por usted. Le corresponde a usted la responsabilidad de interpretar correctamente los resultados, en el contexto de su estudio.

TTEST ON THE AVERAGE TOTAL PROTEIN LEVEL

Variable	Mean	Std Dev	Std Error	T	Prob > |T|
X	7.2675000 ①	0.0291548 ②	0.0103078 ③	705.0510620	0.0001
TRANSFRM	0.0175000 ④	0.0291548 ⑤	0.0103078 ⑥	1.6977494 ⑦	0.1334 ⑧

Distribución ji-cuadrado e inferencias sobre la varianza

En este capítulo, presentamos otra importante familia de variables aleatorias continuas. Estas variables aleatorias, llamadas ji-cuadrado, se utilizan en muchos de los diferentes entornos estadísticos. Aquí se presenta su aplicación para sacar conclusiones relacionadas con la varianza de la población.

7.1. DISTRIBUCIÓN JI-CUADRADO Y ESTIMACIÓN POR INTERVALO DE LA VARIANZA POBLACIONAL

Ya se ha hablado de que la varianza muestral S^2 es un estimador puntual lógico de la varianza de la población σ^2. Las estimaciones de σ^2 obtenidas en repetidas muestras de la misma población fluctúan alrededor de la verdadera varianza de la población. Hasta aquí, las estimaciones de σ^2 han sido objeto de interés ya que son necesarias para evaluar el estadístico *T,* utilizado para hacer inferencias sobre la media de la población.

En algunas ocasiones, el interés no se centra en la media sino en la varianza misma. Por lo tanto, no sólo es necesario poder hallar una estimación puntual de σ^2, sino también construir un intervalo de confianza o hacer un contraste de hipótesis para este parámetro.

Ejemplo 7.1.1. El cobre, mineral requerido en algún grado por la mayoría de las plantas, se considera un micronutriente. Su concentración en una planta se mide en partes por millón, y se determina quemando totalmente la planta y analizando las cenizas. La concentración de cobre varía de una especie a otra. Se diseña un experimento para estimar esta variabilidad.

Ejemplo 7.1.2. El Lhasa Apso es un perro de ascendencia tibetana. La raza fue introducida en Inglaterra en 1921, y criada para exhibición por primera vez en 1933. La alzada deseada en los machos es de 10.5 pulgadas. En todo caso, hay mucha variabilidad en la altura. Un criador está intentando reducirla mediante una crianza selectiva.

Ejemplo 7.1.3. Al manufacturar un fármaco, su potencia no permanece constante. Esta variabilidad no debe ser demasiado grande. Una varianza muy grande puede dar como resultado que algunos lotes contengan fármaco demasiado débil para que sea efectivo, mientras que otros contengan fármaco demasiado fuerte y potencialmente peligroso. Se realizan pruebas periódicas para controlar la varianza del fármaco que está siendo producido.

En cada uno de estos tres ejemplos hay que inferir la varianza de la población. Para construir un intervalo de confianza de σ^2 o para contrastar hipótesis relativas a su valor, es necesario introducir otra familia de variables aleatorias continuas. Las características generales de las variables de esta familia, llamadas *variables ji-cuadrado* (X^2), son las siguientes:

1. Hay un número infinito de variables aleatorias ji-cuadrado, identificada cada una por un parámetro γ, llamado *grados de libertad.* El parámetro γ es siempre un entero positivo. La notación $\mathbf{X}^2_\gamma$, designa una variable ji-cuadrado con γ grados de libertad.
2. Cada variable ji-cuadrado es continua.
3. La gráfica de la densidad de cada variable ji-cuadrado con $\gamma \geqslant 2$ es una curva asimétrica que presenta generalmente la forma que aparece en la Figura 7.1a; cuando $\gamma = 1$, la gráfica adopta la forma que se muestra en la Figura *7.1b.*
4. Las variables ji-cuadrado no pueden tomar valores negativos.
5. El parámetro γ es, al mismo tiempo, un parámetro de forma y un parámetro de localización, en el que

$$E[\mathbf{X}^2_\gamma] = \gamma \quad \text{y} \quad \text{Var } \mathbf{X}^2_\gamma = 2\gamma$$

Es decir, el valor medio de una variable aleatoria ji-cuadrado es el mismo que sus grados de libertad y su varianza es el doble de sus grados de libertad.

En la Tabla VIII del Apéndice B se presenta una recopilación parcial de valores de la función de distribución acumulativa con varios grados de libertad para las variables ji-cuadrado En esta tabla, los grados de libertad aparecen encabezando las filas, las probabilidades encabezando las columnas, y los puntos asociados con esas probabilidades en el cuerpo de la tabla.

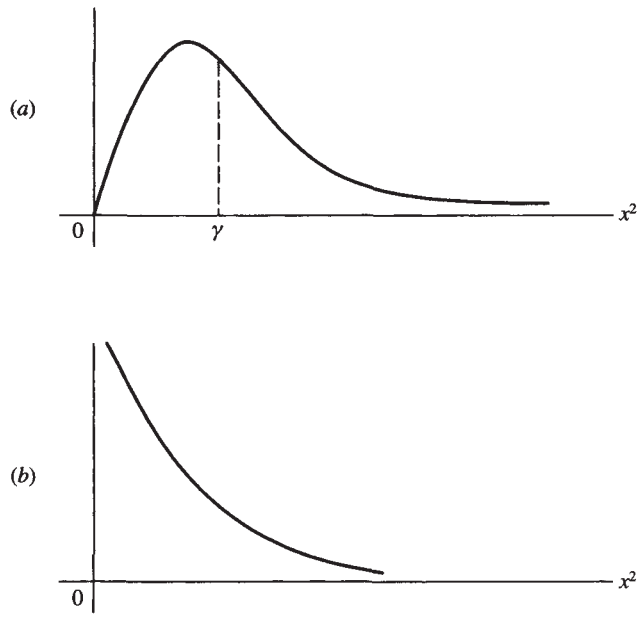


Figura 7.1. (a) Curva ji-cuadrado típica con $\gamma \geqslant 2$. El punto de equilibrio de la curva es γ, sus grados de libertad. (b) Curva ji-cuadrado con $\gamma = 1$.

Ejemplo 7.1.4. Considérese una variable aleatoria ji-cuadrado con diez grados de libertad. Esta variable aleatoria se indica por $\mathbf{X}^2_{10}$. Su valor medio es 10, el mismo que sus grados de libertad, y su varianza es 20, el doble de sus grados de libertad.

a) Hallar $P[\mathbf{X}^2_{10} \leqslant 2.56]$. Para calcularla miramos en la fila 10 de la Tabla VIII. El valor 2.56 se encuentra en la columna 0.10. Por lo tanto, $P[\mathbf{X}^2_{10} \leqslant 2.56] = F(2.56) = 0.01$.

b) Hallar $P[\mathbf{X}^2_{10} \geqslant 16]$. Puesto que 16 se encuentra en la columna 0.900 y en la fila 10, $P[\mathbf{X}^2_{10} \leqslant 16] = 0.900$. Así, $P[\mathbf{X}^2_{10} \geqslant 16] = 1 - F(16) = 1 - 0.90 = 0.1$.

c) El punto con el 5 % del área a su derecha y el 95 % a su izquierda se muestra en la Figura 7.2. Su valor numérico, hallado en la fila 10 y la columna 0.95 de la tabla de ji-cuadrado, es 18.3.

d) El punto con el 95 % del área a su derecha y el 5 % a su izquierda se halla en la fila 10 y en la columna 0.05. Su valor numérico es 3.94.

e) Hallar el área bajo la gráfica de la densidad de $\mathbf{X}^2_{10}$ entre 3.25 y 20.5. Este área, que es igual a $P[3.25 \leqslant \mathbf{X}^2_{10} \leqslant 20.5]$, se muestra en la Figura 7.3a. Como el área a la izquierda de 20.5 es 0.975 y el área a la derecha de 3.25 es 0.025, el área entre estos valores es 0.975 - 0.025 = 0.950 (véase Fig. 7.36).

Intervalo de confianza para σ^2 (opcional)

Para construir un intervalo de confianza de σ^2 necesitamos una variable aleatoria que contenga en su expresión a este parámetro y cuya distribución se conozca. Si es posible hallar dicha variable aleatoria, se puede utilizar un procedimiento algebraico, similar al utilizado para construir un intervalo de confianza de μ, para determinar los límites de confianza de σ^2. El Teorema 7.1.1 proporciona la variable aleatoria necesaria.

Teorema 7.1.1. Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria de tamaño n de una distribución que es normal, con media μ y varianza σ^2. La variable aleatoria $(n - 1)S^2/\sigma^2$ se distribuye como una variable aleatoria ji-cuadrado con n - 1 grados de libertad.

El Ejemplo 7.1.5 ilustra la forma de aplicar este teorema para construir un intervalo de confianza de σ^2.

Ejemplo 7.1.5. Se extrajo una muestra aleatoria de 16 plantas para estimar la varianza en la concentración de cobre en las plantas halladas en el New River Valley. Se quemaron las plantas, se analizaron sus cenizas y se obtuvieron las siguientes observaciones con respecto *a X,*

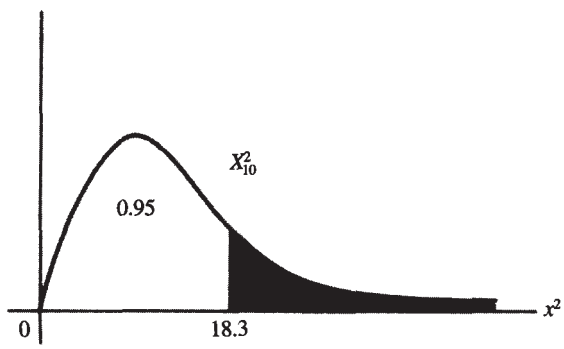


Figura 7.2. $P[X^2_{10} \geqslant 18.3] = 0.05$.

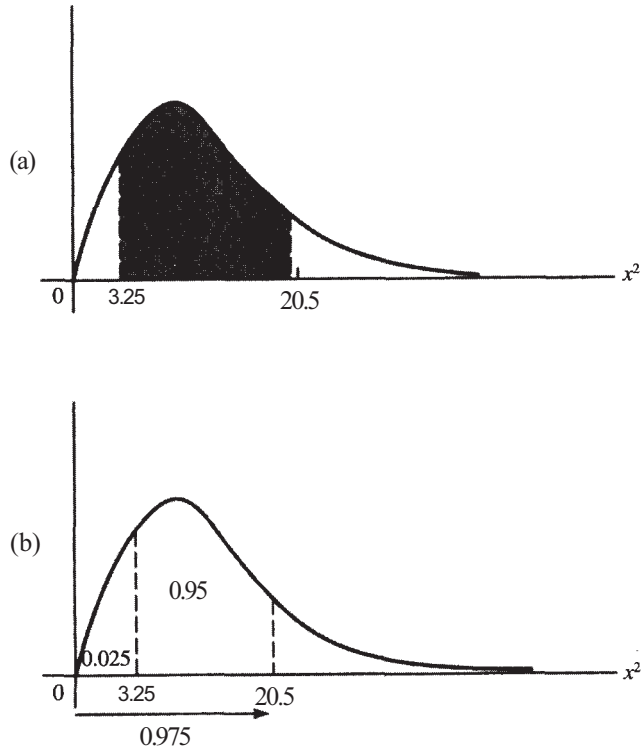


Figura 7.3. (a) Área sombreada = $P[3.25 \leqslant X^2_{10} \leqslant 20.5]$. *(b)* $P[3.25 \leqslant X^2_{10} \leqslant 20.5] = 0.95$.

concentración de cobre (en partes por millón) (supóngase que X está normalmente distribuida):

5	3	34	18	27	14
8	50	38	43	35	
20	70	25	60	19	

Para estos datos $s^2 = 377.30$. La partición de la curva X^2_{15}, que se necesita para construir un intervalo de confianza para σ^2 del 90 % se muestra en la Figura 7.4.

Obsérvese que, puesto que la muestra es de tamaño 16, el número de grados de libertad es 15, el tamaño de la muestra menos 1. Como se hizo antes, se obtiene un intervalo de confianza del 90 % mediante la partición de la curva, de forma que el 90 % del área se encuentre en el centro y el 10 % restante dividido en dos porciones iguales. En la Figura 7.4 se ve claramente que

$$P[7.26 \leqslant X^2_{15} \leqslant 25] = 0.90$$

En este caso, la variable aleatoria ji-cuadrado es

$$\frac{(n-1)S^2}{\sigma^2} = \frac{15S^2}{\sigma^2}$$

Por lo tanto, sabemos que

$$P\left[7.26 \leqslant \frac{15S^2}{\sigma^2} \leqslant 25\right] = 0.90$$

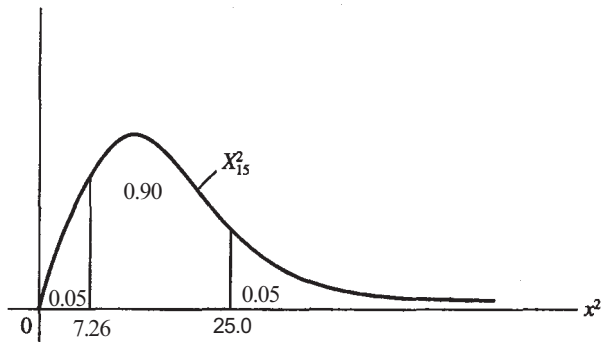


Figura 7.4. Partición de X^2_{15} para obtener un intervalo de confianza de σ^2 del 90%.

Nuestro objetivo es aislar σ^2 en el centro de la desigualdad. Para hacerlo, primero se invierte cada miembro de la desigualdad y, a continuación, se multiplica por $15S^2$. Esto da como resultado la siguiente secuencia de pasos:

$$P\left[7.26 \leqslant \frac{15S^2}{\sigma^2} \leqslant 25\right] = 0.90$$

$$P\left[\frac{1}{25} \leqslant \frac{\sigma^2}{15S^2} \leqslant \frac{1}{7.26}\right] = 0.90$$

$$P\left[\frac{15S^2}{25} \leqslant \sigma^2 \leqslant \frac{15S^2}{7.26}\right] = 0.90$$

El límite inferior del intervalo de confianza es

$$\frac{15S^2}{25} = \frac{15(377.30)}{25.0} = 226.38$$

y el límite superior es

$$\frac{15S^2}{7.26} = \frac{15(377.30)}{7.26} = 779.55$$

Tenemos un 90 % de confianza en que el verdadero valor de la varianza de la concentración de cobre, en el New River Valley, está comprendida entre 226.38 y 779.55. Para hallar un intervalo de confianza del 90 % de la desviación típica de la población, necesitamos solamente extraer la raíz cuadrada de esos límites numéricos. De este modo, podemos decir que tenemos un 90 % de confianza en que la verdadera desviación típica de la población esté entre $\sqrt{226.38} = 15.04$ y $\sqrt{779.54} = 27.92$ partes por millón.

La técnica utilizada en el Ejemplo 7.1.5 puede generalizarse para obtener cualquier nivel de confianza deseado, eligiendo adecuadamente los puntos de la tabla de la ji-cuadrado. Obsérvese que, dado que la distribución ji-cuadrado no es simétrica, deben leerse en la tabla dos puntos diferentes. El punto que aparece en el límite inferior se designará mediante χ^2_1. Es el más grande de los dos puntos leídos en la tabla. El punto que aparece en el límite superior se indicará mediante χ^2_2. Los límites de confianza generales se dan en el Teorema 7.1.2.

Teorema 7.1.2. Intervalo de confianza de σ^2. Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria de tamaño n de una distribución normal con media μ y varianza σ^2. Los límites inferior y superior, respectivamente, de un intervalo de confianza para σ^2 son

$$L_1 = \frac{(n-1)S^2}{\chi_1^2} \qquad L_2 = \frac{(n-1)S^2}{\chi_2^2}$$

donde los puntos x^2 lo son de la distribución ji-cuadrado con n - 1 grados de libertad.

Al igual que en el caso de la distribución *T,* pueden hallarse aproximaciones de los puntos ji-cuadrado, para muestras grandes, utilizando la distribución normal estándar. El procedimiento se muestra en el Ejercicio 5, al final de esta sección. Este ejercicio posibilita la construcción de intervalos de confianza sobre σ^2 para muestras mayores que las que pueden manejarse utilizando la tabla de ji-cuadrado. Los puntos ji-cuadrado para muestras grandes pueden obtenerse también con la mayoría de los paquetes estadísticos y con la calculadora TI83.

EJERCICIOS 7.1

1. Considérese la variable aleatoria $\mathbf{X}_9^2$.
   a) Hallar $P[\mathbf{X}_9^2 \leqslant 2.09]$.
   b) Hallar $P[\mathbf{X}_9^2 \geqslant 11.4]$.
   c) Hallar $P[14.7 \leqslant \mathbf{X}_9^2 \leqslant 16.9]$.
   d) Encontrar el punto que deja a su derecha un área de 0.025.
   e) Encontrar el punto que deja a su izquierda un área de 0.01.
   f) Encontrar puntos χ_1^2 y χ_2^2 tales que el área a la derecha de χ_2^2 sea igual al área a la izquierda de χ_1^2 y $P[\chi_1^2 \leqslant \mathbf{X}_9^2 \leqslant \chi_2^2] = 0.90$. *{Sugerencia:* Hacer un dibujo de estos puntos.)
2. Considérese la variable aleatoria $\mathbf{X}_{15}^2$. Calcular:
   a) $E[\mathbf{X}_{15}^2]$.
   b) Var $\mathbf{X}_{15}^2$.
   c) El punto que deja a su derecha un área de 0.05.
   d) El punto que deja a su izquierda un área de 0.10.
   e) $P[\mathbf{X}_{15}^2 \leqslant 7.26]$.
   f) $P[\mathbf{X}_{15}^2 \geqslant 27.5]$.
3. En los inviernos rigurosos, se utiliza sal para quitar el hielo de las carreteras. Para hallar la cantidad aproximada de sal que se está introduciendo en el medio ambiente por esta causa, se realizó un estudio en New England. Se obtuvieron las siguientes observaciones sobre la variable aleatoria *X,* número de toneladas métricas de sal utilizadas sobre las carreteras por semana, en distritos aleatoriamente seleccionados, a lo largo de la región:

   3900 3875 3820 3860 3840
   3852 3800 3825 3790

   a) Establecer una estimación puntual de μ.
   b) Establecer una estimación puntual de σ^2
   c) Supóngase que *X* está normalmente distribuida. Hallar un intervalo de confianza de μ del 90%.
   d) Establecer intervalos de confianza del 90 % para σ^2 y para σ.
4. La célula típica de una planta tiene una gran cantidad de citoplasma limitado por una membrana celular llamada membrana plasmática. El espesor medio de esta membrana

varía de una especie a otra. Una muestra aleatoria de 20 especies proporcionó las siguientes observaciones sobre *X,* espesor medio de la membrana celular angstrom:

80	90	85	82	75	58	70
84	87	81	87	61	73	84
85	70	78	95	77	52	

d) Establecer una estimación puntual de σ^2.
b) Supóngase que X está distribuida normalmente. Encontrar intervalos de confianza del 95 % de μ, σ^2 y σ.

5. *Aproximación normal para* X^2. Obsérvese que en la tabla de ji-cuadrado aparecen valores de γ del 1 al 30. Por lo tanto, puede utilizarse para tamaños de muestras de 2 a 31. Para muestras mayores, puede hallarse la aproximación de los puntos ji-cuadrado mediante la fórmula

$$\chi_r^2 \cong \left(\tfrac{1}{2}\right)\left[z_r + \sqrt{2\gamma - 1}\right]^2$$

donde el subíndice r indica el área deseada a la *derecha* del punto. Recuérdese que y representa los grados de libertad asociados con el punto ji-cuadrado. Por ejemplo, para una muestra de tamaño 50, el punto $\chi^2_{0.025}$ viene dado por

$$\begin{aligned}\chi^2_{0.025} &\cong \left(\tfrac{1}{2}\right)\left[z_{0.025} + \sqrt{2\gamma - 1}\right]^2 \\ &= \left(\tfrac{1}{2}\right)\left[1.96 + \sqrt{2(49) - 1}\right]^2 \\ &\cong 69.72\end{aligned}$$

Hallar la aproximación de cada uno de los puntos enumerados de *a* a c.
a) $\chi^2_{0.05}$ y $\chi^2_{0.90}$; $\gamma = 79$.
b) $\chi^2_{0.025}$ y $\chi^2_{0.975}$; $\gamma = 99$.
c) $\chi^2_{0.005}$ y $\chi^2_{0.995}$; $\gamma = 74$.

6. Se ha efectuado un estudio sobre la obesidad en niños menores de 12 años. Se ha obtenido una muestra de 100 niños obesos y se ha averiguado la edad de cada niño cuando comenzó a sufrir la obesidad. Se ha determinado que la media muestral es de 4 años, con una desviación típica muestral de 1.5 años.
a) Encontrar un intervalo de confianza del 95 % para la edad media de inicio de la obesidad de los niños.
b) Determinar un intervalo de confianza del 95 % para la varianza de la edad de inicio de la obesidad.
c) Determinar un intervalo de confianza del 95 % para la desviación típica de la edad de inicio de la obesidad.
(Basado en la información hallada en Rebecca Unger, Lisa Kreeger y Catherine Christoffel, «Childhood Obesity», *Clinical Pediatrics,* julio 1990, págs. 368-373).

7. Se ha realizado un estudio para estimar las características del terreno inundable de una parte del río Mississippi. Una de las variables estudiadas es *X,* anchura del terreno inundable. Una muestra de las mediciones obtenidas en 61 lugares seleccionados aleatoriamente dio como resultado $\bar{x} = 3400$ m y $s = 100$ m.
a) Hallar un intervalo de confianza del 90 % para la anchura media del terreno inundable en la región.

b) Hallar intervalos de confianza del 90 % para la varianza y la desviación típica de *X*. (Basado en la información hallada en Frederick Swanson y Richard Sparks, «Long Term Ecological Research and the Invisible Place», *Bioscience,* agosto 1990, págs. 502-508.)

7.2. CONTRASTES DE HIPÓTESIS DE LA VARIANZA POBLACIONAL (OPCIONAL)

Si existe una idea preconcebida en relación al valor de la varianza o de la desviación típica de una variable aleatoria, es adecuado tratar el problema con las técnicas del contraste de hipótesis.

Los contrastes de hipótesis para σ^2 toman las mismas formas generales que los relativos a la media. Los describimos a continuación, representando con σ_0^2 el valor hipotético de la varianza poblacional.

I	II	III
H_0: $\sigma^2 \leqslant \sigma_0^2$	H_0: $\sigma^2 \geqslant \sigma_0^2$	H_0: $\sigma^2 = \sigma_0^2$
H_1: $\sigma^2 > \sigma_0^2$	H_1: $\sigma^2 < \sigma_0^2$	H_1: $\sigma^2 \neq \sigma_0^2$
Contraste con cola a la derecha	Contraste con cola a la izquierda	Contraste con dos colas

El estadístico para contrastar cada una de ellas es

$$\frac{(n-1)S^2}{\sigma_0^2}$$

Este estadístico, bajo el supuesto de que el valor hipotético de σ^2 es correcto, sigue una distribución ji-cuadrado con n - 1 grados de libertad.

En un contraste con cola a la derecha, se rechaza H_0 para los valores del estadístico del contraste que son demasiado grandes para haberse producido por azar; se hace un contraste con cola a la izquierda, rechazando H_0 para los valores pequeños del estadístico del contraste. Como en el caso de los contrastes *T,* la decisión se realiza basándose en el valor *P* calculado. El Ejemplo 7.2.1 ilustra el contraste y explica la lógica existente en las normas de rechazo.

Ejemplo 7.2.1. La varianza habitual con respecto a la altura de los machos de Lhasa Apso es de 0.25. Un criador está intentando reducir esta cifra. Después de un período de crianza selectiva, hay que elegir y medir de entre los animales una muestra aleatoria de 15 machos. Puesto que el argumento del investigador se toma como hipótesis alternativa, el propósito del experimento es contrastar

$$H_0: \sigma^2 \geqslant 0.25 \qquad H_1: \sigma^2 < 0.25$$

El estadístico del contraste a utilizar es

$$\frac{(n-1)S^2}{\sigma_0^2} = \frac{14S^2}{0.25}$$

que, si H_0 es cierta, tiene una distribución ji-cuadrado con n - 1 = 14 grados de libertad. Puesto que S^2 es un estimador para σ^2, si H_0 es cierta, esperamos que el valor numérico de S^2

esté próximo a 0.25, valor hipotético de σ^2. Ello fuerza al cociente $S^2/0.25$ a estar próximo a 1, y al valor del estadístico del contraste a estar próximo a 14, su valor esperado. Si, en todo caso, la alternativa es cierta y la varianza de la población es efectivamente más pequeña que el valor hipotético 0.25, esperamos que S^2 tenga un valor más pequeño que 0.25. Esto, a su vez, puede forzar al cociente $S^2/0.25$ a ser menor que 1, obteniéndose un valor menor que 14 para el estadístico del contraste. De este modo, es lógico rechazar H_0 en favor de H_1 si el valor observado del estadístico es demasiado pequeño para que se deba al azar. Una vez realizado el experimento, se obtuvo una varianza muestral de 0.21. El valor del estadístico del contraste es $14(0.21)/0.25 = 11.76$.

El valor *P* es la probabilidad de observar un valor de 11.76 o menor. Esta probabilidad se ilustra en la Figura 7.5. Su valor aproximado puede hallarse a partir de la fila 14 de la tabla de ji-cuadrado. Dado que el valor 11.76 cae entre 10.2 (área a la izquierda 0.25) y 13.3 (área a la izquierda 0.50), el valor *P* está entre 0.25 y 0.50. Puesto que esta probabilidad es grande, no rechazaremos H_0. No tenemos suficientes pruebas para sostener la afirmación de que la crianza selectiva haya reducido la variabilidad en las alturas de los machos Lhasa Apsos.

Al igual que en el caso de los contrastes *T,* si se desea, puede contrastarse H_0 utilizando un nivel α prefijado. Si es así, rechazamos H_0 en caso de que el valor *P* calculado sea, a lo sumo, igual a α.

Obsérvese que, al igual que, en el caso de los contrastes *T,* se ha hecho una suposición de normalidad. En el caso del estadístico ji-cuadrado, la violación de esta suposición puede producir serios errores. Si la población sobre la que se realiza el muestreo no es normal, debe utilizarse otro método para contrastar las hipótesis sobre σ^2. En [2] se explica dicho método.

EJERCICIOS 7.2.

1. Una variable estudiada por los biólogos es la temperatura interna del cuerpo en los animales poiquilotermos (animales cuya temperatura corporal fluctúa con el ambiente circundante). El nivel letal (DL_{50}) para los lagartos del desierto es de 45 °C. Se ha observa-

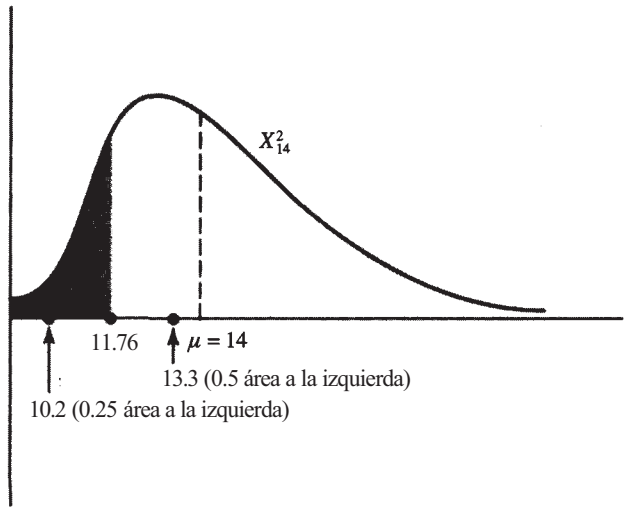


Figura 7.5

do que la mayor parte de estos animales se oculta del calor en verano para evitar aproximarse a este nivel letal. Se realiza un experimento para estudiar *X,* tiempo (en minutos) que se requiere para que la temperatura del cuerpo de un lagarto del desierto alcance los 45 °C, partiendo de la temperatura normal de su cuerpo mientras está a la sombra. Se obtuvieron los siguientes datos:

10.1	12.5	12.2	10.2	12.8	12.1
11.2	11.4	10.7	14.9	13.9	13.3

a) Hallar estimaciones puntuales de μ, σ^2 y σ.
b) Supóngase que X es normal. Basándose en estos datos, ¿puede concluirse que el tiempo medio requerido para alcanzar la dosis letal es menor que trece minutos? Explicar la respuesta, a partir del valor P.
c) ¿Puede concluirse que la desviación típica de X es menor de un minuto y medio? Explicar la respuesta en función del valor P. *Sugerencia:* Contrastar

$$H_0: \sigma \geqslant 1.5 \qquad H_1: \sigma < 1.5$$

es equivalente a contrastar

$$H_0: \sigma^2 \geqslant (1.5)^2 \qquad H_1: \sigma^2 < (1.5)^2$$

2. El calcio se presenta normalmente en la sangre de los mamíferos en concentraciones de alrededor de 6 mg/100 mL de sangre total. La desviación típica normal de esta variable es 1 mg de calcio por cada 100 mL de sangre total. Una variabilidad mayor que ésta puede ocasionar graves trastornos en la coagulación de la sangre. Una serie de nueve pruebas realizadas sobre un paciente revelaron una media muestral de 6.2 mg de calcio por 100 mL de sangre total y una desviación típica muestral de 2 mg de calcio por cada 100 mL de sangre. ¿Hay alguna evidencia, con $\alpha = 0.05$, de que el nivel medio de calcio para este paciente sea más alto del normal? ¿Hay alguna evidencia, a un nivel $\alpha = 0.05$ de que la desviación típica del nivel de calcio sea más alta de la normal?
3. Para satisfacer las necesidades respiratorias de los peces de agua caliente, el contenido de oxígeno disuelto debería presentar un promedio de 6.5 partes por millón (ppm), con una desviación típica no mayor de 1.2 ppm. Cuando la temperatura del agua crece, el oxígeno disuelto decrece. Esto causa la asfixia del pez. Se realiza un estudio sobre los efectos del calor en verano sobre un gran lago. Después de un período particularmente caluroso, se toman muestras de agua en 25 lugares aleatoriamente seleccionados en el lago, y se determina el contenido de oxígeno disuelto. Resultan una media muestral de 6.3 ppm y una desviación típica muestral de 1.7.
   a) ¿Hay evidencia, al nivel $\alpha = 0.05$, de que el contenido medio de oxígeno en el lago ha descendido del nivel aceptable de 6.5 ppm?
   b) Contrastar

$$H_0: \sigma \leqslant 1.2$$
$$H_1: \sigma > 1.2$$

   Basándose en los datos dados, ¿puede rechazarse H_0 al nivel $\alpha = 0.05$?
4 Se cree que el crecimiento medio de los abetos adultos cerca de Whitewater Bay, Alaska, es de 19 pulgadas al año, con una desviación estándar de 0.5 pulgadas. Los investigadores

creen que las duras condiciones de sequía han producido una disminución de ambos valores. Una muestra de 50 árboles arroja un crecimiento medio de 18.9 pulgadas, con una desviación estándar muestral de 0.15 pulgadas.

a) Establecer la hipótesis de investigación en relación con el crecimiento medio. ¿Son compatibles los datos con esta teoría? Explicarlo, basándose en el valor *P* del contraste.

b) Establecer la hipótesis de investigación concerniente a la desviación estándar. ¿Son compatibles los datos con esta teoría? Explicarlo, basándose en el valor *P* del contraste.

c) ¿Cree usted, desde un punto de vista práctico, que la diferencia en el crecimiento medio es suficiente para preocuparse?

(Basado en la información hallada en Herbert McClean, «Return to Admirality», *American Forests,* agosto 1989, págs. 21-24.)

Inferencias sobre proporciones

En este capítulo, estudiaremos inferencias sobre una proporción y la comparación de dos proporciones. El Teorema central del límite, estudiado en el Capítulo 6, nos dará la justificación teórica para los procedimientos que aquí se presentan. A lo largo de todo este capítulo, supondremos que los tamaños de las muestras son lo suficientemente grandes para justificar el uso de este teorema.

8.1. ESTIMACIÓN PUNTUAL

Consideremos la situación siguiente: en una población de interés se está estudiando un rasgo particular y cada miembro de la población puede clasificarse según que posea o no ese rasgo. Las inferencias se hacen con respecto al parámetro p, proporción de la población que tiene el rasgo. (Véase Fig. 8.1.)

Obsérvese que en este caso no utilizamos nuestra notación habitual para designar los parámetros de población mediante letras griegas. No existe una letra griega empleada conveniente y comúnmente para la proporción de la población. Utilizamos la letra p porque es una elección natural para representar una proporción, un porcentaje o una probabilidad.

El Ejemplo 8.1.1. ilustra una situación en la que el interés se centra en la estimación de una proporción de la población.

Figura 8.1. Partición de una población con respecto a un rasgo.

Ejemplo 8.1.1. Hay más varones que mujeres que padecen alguna forma de retraso mental Recientemente, se ha propuesto el síndrome del cromosoma X frágil como posible causa del fenómeno. Este defecto hereditario se transmite generalmente de madre a hijo y aparece como un defecto del cromosoma X, o cromosoma femenino. Se efectúa un estudio para estimar la proporción de varones que están en instituciones para retrasados con este problema. En este caso, la población de interés es la de varones en instituciones para retrasados; el rasgo en estudio es el defecto en el cromosoma X. Queremos obtener una estimación puntual para la proporción p de varones en estas instituciones que tienen este defecto particular. Esquemáticamente, la situación se representa en la Figura 8.2.

¿Cuál es un estimador puntual lógico para p? El sentido común indica que deberíamos extraer una muestra aleatoria de la población de interés, determinar la proporción de objetos

"OH, OH, *ACABO* DE DESCUBRIR QUE EL 79% DE MIS RATAS TIENEN CÁNCER...
¡Y AÚN NO LES HE INYECTADO NADA!

con el rasgo en la muestra y utilizar esta «proporción muestral» como estimación de la proporción p de la población. Es decir,

$$\frac{X}{n} = \frac{\text{número de objetos en la muestra con el rasgo}}{\text{tamaño de la muestra}} = \text{proporción}$$

es un estimador lógico para p.

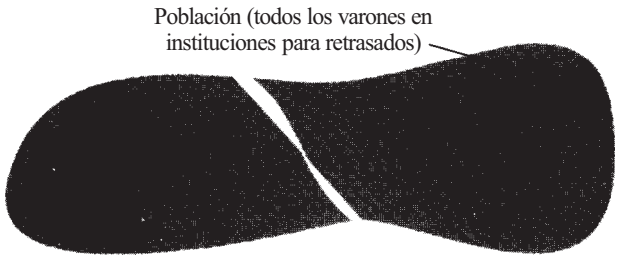


Figura 8.2. Partición de la población relativa a la presencia del cromosoma X frágil.

Para remarcar el hecho de que la proporción muestral es un estimador de la proporción de la población, utilizaremos la notación de «acento circunflejo». Con esta notación, se coloca un acento circunflejo **(^)** sobre el parámetro para indicar que el estadístico dado es un estimador puntual del parámetro. En este caso, indicamos que el estadístico X/n es un estimador de p escribiendo

$$\frac{X}{n} = \hat{p}$$

Ejemplo 8.1.2. Se obtuvo una muestra aleatoria de 150 varones en instituciones para retrasados mentales. De éstos, se encontró que cuatro presentaban el cromosoma X frágil. Basándose en esta muestra,

$$\hat{p} = \frac{x}{n} = \frac{4}{150} = 0.027$$

Observe que, multiplicando por 100, esta proporción puede convertirse en un porcentaje (2.7 %). Si esta estimación refleja exactamente el verdadero porcentaje de varones retrasados que presentan este defecto, ello puede significar que esta alteración es la segunda en importancia de las causas conocidas de retraso mental después del síndrome de Down (mongolismo).

En la siguiente sección, se utilizará la proporción muestral para la obtención del intervalo de confianza de la proporción de la población. Para ello, debemos preguntarnos cuál es la distribución de $\hat{p}$. Esta pregunta será contestada por el Teorema 8.1.1. En el Ejercicio 7 se esboza la prueba de este teorema.

Teorema 8.1.1. Para muestras de tamaño grande, la proporción de la muestra

$$\hat{p} = \frac{X}{n} = \frac{\text{número de objetos en la muestra con el rasgo}}{\text{tamaño de la muestra}}$$

tiene una distribución aproximadamente normal. Además, la media de $\hat{p}$ es p y su varianza es $p(1 - p)/n$.

EJERCICIOS 8.1

1. El virus del herpes simple, un virus muy común, produce una de las formas más molestas de enfermedad venérea. Un estudio reciente probó una pomada que contenía el azúcar 2-desoxi-D-glucosa, sobre 36 mujeres con infecciones genitales de herpes. En el transcurso de cuatro días, los síntomas disminuyeron en 32 de los 36 casos. Encontrar una estimación puntual para p, proporción de mujeres para las que este tratamiento se mostrará eficaz.
2. Se realiza un estudio de ámbito nacional para estimar la proporción p de individuos de dieciséis años o menos que fuman regularmente. De 1000 individuos entrevistados, 200 fumaban regularmente. Encontrar una estimación puntual para p.
3. Un estudio reciente indica que el estrés agudo puede inducir a que se produzcan alteraciones en el corazón capaces de ocasionar la muerte. Se obtuvo confirmación de ello

examinando 15 casos en los que los individuos murieron después de una agresión física, a pesar de que los daños por sí solos no fueron lo suficientemente graves como para causar la muerte. De ellos, 11 fueron consecuencia de una degeneración de las células del corazón, llamada degeneración miofibrilar. Encontrar una estimación puntual para *p,* proporción de muertes debidas a degeneración miofibrilar en casos de agresiones.

4. Considérese un pequeño estudio piloto diseñado para estimar la proporción de pacient0s, que actualmente trabajan, atendidos en una clínica gratuita. Se ha obtenido una muestra de 20 individuos. Se clasifica cada dato indicándolo por $X_i = 1$ si la i-ésima persona elegida trabaja y por $X_i = 0$ en cualquier otro caso. Los datos obtenidos se muestran en la Tabla 8.1.
   a) Calcular $\sum x_i$.
   b) Comprobar que $\sum x_i = x$, número de personas de la muestra que están empleadas.
   c) Estimar la proporción de personas de la muestra que están empleadas.
5. Extraer una muestra aleatoria de 30 números de un dígito de la tabla de números aleatorios (Tabla IV del Apéndice B). Sea $X_i = 1$ si el i-ésimo número extraído es un 5, y $X_i = 0$ si es cualquier otro. Estimar la proporción de cincos de la tabla.
6. Extraer una muestra aleatoria de tamaño 10 de los árboles de la Tabla V del Apéndice 3. Sea $X_i = 1$ si el árbol elegido tiene un DMAP (diámetro medio a la altura del pecho) mayor que 5 y $X_i = 0$ en cualquier otro caso.
   a) Hallar $\sum x_i$.
   b) Calcular el número de árboles de la muestra con un DMAP mayor que 5.
   c) Estimar la proporción de árboles de la muestra con un DMAP mayor que 5.
   d) Estimar el número de árboles de la tabla con un DMAP mayor que 5.
7. Definir un conjunto de variables aleatorias independientes $X_1, X_2, \ldots, X_n$ de manera que $X_i = 1$ si el i-ésimo elemento de la muestra tiene la característica de interés y $X_i = 0$ si no la tiene. Por ejemplo, si el primer varón seleccionado en el Ejemplo 8.1.2 tiene el cromosoma X frágil, entonces $x_i = 1$, y si no la tiene, entonces $x_1 = 0$. Si la probabilidad de que aparezca la característica es *p,* la probabilidad de que no lo haga es $1 - p$. Ello implica que la densidad de la variable aleatoria discreta X_i es

X_i	0	1
$f(X_i)$	$1 - p$	p

Tabla 8.1 Resultados de un estudio de 20 personas atendidas en una clínica gratuita

Persona	¿Empleada?	Persona	¿Empleada?
1	**Sí $x_1 = 1$**	11	No $x_{11} = 0$
2	**Sí $x_2 = 1$**	12	Sí $x_{12} = 1$
3	**No $x_3 = 0$**	13	No $x_{13} = 0$
4	**No $x_4 = 0$**	14	No $x_{14} = 0$
5	**No $x_5 = 0$**	15	No $x_{15} = 0$
6	**No $x_6 = 0$**	16	No $x_{16} = 0$
7	**No $x_7 = 0$**	17	Sí $x_{17} = 1$
8	**No $x_8 = 0$**	18	Sí $x_{18} = 1$
9	**No $x_9 = 0$**	19	No $x_{19} = 0$
10	No $x_{10} = 0$	20	Sí $x_{20} = 1$

a) Utilizar el método de la Sección 4.2 para verificar que $E[X_i] = p$ $E[X_i^2] = p$ y Var $X_i = p(1 - p)$.

b) Comprobar que X, número de elementos de la muestra con la característica, viene dado por

$$X = \sum_{i=1}^{n} X_i$$

c) Supóngase que n es lo suficientemente grande para que sea aplicable el Teorema central del límite. Utilizar dicho teorema para razonar que X/n tiene una distribución aproximadamente normal con media p y varianza $p(1 - p)/n$, como afirma el Teorema 8.1.1.

d) Para muestras grandes, indicar cuál es la distribución aproximada de la variable aleatoria

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

8. *Captura-recaptura. Método para muestrear poblaciones de animales salvajes.* Un parámetro de interés para los ingenieros de montes es N, tamaño de la población de ciertas especies de animales salvajes. Por ejemplo, es importante conocer el número de osos en estado salvaje en las Great Smoky Mountains, para que sea posible proteger o controlar su población. Desgraciadamente, debido a la gran movilidad de estos animales y a la extensión del área geográfica implicada, es imposible hacer un censo. De aquí que el valor de N deba ser aproximado estadísticamente. El método de captura-recaptura consiste en capturar un determinado número de animales, marcarlos y liberarlos de nuevo, permitiendo que se dispersen. Varios días o semanas después, se extrae una segunda muestra de animales y se determina la proporción de animales marcados en ella. Se utiliza la proporción muestral como estimación de la proporción de animales marcados de la población. Sea

T = número de animales capturados y marcados originalmente

p = proporción de animales marcados en la población = $\frac{T}{N}$

C = número de animales capturados en la segunda muestra

R = número de animales vueltos a capturar
= número de animales marcados en la segunda muestra

$\hat{p}$ = proporción de animales marcados en la segunda muestra = $\frac{R}{C}$

Puesto que $\hat{p}$ estima a p, $\hat{p} \cong p$. Sustituyendo, podemos concluir que

$$\frac{R}{C} \cong \frac{T}{N}$$

Resolviendo esta ecuación para N, vemos que

$$N \cong \frac{CT}{R}$$

Por ejemplo, si originalmente se marcaron $T = 20$ osos, se capturaron $C = 15$ en la segunda muestra y, de ellos, $R = 3$ estaban marcados, entonces el número de osos que se estima que viven en estado salvaje es

$$\hat{N} = \frac{CT}{R} = \frac{15(20)}{3} = 100$$

a) Se realiza un estudio sobre halcones en el noroeste de Canadá. Se marcan treinta halcones y se capturan veinte en la segunda muestra; de éstos, sólo dos estaban marcados. Utilizar esta información para estimar *N,* número de halcones que viven en libertad en el área.

b) A mediados de la década de 1930, el cisne trompetero estaba en peligro de extinción en Norteamérica. En esta época, se designaron como refugio el Yellowstone Park y un área adyacente de fuentes termales en Red Rock Lakes, Montana. A finales de los años 60 se efectuó un estudio de los cisnes en el área. Se capturaron y marcaron cincuenta cisnes. Una segunda muestra de 30 cisnes contenía 5 que habían sido marcados. Utilizar esta información para estimar el total de la población de cisnes en la región, en esas fechas.

c) El método de captura-recaptura se emplea para hacer una estimación aproximada del número de drogadictos en la ciudad de Nueva York. Durante seis meses se registraron los drogadictos arrestados por primera vez por delitos menores. Después fueron puestos en libertad. De esta forma, 500 adictos quedaron «marcados». Durante el año siguiente se arrestaron por diversos delitos 1800 drogadictos y 20 eran del grupo de los marcados. Basándose en esta información, estimar el número total de drogadictos en la ciudad de Nueva York.

8.2. ESTIMACIÓN POR INTERVALO DE p

Volvemos ahora al problema de la estimación del intervalo de confianza de *p.* Es decir, nos gustaría extender la estimación puntual de *p* de que disponemos a un intervalo de valores, de forma que pueda señalarse un nivel de confianza. Para hacerlo así, hemos de encontrar una variable aleatoria cuya expresión contenga a *p y* cuya distribución de probabilidad se conozca, al menos aproximadamente. Esto se hace fácilmente.

A partir del Teorema 8.1.1, sabemos que $\hat{p}$ está distribuida aproximadamente como una normal con media *p y* varianza *p(1 - p)/n.* El Teorema 5.3.1 nos permite concluir que la variable aleatoria tipificada

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

sigue aproximadamente la Z o distribución normal tipificada.

Recordemos que para hallar el intervalo de confianza de la media, comenzábamos con la variable aleatoria Z.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

La variable aleatoria

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

tiene la misma estructura algebraica. Aquí, $\hat{p}$ desempeña el papel de $\bar{X}$, p corresponde μ y $\sqrt{p(1-p)/n}$ es equivalente a $\sigma/\sqrt{n}$. El argumento algebraico dado en la Sección 6.4 para obtener los límites de confianza $\bar{X} \pm z\sigma/\sqrt{n}$ pueden repetirse para obtener los límites para un intervalo de confianza de p:

$$L_1 = \hat{p} - z\sqrt{\frac{p(1-p)}{n}}$$

$$L_2 = \hat{p} + z\sqrt{\frac{p(1-p)}{n}}$$

No obstante, tenemos un problema que no se ha presentado antes. Los límites L_1 y L_2 han de ser *estadísticos*. Desgraciadamente, éste no es el caso. Según están escritos, L_1 y L_2 contienen el parámetro desconocido p. Esto significa que estamos intentando utilizar p para estimar p, ¡una situación aparentemente absurda! El problema puede abordarse de una manera natural. Es decir, podemos estimar $p(1-p)$ reemplazando p por su estimador $\hat{p}$. Recuérdese que, cuando se reemplaza la desviación típica verdadera por un estimador, la distribución cambia de Z a T_{n-1} Aquí no haremos este cambio porque, en las situaciones reales, los tamaños muestrales para estimar las proporciones son grandes. Para muestras grandes, los puntos t son aproximados por los puntos z. Por lo tanto, los límites de confianza para p vienen dados por

$$\boxed{L_1 = \hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

$$\boxed{L_2 = \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

El Ejemplo 8.2.1 ilustra el empleo de este intervalo.

Ejemplo 8.2.1. La mosca adulta del gusano tornillo es de color azul metálico y su tamaño triplica al de la mosca común. El gusano tornillo pone los huevos en las heridas de animales de sangre caliente y produce una grave infección. Se realizó un experimento con el objetivo de controlar esta población. Se expuso a las crisálidas del gusano tornillo a una dosis de radiación de 2500 rad con la esperanza de esterilizar a la mayor parte de los machos. Dado que las hembras se aparean sólo una vez, si lo hacen con un macho estéril producirán huevos estériles. Se encontró que, tras la radiación, 415 de los 500 apareamientos observados dieron como resultado huevos estériles. La estimación puntual para la proporción de apareamientos estériles producidos por esta dosis es $\hat{p} = \frac{415}{500} = 0.83$. Para construir un intervalo de confianza de p del 95 %, consideramos la partición de la curva normal tipificada de la Figura 8.3. Un intervalo de confianza del 95 % viene dado por

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.83 \pm 1.96\sqrt{\frac{0.83(0.17)}{500}}$$

$$= 0.83 \pm 0.03$$

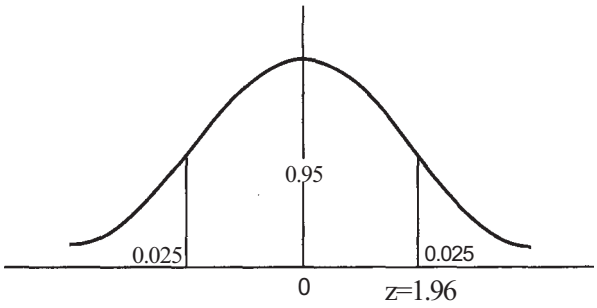


Figura 8.3. Partición de Z para obtener un intervalo de confianza de p del 95%.

Pasándolo a porcentajes, podemos tener una confianza del 95 %, aproximadamente, de que el verdadero porcentaje de apareamientos estériles resultante de este nivel de radiación esté entre el 80 y el 86 %.

EJERCICIOS 8.2

1. Utilizar los datos del Ejemplo 8.1.2 para hallar un intervalo de confianza del 95 % de la proporción de varones con cromosoma X frágil en las instituciones para retrasados mentales. ¿Será mayor un intervalo de confianza del 90 % de esta proporción que el intervalo anteriormente construido o, por el contrario, será menor?
2. Utilizar los datos del Ejercicio 2 de la Sección 8.1 para obtener un intervalo de confianza del 99 % de la proporción de individuos de 16 años o menos que fuman regularmente. ¿Le sorprendería leer un artículo que diga que esta proporción es de 0.23? Explicarlo.
3. Al realizar un recuento de leucocitos, se extiende una gota de sangre en una capa fina y uniforme sobre un portaobjetos, se tiñe con tinción de Wright, y se examina con el microscopio. De los 200 leucocitos contabilizados, 125 fueron neutrófilos, unos leucocitos que se producen en la médula ósea y cuya función, en parte, es la de eliminar agentes infecciosos de la sangre.
   a) Encontrar una estimación puntual para *p,* proporción de neutrófilos hallados entre los leucocitos del individuo.
   b) Obtener un intervalo de confianza del 90 % para *p.*
   c) En un individuo de salud normal, el porcentaje de neutrófilos entre los leucocitos es del 60 al 70 %. Basándose en el intervalo obtenido en el apartado *b,* ¿hay signos claros de un desequilibrio de neutrófilos en el individuo? Razonar la respuesta.
4. Una defensa que ciertos organismos tienen contra los depredadores es la de hacerse repugnantes, es decir, adquirir un sabor, olor, o lanzar un líquido desagradables. Un insecto con esta característica es el insecto palo. De 50 arrendajos azules a los que un insecto palo lanzó el líquido, 42 se mantuvieron a distancia del insecto durante al menos dos semanas. Utilizar esta información para hallar un intervalo de confianza del 94 % de *p,* proporción de arrendajos azules que entraron en sucesivos contactos con el insecto palo durante dos semanas.
5. En 1987 se reforestaron más de 3 millones de acres con dos mil millones de plantas de vivero. Una grave sequía durante la siguiente estación mató muchas de estas plantas. Se obtuvo una muestra de 1000 plantas y se descubrió que 300 estaban muertas. Obtener un intervalo de confianza del 90 % de la proporción de plantas de vivero muertas. Utilizar esta información para estimar el número de plantas de vivero muertas en la población. (Basado en la información hallada en Howard Burnett, «A Report on Our Stressed-Out Forests», *American Forests,* abril 1989, págs. 21-25.)

6. Se ha realizado un estudio sobre niños que padecen de dolor en el pecho. Se ha hallado que, de 137 niños que tenían dolor en el pecho, 100 mostraban radiografías normales. Obtener un intervalo de confianza del 95 % de la proporción de niños con dolor en el pecho que mostraron radiografías de tórax normales. (Basado en la información hallada en Steven Selbst, Richard Ruddy y B. J. Clark, «Chest Pain in Children», *Clinical Pediatrics,* vol. 29, núm. 7, julio de 1990, págs. 374-377.)
7. Considérese el estudio del Ejercicio 6. De 191 niños con dolor en el pecho, 160 presentan un electrocardiograma normal. Obtener un intervalo de confianza del 95 % de la proporción de niños con dolor en el pecho que tienen un electrocardiograma normal.
8. Utilizar los datos del Ejercicio 3 de la Sección 8.1 para obtener un intervalo de confianza del 95 % sobre la proporción de muertes debidas a degeneración miofibrilar por agresión. ¿Es n lo suficientemente grande para hacerse una buena idea del valor de p a partir de este intervalo?

8.3. TAMAÑO MUESTRAL PARA LA ESTIMACIÓN DE *p*

Como anteriormente, la cuestión del tamaño de la muestra debería ser considerada en los estadios de planificación del experimento. El Ejemplo 8.3.1 ilustra el problema que puede surgir en un experimento no planificado.

Ejemplo 8.3.1. El más novedoso avance en el tratamiento del acné es un fármaco llamado ácido *cis*- 13-retinoico. En un reciente estudio, se probó este fármaco en 14 pacientes afectados de un acné bastante grave. En 13 de estos pacientes se produjo una limpieza radical de sus lesiones activas. Para construir un intervalo de confianza de p del 99 %, proporción de pacientes sobre los que el fármaco será eficaz, se necesita la partición de la curva normal tipificada dibujada en la Figura 8.4. El intervalo de confianza viene dado por

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

donde $\hat{p} = \frac{13}{14} = 0.93$. E n este caso, los límites d e confianza $0.93 \pm 2.575\sqrt{0.93(0.07)/14}$ ó bien 0.93 ± 0.18. Podemos tener un 99 % de confianza en que el fármaco será eficaz en el 75 % de aquellos pacientes sobre los que se aplique.

Interesa hacer una puntualización con respecto al Ejemplo 8.3.1. Es obvio que el intervalo obtenido no es muy informativo. Es demasiado largo para dar al experimentador un claro indicio del valor efectivo de p. El problema es consecuencia de dos factores: el grado de confianza requerido es muy alto (99 %) y el tamaño de la muestra es demasiado pequeño. Podemos corregirlo, por tanto, reduciendo el nivel de confianza o aumentando el tamaño de la muestra, o haciendo ambas cosas a la vez.

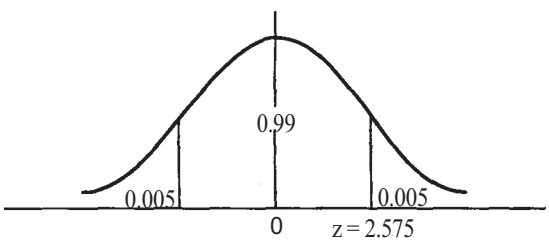


Figura 8.4. Partición de Z para obtener un intervalo de confianza de p del 99%.

Esto plantea otra cuestión importante. ¿De qué tamaño debería seleccionarse la muestra para que p estuviera dentro de una distancia específica d de p, con un grado de confianza establecido? Hay dos formas de responder a esta pregunta. La primera es aplicable cuando disponemos de una estimación de p basada en algún experimento previo. Este método se explica esquemáticamente en la Figura 8.5.

Puesto que se espera que p caiga dentro del intervalo de confianza, p y p están a una distancia máxima d, donde d viene dado por

$$d = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Esta ecuación se resuelve para n:

$$d = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$d^2 = z^2\frac{\hat{p}(1-\hat{p})}{n}$$

$$n = \frac{z^2\,\hat{p}(1-\hat{p})}{d^2}$$

De este modo, obtenemos la siguiente fórmula para calcular el tamaño de la muestra necesario para estimar p, con un grado de precisión y confianza establecidos, cuando se dispone de una estimación previa de p:

$$n \cong \frac{z^2\,\hat{p}(1-\hat{p})}{d^2} \quad \text{(estimación previa disponible)}$$

El empleo de esta fórmula se detalla en el Ejemplo 8.3.2.

Ejemplo 8.3.2. Si deseamos hacer pruebas posteriores con el fármaco que combate el acné del Ejemplo 8.3.1, ¿qué tamaño deberá tener la muestra para estimar p con $d = 0.02$ y una confianza del 90 %? Puesto que disponemos de una estimación previa $p = 0.93$, podemos aplicar la fórmula ya obtenida. El punto z de la fórmula es el mismo que el requerido para

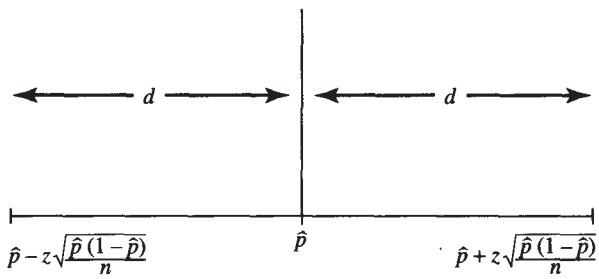


Figura 8.5. Intervalo de confianza de p.

construir un intervalo de confianza de p del 90 %. Este punto se señala en la Figura 8.6. Así, el tamaño de la muestra será

$$n = \frac{z^2 \hat{p}(1-\hat{p})}{d^2} = \frac{(1.645)^2(0.93)(0.07)}{(0.02)^2}$$
$$\cong 441$$

El segundo método de determinación del tamaño de la muestra para estimar proporciones se basa en el hecho de que el término $\hat{p}(1-\hat{p})$ nunca puede ser mayor que $\frac{1}{4}$, independientemente del valor de $\hat{p}$. Esto puede verificarse utilizando técnicas de cálculo infinitesimal (Ejercicio 7). Si no se dispone de una estimación previa de p, puede hallarse un tamaño muestral adecuado sustituyendo $\hat{p}(1-\hat{p})$ por $\frac{1}{4}$, en la ecuación anterior. De este modo, la fórmula para hallar el tamaño de la muestra necesario para estimar p con un determinado grado de precisión y confianza, cuando *no* se dispone de una estimación previa de p, es

$$n \cong \frac{z^2}{4d^2} \quad \text{(sin estimación previa disponible)}$$

El uso de esta fórmula se explica en el Ejemplo 8.3.3.

Ejemplo 8.3.3. Los hematíes normales de la sangre de los seres humanos tienen forma de discos bicóncavos. En ocasiones, la hemoglobina, una proteína que se combina fácilmente con el oxígeno, está formada en la célula de una manera imperfecta. Un tipo de hemoglobina imperfecta hace que las células presenten un hundimiento o un aspecto en «forma de hoz». Tales células «falciformes» son menos eficaces para transportar el oxígeno que las normales y causan un déficit de oxígeno llamado anemia falciforme o drepanocítica. Esta afección predomina significativamente entre los individuos de raza negra. Se efectúa un estudio para estimar el porcentaje de individuos de raza negra afectados en Virginia. ¿De qué tamaño debería elegirse una muestra para estimar este porcentaje, a una distancia de 1 punto porcentual, con una confianza del 98 %? Se supone que no se dispone de una estimación previa de p. El punto z buscado aparece en la Figura 8.7. Puesto que un punto porcentual es 0.01, el tamaño deseado de la muestra es

$$n \cong \frac{z^2}{4d^2} = \frac{(2.33)^2}{4(0.01)^2} = 13\,573$$

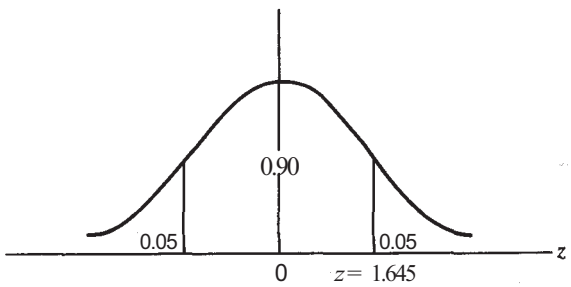


Figura 8.6. Punto necesario para obtener un intervalo de confianza de p del 90%.

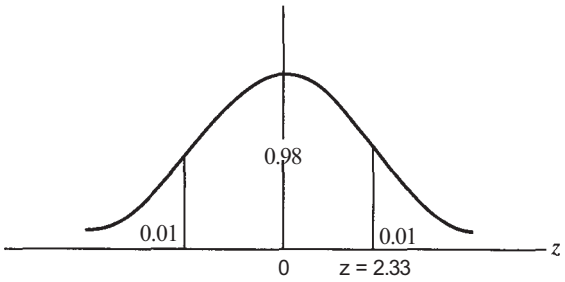


Figura 8.7. Punto necesario para obtener un intervalo de confianza de *p* del 98%.

EJERCICIOS 8.3

1. Cuando se realizó el estudio del Ejemplo 8.3.3, de los 13 573 individuos de raza negra muestreados, se encontró que 1085 tenían anemia falciforme. Utilizar esta información para hallar un intervalo de confianza del 98 % del porcentaje de individuos de raza negra en el estado de padecer la enfermedad.
2. ¿Qué tamaño muestral se requeriría para estimar la proporción de muertes debidas a degeneración miofibrilar en casos de agresión, a la distancia de 0.02 y con una confianza del 95 % (Utilizar los datos del Ejercicio 3 de la Sección 8.1 para obtener una estimación previa de *p).*
3. El estampido sónico es un problema asociado con la utilización del transporte supersónico (SST). A finales de los años 60 y principios de los 70, se realizaron pruebas preliminares sobre Oklahoma City, St. Louis y otros lugares. Después de hacer las pruebas, se efectuó un estudio para estimar el porcentaje de personas que pensaban que no podrían vivir con estampidos sónicos. ¿De qué tamaño debería haberse elegido la muestra para estimar el porcentaje a una distancia de 3 puntos porcentuales, con una confianza del 94 %?
4. La agencia de Protección del Medio Ambiente identificó recientemente en Estados Unidos 30 000 vertederos de basura considerados al menos potencialmente peligrosos. ¿Qué tamaño muestral se necesita para estimar el porcentaje de estos lugares que suponen una amenaza para la salud, a una distancia de 2 puntos porcentuales y con una confianza del 90%?
5. ¿Qué tamaño debería tener la muestra para estimar la proporción de arrendajos azules que evitarán tener contacto con el insecto palo durante dos semanas, con un error máximo de 3 % y una confianza del 94 %? (*Sugerencia:* Utilizar los datos del Ejercicio 4, Sección 8.2, para obtener una estimación de *p.)*
6. ¿Cuántos pacientes deben formar la muestra, con el fin de estimar la proporción de niños con dolor de pecho que presentaban radiografías de tórax normales? *(Sugerencia:* Utilizar los datos del Ejercicio 6, Sección 8.2, como estudio piloto.)
7. Se ha afirmado que $p(1 - p) < \frac{1}{4}$. Considérese la función $g(p) = p(1 - p)$.
   a) Hallar la derivada de *g.*
   b) Calcular el valor máximo de $g(p)$ haciendo $g'(p) = 0$ y resolviendo para *p.*
   c) Utilizar el criterio de la segunda derivada para comprobar que $p = \frac{1}{2}$ hace máxima la función $g(p)$.

8.4. CONTRASTE DE HIPÓTESIS SOBRE p

Consideramos ahora el problema de contrastar hipótesis sobre una proporción *p.* Esto implica que se ha propuesto un valor de *p* previamente a la realización del estudio. El objetivo del

experimento es obtener evidencia estadística que apoye o rechace este valor. Las hipótesis a comprobar pueden adoptar una cualquiera de tres formas, dependiendo del propósito del estudio. Designemos p_0 al valor hipotético de p. Estas formas son las siguientes:

I	II	III
$H_0: p \leq p_0$	$H_0: p \geq p_0$	$H_0: p = p_0$
$H_1: p > p_0$	$H_1: p < p_0$	$H_1: p \neq p_0$
Contraste con cola a la derecha	Contraste con cola a la izquierda	Contraste con dos colas

Ejemplo 8.4.1. Una de las teorías del aprendizaje que ha sido causa de gran controversia es la de «transferencia de enseñanza por canibalismo». En un estudio para ganar consistencia estadística para esta teoría, se adiestró a un grupo de planarias para que evitasen un *shock* eléctrico. Luego fueron trituradas y suministradas como alimento a un grupo de 100 planarias no adiestradas. Si no se transfiere la enseñanza a las planarias no tratadas, se admite que la probabilidad de que una planaria tratada pueda evitar el *shock* es de $\frac{1}{2}$; si se transfiere, la probabilidad es mayor que $\frac{1}{2}$. Puesto que el propósito de este estudio es conseguir apoyo estadístico para la teoría, el supuesto $p > \frac{1}{2}$ se transforma en la hipótesis alternativa. De esta forma estamos contrastando

$$H_0: p \leq \frac{1}{2}$$

$$H_1: p > \frac{1}{2}$$

Ejemplo 8.4.2. Hasta muy recientemente, p, la tasa de mortalidad causada por una infección vírica del cerebro altamente mortal, la encefalitis producida por el virus del herpes simple, ha sido del 70 %. Se realiza un estudio para probar un nuevo fármaco, la vidarabina, para utilizarlo en el tratamiento de la enfermedad. Puesto que se espera que la vidarabina reduzca la tasa de mortalidad, este supuesto se transforma en la hipótesis alternativa. Es decir, estamos contrastando

$$H_0: p \geq 0.70$$
$$H_1: p < 0.70$$

Para contrastar hipótesis sobre p, ha de desarrollarse un estadístico. El estadístico deberá ser lógico. Además, para encontrar el valor P del contraste, debe conocerse, al menos aproximadamente, su distribución de probabilidad, bajo el supuesto de que la hipótesis nula es cierta. De nuevo, el estadístico elegido es el mismo que la variable aleatoria utilizada para generar los límites de confianza para p. En particular, servirá

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Este estadístico es lógico, porque básicamente compara el valor estimado para p, es decir, $\hat{p} = X/n$, con el valor hipotético de p, es decir, p_0. Si sus valores son próximos, indicando que H_0 es cierta, el valor observado del estadístico será cercano a cero. Si $\hat{p}$ y p_0 difieren notable-

mente, lo que indica que H_0 es falsa, entonces el valor observado del estadístico será o muy grande con valor positivo, o muy grande con valor negativo. Para muestras grandes, el estadístico tiene una distribución que es aproximadamente normal tipificada.

Ejemplo 8.4.3. La teoría del aprendizaje por canibalismo se confirma contrastando

$$H_0: p \leq 0.5$$
$$H_1: p > 0.5 \quad \text{(el canibalismo aumenta la probabilidad de evitar el } \textit{shock)}$$

Cuando se realizó el experimento descrito en el Ejemplo 8.4.1,57 de las 100 planadas puestos a prueba rehuyeron el *shock*. El valor observado del estadístico es

$$\frac{x/n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\frac{57}{100} - 0.5}{\sqrt{0.5(0.5)/100}} = 1.4$$

Dado que el contraste es de cola a la derecha, el valor P viene dado por

$$P = P[Z \geq 1.4]$$

Este valor P se ilustra en la Figura 8.8. Puede hallarse a partir de la tabla de Z, Tabla III del Apéndice B. En este caso,

$$\begin{aligned} P = P[Z \geq 1.4] &= 1 - P[Z \leq 1.4] \\ &= 1 - 0.9192 \\ &= 0.0808 \end{aligned}$$

El método para los contrastes de hipótesis relativos a p indicados aquí supone que los tamaños muestrales son lo suficientemente grandes para poder aplicar el Teorema central del límite. (Véase Teorema 6.2.2.) Como se ha indicado en la Sección 6.2, las muestras pequeñas con un tamaño alrededor de 25 son generalmente buenas. Recuerde que la potencia se ve afectada por el tamaño de la muestra. Generalmente, es difícil detectar diferencias pequeñas, pero quizás importantes, entre p y p_0 si las muestras son pequeñas. Sin embargo, en ciertas ocasiones, especialmente en medicina, las muestras son pequeñas por necesidad. Por ejemplo, al estudiar una afección rara, quizá sólo existan 8 ó 10 casos conocidos. El investigador debe hacer todo lo que pueda con los recursos disponibles. En el Capítulo 13 se presenta un contraste basado en la distribución binomial adecuado para muestras pequeñas.

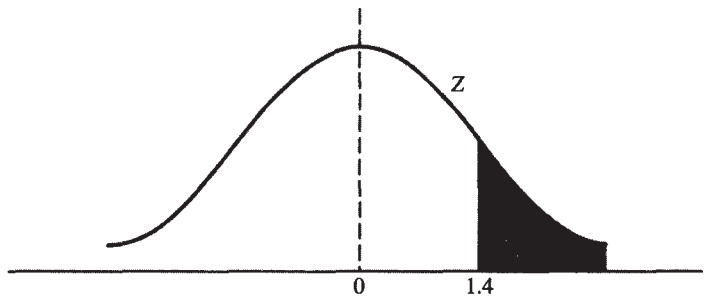


Figura 8.8. Valor P para el contraste con cola a la derecha del Ejemplo 8.4.3. $P = P[Z \geq 1.4] = 0.0808$.

EJERCICIOS 8.4

1. Los que se oponen a la construcción de una presa en el New River pretenden que la mayor parte de los residentes que viven a lo largo del río también se oponen a su construcción. Se realiza un estudio para conseguir apoyo para este punto de vista.
   a) Construir las hipótesis nula y alternativa apropiadas.
   b) De 500 personas estudiadas, 270 se opusieron a la construcción. ¿Es suficiente evidencia para afirmar que una mayoría de residentes se opone? Explicar la respuesta basándose en el valor *P* observado del estadístico. ¿A qué tipo de error se está expuesto? Comentar las consecuencias prácticas de cometer tal error.
2. Se está probando un nuevo tipo de trampa japonesa para escarabajos. El fabricante asegura que su trampa atrae y mata a más del 90 % de los escarabajos que llegan a una distancia de 30 pies de la trampa. Se realiza un experimento para apoyar tal afirmación.
   a) Construir las hipótesis nula y alternativa apropiadas.
   b) El experimento se efectúa dejando 900 escarabajos cerca de la trampa. Si es cierta, ¿cuál es el número máximo de escarabajos que se esperaría que fueran atraídos por la trampa? De los 900, fueron atraídos 825 por la trampa y murieron. ¿Es esta evidencia suficiente para sostener la pretensión? Explicarlo basándose en el valor *P* del contraste.
   c) ¿A qué tipo de error se está ahora expuesto? Comentar las consecuencias de cometer tal error.
3. Considérese el Ejemplo 8.4.2. De 50 sujetos en los que se probó la vidarabina, 14 murieron. Encontrar el valor *P* para este contraste.
4. Entre los pacientes con cáncer de pulmón, el 90% o más muere generalmente en el espacio de tres años. Como resultado de nuevas formas de tratamiento, se cree que esta tasa se ha reducido.
   a) Construir las hipótesis nula y alternativa necesarias para sostener este argumento.
   b) En un reciente estudio sobre 150 pacientes diagnosticados de cáncer de pulmón, 128 murieron en el espacio de tres años. ¿Puede ser rechazada H_0 al nivel $\alpha = 0.1$? Al nivel $\alpha = 0.05$, ¿piensa que hay pruebas suficientes para pretender que los nuevos métodos de tratamiento son más eficaces que los antiguos? Explicar la respuesta.
5. El método usual para tratar la leucemia mieloblástica aguda consiste en someter al paciente a quimioterapia intensiva en el momento del diagnóstico. Históricamente, esto ha producido una tasa de remisión del 70 %. Estudiando un nuevo método de tratamiento, se utilizaron 50 voluntarios. ¿Cuántos de los pacientes deberían haber remitido para que los investigadores pudiesen afirmar al nivel a = 0.025, que el nuevo método produce remisiones más altas que el antiguo?
6. La grave sequía de 1987 afectó tanto a la tasa de mortalidad de las plantas de vivero como a la tasa de crecimiento de los árboles ya establecidos. Se piensa que la mayor parte de los árboles de las zonas afectadas tienen un anillo de crecimiento de 1987 menor a la mitad de los anillos de crecimiento de los demás. Una muestra de 250 árboles, ha dado 150 con esta característica. ¿Apoyan los datos esta idea? Explicarlo. (Basado en la información hallada en Howard Burnett, «A Report on Our Stressed-Out Forests», *American Forests,* abril 1989, págs. 21-25.)
7. Se cree que más del 85 % de todos los niños con dolor torácico presentará, no obstante, un ecocardiograma normal. Una muestra de 139 de estos niños ha dado 123 con ecocardiogramas normales.
   a) Establecer la hipótesis de investigación.
   b) Hallar una estimación puntual para la proporción de niños con dolor torácico que presentan ecocardiogramas normales. Basándose en esta estimación, ¿piensa usted que la hipótesis de investigación se sostendrá después del contraste?

c) Contrastar H_0: $p \leq 0.85$ frente a H_1: $p > 0.85$. ¿Se puede rechazar H_0 con un nive $\alpha = 0.10$?

d) Explicar cualquier aparente contradicción entre los resultados de los apartados *b y c.* (Basado en la información hallada en Steven Selbst, Richard Ruddy y B.J. Clark, «Chest Pain in Children», *Clinical Pediatrics,* vol. 29, núm. 7, julio de 1990, 374-377.)

8. *Potencia.* Supongamos que queremos contrastar

$$H_0: p = 0.5$$

Acordamos rechazar H_0 si el valor observado del estadístico Z es superior a 1.96. Puesto que $P[Z > 1.96] = 0.025$, estamos prefijando a en 0.025. Recordemos que la potencia viene dada por

$$\text{Potencia} = \text{P[rechaza } H_0 \mid H_1 \text{ es verdadera]}$$

En este caso,

$$\text{Potencia} = P[Z > 1.96 \mid p = 0.6]$$

Por ejemplo, supongamos que $n = 20$. ¿Cuál es la potencia del contraste? En este caso

$$\text{Potencia} = P\left[\frac{\hat{p} - 0.5}{\sqrt{0.5(0.5)/20}} > 1.96 \mid p = 0.6\right]$$

Primero despejamos $\hat{p}$ de la forma siguiente:

$$\begin{aligned}\text{Potencia} &= P\left[\hat{p} - 0.5 > 1.96\sqrt{\frac{0.5(0.5)}{20}} \mid p = 0.6\right] \\ &= P\left[\hat{p} > 0.5 + 1.96\sqrt{\frac{0.5(0.5)}{20}} \mid p = 0.6\right] \\ &= P\left[\hat{p} > 0.72 \mid p = 0.6\right]\end{aligned}$$

Si $p = 0.6$, podemos tipificar $\hat{p}$ restando su media de 0.6 y dividiéndolo por su desviación típica, $\sqrt{0.6(0.4)/20}$. Vemos que

$$\text{Potencia} = P\left[\frac{\hat{p} - 0.6}{\sqrt{0.6(0.4)/20}} > \frac{0.72 - 0.6}{\sqrt{0.6(0.4)/20}} \mid p = 0.6\right]$$

Si $p = 0.6$, la variable aleatoria a la izquierda de esta desigualdad sigue una distribución normal tipificada. Por lo tanto,

$$\begin{aligned}\text{Potencia} = P[Z > 1.09] &= 1 - P[Z \leq 1.09] \\ &= 1 - 0.8621 \\ &= 0.1379\end{aligned}$$

Obsérvese que una muestra de tamaño 20 deja muy pocas posibilidades de detectar la diferencia entre las proporciones 0.5 y 0.6. Para ver el efecto que tiene el tamaño de la muestra sobre la potencia, hallar la potencia del contraste anterior para cada una de las muestras de los apartados a a c.

a) $n = 50$
b) n = 200
c) $n = 1000$
d) Si cambiáramos el contraste de forma que $\alpha = 0.1$, ¿aumentará la potencia, disminuirá o permanecerá igual? Explicarlo.

8.5. COMPARACIÓN DE DOS PROPORCIONES: ESTIMACIÓN

En los estudios médicos y biológicos surge frecuentemente el problema de comparar dos proporciones. La situación general puede describirse así: hay dos poblaciones de interés; en cada población, se estudia el mismo rasgo; cada miembro de cada población puede clasificarse en función de que lo tenga o no, y en cada población es conocida la proporción de los que lo tienen. Se harán inferencias sobre p_1, p_2 y $p_1 - p_2$, donde p_1 y p_2 son las proporciones de individuos que presentan el rasgo en la primera y segunda poblaciones, respectivamente. (Véase Fig. 8.9.)

> **Ejemplo 8.5.1**. Anualmente, la insuficiencia renal amenaza la vida de muchas personas. Se realiza un estudio en pacientes renales para comparar la tasa de insuficiencia renal entre los tratados con el esteroide prednisona y los que reciben un placebo. Aquí, las dos poblaciones de interés son los pacientes renales tratados con el fármaco y los que no lo reciben. El rasgo bajo estudio en cada caso es el de sufrir insuficiencia renal. (Véase Fig. 8.10.)

El problema de la estimación puntual de la diferencia entre las dos proporciones se resuelve de manera obvia. Simplemente estimamos p_1 y p_2 individualmente, y después tomamos $P_1 - P2$ como estimación para la diferencia entre los dos. Es decir,

$$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

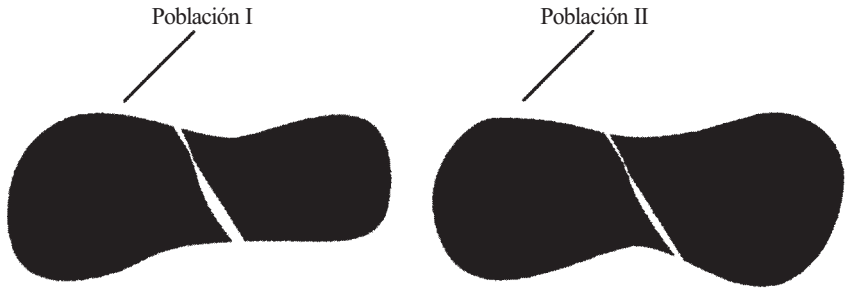


Figura 8.9. Comparación de dos proporciones. ¿Cuál es la diferencia entre p_1 y p_2?

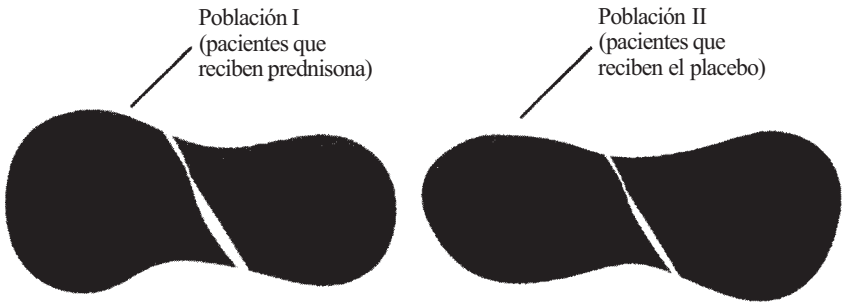


Figura 8.10. $p_1 - p_2 = ?$

donde n_1 y n_2 son los tamaños de las muestras extraídas de las dos poblaciones, y X_1 y X_2 indican, respectivamente, el número de objetos en las muestras que poseen el rasgo.

Ejemplo 8.5.2. En un estudio sobre el uso de la prednisona en el tratamiento de pacientes renales, se utilizaron 72 sujetos en 19 hospitales. De los 34 pacientes tratados con prednisona, sólo uno sufrió insuficiencia renal. Sin embargo, de los 38 que recibieron un placebo, se produjo insuficiencia renal en 10. Basándose en este estudio, $\hat{p}_1 = \frac{1}{34} \cong 0.03$ y $\hat{p}_2 = \frac{10}{38} \cong 0.26$. La diferencia estimada entre las dos proporciones poblacionales es

$$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = 0.03 - 0.26 = -0.23$$

Intervalo de confianza de la diferencia de dos proporciones

Para extender el estimador puntual $\hat{p}_1 - \hat{p}_2$ a un intervalo, debemos considerar la distribución de probabilidad de esta variable aleatoria. Su distribución aproximada se establece en el Teorema 8.5.1. Este teorema puede probarse parcialmente utilizando las leyes para la esperanza matemática y la varianza del Apéndice A. La demostración se presenta en el Ejercicio 12, al final de esta sección.

Teorema 8.5.1. Para muestras de gran tamaño, el estimador $\hat{p}_1 - \hat{p}_2$ es aproximadamente normal con media $p_1 - p_2$ y varianza

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Para construir un intervalo de confianza para $p_1 - p_2$, hemos de encontrar una variable aleatoria cuya expresión contenga este parámetro y cuya distribución de probabilidad sea conocida al menos aproximadamente. Esto es ahora fácil de hacer. Simplemente, tipificamos la variable $\hat{p}_1 - \hat{p}_2$ para concluir que la variable aleatoria

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

es aproximadamente normal *tipificada.* Para no repetir un argumento algebraico dado previamente, consideraremos tres intervalos que ya han sido obtenidos y señalaremos sus semejanzas:

Parámetro estimado	**Obtenido con**	**Distribución**	**Límites**
μ (σ^2 conocida)	$\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	Z	$\bar{X} \pm z\dfrac{\sigma}{\sqrt{n}}$
μ (σ^2 desconocida)	$\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$	T	$\bar{X} \pm t\dfrac{S}{\sqrt{n}}$
p	$\dfrac{\hat{p}-p}{\sqrt{p(1-p)/n}}$	$\sim Z$ (aproximadamente Z)	$\hat{p} \pm z\sqrt{\dfrac{p(1-p)}{n}}$

La estructura algebraica de cada una de las variables es la misma y tiene la forma

$$\frac{\text{Estimador - parámetro}}{D}$$

donde D es la desviación típica o el estimador para la desviación típica del estimador que hay en el numerador. Ésta es también la forma algebraica que toma la variable

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim Z$$

Los límites de confianza en el caso previo eran

$$\text{Estimador} \pm \text{probabilidad puntual x } D$$

Aplicándolo a la variable aleatoria anterior, los límites de confianza propuestos para un intervalo de $p_1 - p_2$ son

$$(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

De nuevo hay un pequeño problema. Los límites propuestos no son *estadísticos.* Incluyen las proporciones poblacionales desconocidas p_1 y p_2. Como en el caso de una muestra, este problema puede obviarse reemplazando las proporciones poblacionales por sus estimadores $\hat{p}_1$ y $\hat{p}_2$. Así, se obtiene la siguiente fórmula para hallar los intervalos de confianza para la diferencia entre dos proporciones poblacionales:

$$\boxed{(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Esta fórmula se explica en el Ejemplo 8.5.3.

Ejemplo 8.5.3. Para construir un intervalo de confianza del 95 % de la diferencia en la tasa de insuficiencia renal entre los que reciben prednisona y los que no reciben el fármaco, se necesita la partición de la curva normal tipificada dibujada en la Figura 8.3. Del Ejemplo 8.5.2, $n_1 = 34$, $n_2 = 38$, $\hat{p}_1 = 0.0$ $\hat{p}_2 = 0.26$ y $\hat{p}_1 - \hat{p}_2 = -0.23$. Los límites de confianza buscados son

$$(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = -0.23 \pm 1.96\sqrt{\frac{0.03(0.97)}{34} + \frac{0.26(0.74)}{38}}$$

$$= -0.23 \pm 0.15$$

Podemos tener un 95 % de confianza en que la diferencia de tasas en los procesos está entre -38 % y -8 %. Obsérvese que el cero no está contenido en este intervalo. Esto es importante, ya que podemos interpretar que ello significa que tenemos un 95 % de confianza en que las dos tasas, de hecho, son diferentes. Puesto que ambos límites son negativos, es posible, además, inferir que la tasa de insuficiencia en aquellos a los que se suministra el fármaco es menor que en aquellos a los que no se les suministra en, al menos, un 8 %.

EJERCICIOS 8.5

1. El fármaco Anturane, disponible en el mercado desde 1959 para el tratamiento de la gota, está siendo estudiado para utilizarlo en la prevención de muertes súbitas por una segunda crisis cardíaca, entre pacientes que ya sufrieron un primer ataque. En el estudio, 733 pacientes recibieron Anturane y a 742 se les dio un placebo. Después de ocho meses, se observó que de 42 muertes por un segundo ataque al corazón, 29 se produjeron en el grupo del placebo y 13 en el grupo de Anturane. Utilizar estos datos para estimar la diferencia en el porcentaje de muertes súbitas entre los pacientes tratados con Anturane y aquellos que no reciben el fármaco.
2. Se está probando un antibiótico llamado doxiciclina para prevenir la «diarrea del viajero». El fármaco fue probado en 38 voluntarios del Cuerpo de Paz que fueron a Kenya. A la mitad se les dio doxiciclina y a la otra mitad una dosis ficticia. De los que recibieron doxiciclina, 17 se libraron del trastorno, mientras que solamente 11 de los del otro grupo se libraron. Encontrar una estimación puntual para la diferencia de las tasas de protección entre los que utilizan doxiciclina y los que no la utilizan.
3. Uno de los ejemplos de selección natural estudiados es el de la polilla moteada. Hasta 1845, todas las especies conocidas presentaban colores claros, pero ese año fue capturada en Manchester una polilla negra. A causa de la industrialización en la zona, los troncos de los árboles, las rocas, e incluso la tierra, se habían ennegrecido por el hollín Esta forma muíante negra se extendió rápidamente. H. B. D. Kettlewell creyó que 1a expansión era debida, en parte, a que el color negro protege a la polilla de sus depredadores naturales, en particular de los pájaros. Los entomólogos de la época declararon que ellos nunca habían visto a un pájaro comerse una polilla moteada de color alguno y desecharon la idea. En un experimento para estudiar la teoría, Kettlewell marcó una muestra de 100 polillas de cada color y las liberó después. Volvió por la noche con trampas con luz y recobró el 40 % de las polillas negras y solamente el 19 % de las de color claro. Supóngase que las polillas no recobradas fueron presas de algún depredador. Encontrar una estimación puntual para la diferencia en las tasas de supervivencia.
4. Construir un intervalo de confianza del 95 % para la diferencia en el porcentaje de las muertes súbitas entre los usuarios de Anturane y los pacientes que no utilizan el fárma-

co, basándose en los datos del Ejercicio 1. Si se construyó un intervalo de confianza del 90 % a partir de los mismos datos, ¿qué intervalo sería más largo? ¿Por qué? Verificar la respuesta. ¿Hay más evidencia de que la tasa de mortalidad por el segundo ataque es más baja entre los pacientes tratados con el fármaco que entre los que no lo han sido? Explicar la respuesta.

5. Utilizando los datos del Ejercicio 2, construir un intervalo de confianza del 90 % para la diferencia entre las tasas de protección entre aquellos que utilizan doxiciclina y los que no la utilizan. ¿Hay más evidencia de que la doxiciclina contribuye a proporcionar protección contra la «diarrea del viajero»? Explicar la respuesta.
6. Utilizar los datos del Ejercicio 3 para construir un intervalo de confianza del 98 % para la diferencia entre las tasas de supervivencia de las polillas negras y las de color claro, en la región de Manchester. ¿Apoya el intervalo la teoría de que el color negro contribuye a proteger a estas polillas de los depredadores? Explicar la respuesta.
7. Se ha realizado un estudio sobre la tasa de supervivencia de los pájaros adultos en los trópicos y en las zonas templadas. Inicialmente, se marcaron 500 pájaros adultos con cintas en las patas y se liberaron en Panamá, una región tropical. Un año después, se volvieron a capturar 445. Si suponemos que los que no se recuperaron fueron víctimas de un depredador, la tasa de supervivencia estimada de 1 año para los pájaros adultos en la región es $\hat{p}_1 = \frac{445}{500} = 0.89$. Un experimento similar en Illinois, en la zona templada, dio como resultado una recuperación de 252 de los 500 pájaros con una tasa de supervivencia estimada de aproximadamente 0.504. Hallar un intervalo de confianza del 90 % de la diferencia en las tasas de supervivencia de un año para las dos regiones.
8. Utilizar un argumento similar al de la Sección 8.3 para demostrar que el tamaño de muestra común $n = n_1 = n_2$ necesario para estimar la diferencia de proporciones en menos de una cantidad d, con un grado de confianza estipulado, viene dado por

$$n = \begin{cases} \dfrac{z^2[\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)]}{d^2} & \text{si existen estimaciones previas de } p_1 \text{ y } p_2 \text{ disponibles} \\ \dfrac{z^2}{2d^2} & \text{si no existen estimaciones previas disponibles} \end{cases}$$

9. Utilizar los datos del Ejercicio 1 como estudio piloto para determinar el tamaño de muestra común necesario para estimar la diferencia en el porcentaje de muertes súbitas entre los usuarios de Anturane y entre pacientes que no consumen este medicamento, hasta una distancia de 2 puntos porcentuales y con una confianza del 95 %. Resolver de nuevo el problema con una confianza del 90 %.
10. Utilizar los datos del Ejercicio 3 como estudio piloto para determinar el tamaño de muestra común necesario para estimar la diferencia de las tasas de supervivencia entre los dos grupos de polillas, a una distancia de hasta 3 puntos porcentuales y con una confianza del 98 %.
11. Debe realizarse un estudio para comparar el porcentaje de consumidores de drogas por vía intravenosa entre personas que son VIH positivos (con riesgo de desarrollar el SIDA) con el porcentaje de consumidores de drogas por vía intravenosa entre personas que son VIH negativos. ¿Qué tamaño de muestra común se debe elegir para estimar la diferencia en porcentajes, con una distancia de hasta 2 puntos porcentuales y una con-

fianza del 95 %? Comparar esta respuesta con la obtenida en el Ejercicio 9. ¿Representa una ventaja práctica el tener estimaciones previas de p_1 y p_2?

12. Se puede demostrar que la suma o diferencia de dos variables aleatorias independientes normalmente distribuidas también está distribuida normalmente. Puesto que $\hat{p}_1$ y $\hat{p}_2$ se basan en muestras aleatorias extraídas independientemente de poblaciones separadas, estas variables aleatorias con distribuciones aproximadamente normales son independientes. Por el resultado anterior, su diferencia también está distribuida normalmente.
   a) Hallar $E[\hat{p}_1 - \hat{p}_2]$.
   b) Utilizar las propiedades de la varianza para hallar $\text{Var}[\hat{p}_1 - \hat{p}_2]$.

8.6. COMPARACIÓN DE DOS PROPORCIONES: CONTRASTE DE HIPÓTESIS

Frecuentemente surgen problemas en los que se afirma, antes de llevar a cabo el experimento, que una proporción o porcentaje difiere de otro en una cantidad específica. La finalidad del experimento es obtener apoyo para dicha afirmación. Las hipótesis toman una cualquiera de las tres formas siguientes, donde $(p_1 - p_2)_0$ representa el valor asignado por hipótesis a la diferencia de proporciones:

I	II
$H_0: p_1 - p_2 \leq (p_1 - p_2)_0$	$H_0: p_1 - p_2 \geq (p_1 - p_2)_0$
$H_1: p_1 - p_2 > (p_1 - p_2)_0$	$H_1: p_1 - p_2 < (p_1 - p_2)_0$
Contraste con la cola a la derecha	Contraste con la cola a la izquierda

III
$H_0: p_1 - p_2 = (p_1 - p_2)_0$
$H_1: p_1 - p_2 \neq (p_1 - p_2)_0$
Contraste con dos colas

Consideremos el Ejemplo 8.6.1.

Ejemplo 8.6.1. Los que recomiendan la vitamina C pretenden que con su ingestión mejora la posibilidad de que los pacientes sobrevivan al cáncer. Además, se cree que el porcentaje de pacientes que presentan una mejoría entre los que toman un suplemento de vitamina C supera al de aquellos que no lo hacen en más de 4 puntos porcentuales. La situación se recoge en la Figura 8.11. Desde el punto de vista de los que recomiendan la vitamina C, estamos contrastando

$$H_0: p_1 - p_2 \leq 0.04 \qquad H_1: p_1 - p_2 > 0.04$$

Para contrastar tales hipótesis, debe encontrarse un estadístico que sea lógico y cuya distribución de probabilidad se conozca, al menos aproximadamente, bajo el supuesto de que la hipótesis nula es cierta. Para obtener un estadístico tal, consideremos la variable aleatoria

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

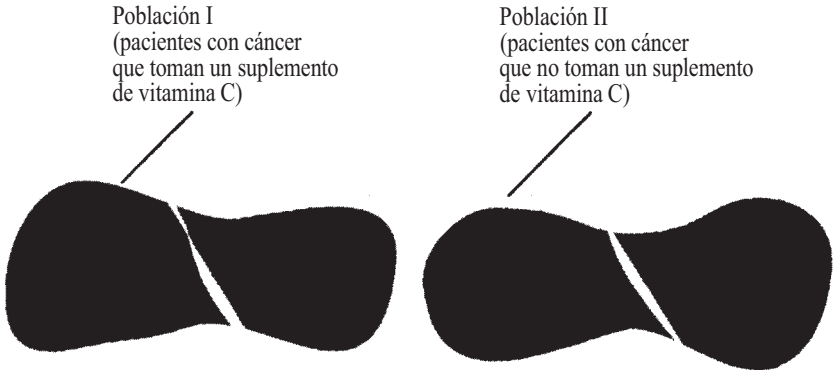


Figura 8.11. ¿Es $p_1 - p_2 > 0.04$?

Ésta es la misma que la utilizada en la Sección 8.5 para construir intervalos de confianza para $p_1 - p_2$. Si la hipótesis nula es cierta, la variable aleatoria es aproximadamente normal tipificada. En todo caso, se nos plantea un problema. ¡Un estadístico para el contraste debe ser un estadístico! La variable aleatoria anterior *no* es un estadístico porque contiene las proporciones poblacionales p_1 y p_2 desconocidas. La forma lógica de resolver este problema es reemplazar p_1 y p_2 por sus estimadores p_1 y p_2, para obtener el *estadístico* aproximadamente normal tipificado:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}$$

Se trata de una elección lógica para un estadístico, puesto que compara la diferencia estimada entre las proporciones $p_1 - p_2$ con la diferencia $(p_1 - p_2)_0$ hipotética. Si el valor hipotético es correcto, la diferencia estimada y la diferencia hipotética tendrán un valor próximo. Ello fuerza al numerador a estar cercano a cero y así proporciona un valor pequeño para el estadístico. Valores grandes positivos o negativos del estadístico indican que la hipótesis nula no es cierta, y que podría ser rechazada en favor de una alternativa apropiada.

La utilización de este estadístico se explica en el Ejemplo 8.6.2.

Ejemplo 8.6.2. Se utiliza un grupo de 150 pacientes para comprobar la teoría de que la vitamina C es una ayuda en el tratamiento del cáncer. Las hipótesis que se contrastan son

$$H_0: p_1 - p_2 \leq 0.04$$

$$H_1: p_1 - p_2 > 0.04$$

Puesto que la desigualdad en la hipótesis alternativa está a la derecha, el contraste es un contraste con cola a la derecha basado en la distribución normal tipificada.

Los 150 pacientes fueron divididos en dos grupos de 75. Un grupo recibió 10 gramos de vitamina C diariamente; el otro recibió un placebo cada día. De los que recibieron la vitamina C, 47 presentaron alguna mejoría en el plazo de cuatro semanas de los que recibieron el

placebo, solamente 43 experimentaron mejoría. Basándose en estos datos, $\hat{p}_1 = \frac{47}{75} \cong 0.63$ y $\hat{p}_2 = \frac{43}{75} \cong 0.57$. El valor observado del estadístico del contraste es

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} = \frac{(0.63 - 0.57) - 0.04}{\sqrt{(0.63)(0.37)/75 + (0.57)(0.43)/75}}$$
$$= 0.25$$

Dado que es con cola a la derecha,

$$P = P[Z \geq 0.25] = 1 - P[Z \leq 0.25]$$
$$= 1 - 0.5987$$
$$= 0.4013$$

Puesto que esta probabilidad es grande, no estamos en condiciones de rechazar H_0. Hablando en términos prácticos, esto significa que no hay suficiente evidencia en este estudio papa mantener el argumento de que la vitamina C ayuda en el tratamiento del cáncer con el alcance pretendido.

Contraste en el que el valor nulo es cero: contraste conjunto

Aunque la diferencia $(p_1 - p_2)_0$ que constituye la hipótesis nula puede tomar cualquiera de los valores propuestos, el que aparece más frecuentemente es el cero. En este caso, las hipótesis consideradas previamente comparan p_1 con p_2 y toman algunas de las siguientes formas:

I	II	III
$H_0: p_1 \leq p_2$	$H_0: p_1 \geq p_2$	$H_0: P_1 = P_2$
$H_1: p_1 > p_2$	$H_1: p_1 < p_2$	$H_1: P_1 \neq P_2$
Contraste con cola a la derecha	Contraste con cola a la izquierda	Contraste con dos colas

Estas hipótesis pueden ser contrastadas utilizando el estadístico del contraste definido anteriormente, sin más que tener en cuenta que $p_1 - p_2 = 0$. No obstante, existe un procedimiento alternativo basado en que si H_0 es cierta y $p_1 = p_2$, entonces $\hat{p}_1$ y $\hat{p}_2$ son estimadores de la misma proporción, que indicaremos con p. Queremos combinar o «reunir» a $\hat{p}_1$ y $\hat{p}_2$ para formar un estimador de p. Se puede tomar simplemente la media de los valores de $\hat{p}_1$ y $\hat{p}_2$, pero entonces no se están teniendo en cuenta las diferencias que pueda haber entre los tamaños de las muestras. Para tenerla en cuenta, se forma un estimador que dé tanta importancia a la proporción muestral como grande sea el tamaño de la muestra de la que se ha obtenido. Es decir, se construye una proporción «ponderada» tomando como pesos los tamaños muestrales. Así, se estima la proporción común de la población p mediante el estimador conjunto $\hat{p}$ donde $\hat{p}$ es dado por

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Obsérvese que p es precisamente el número total de elementos, en las dos muestras combinadas, que tienen el rasgo de interés, dividido por la suma de los tamaños muestrales.

Si ahora se sustituyen p_1 y p_2 por $\hat{p}$ en la fórmula anterior, se obtiene el siguiente estadístico del contraste para contrastar H_0: $p_1 - p_2 = 0$;

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2}}$$

Simplificando

$$\boxed{\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}$$

Ejemplo 8.6.3. Un enemigo importante del caracol *(Cepaea memoralis)* es el zorzal cantor. Estos pájaros los seleccionan en las colonias de caracoles y los llevan a las rocas próximas. Allí, los pájaros parten los caracoles, comen las partes blandas y dejan las conchas. En un estudio de selección natural, se comparó la proporción de conchas no listadas encontradas junto a las rocas con la proporción de caracoles no listados en la colonia próxima, una ciénaga en los alrededores de Oxford, Inglaterra. El fondo de la ciénaga era bastante uniforme. Se pensó que, debido a su capacidad para armonizar con el fondo, los caracoles no listados estarían más protegidos de los depredadores que los individuos listados de la colonia. Esto daría como resultado que la proporción de conchas no listadas en las rocas fuera más pequeña que la de caracoles no listados en la colonia. La situación se esquematiza en la Figura 8.12.

Tomando como hipótesis alternativa la teoría que se pretende sostener, el propósito del estudio es contrastar

H_0: $p_1 \geq p_2$
H_1: $p_1 < p_2$ (los caracoles no listados están protegidos en la ciénaga)

De 863 conchas rotas alrededor de las rocas, 377 eran no listadas. Ello proporciona una estimación puntual para p_1 de $\hat{p}_1 \cong \frac{377}{863} = 0.44$. De 560 individuos recogidos en la ciénaga, 296 eran no listados, lo que da una estimación puntual para p_2 de $\hat{p}_2 = \frac{296}{560} \cong 0.53$. La estimación conjunta de p es

$$\hat{p} = \frac{863(0.44) + 560(0.53)}{863 + 560} \cong 0.47$$

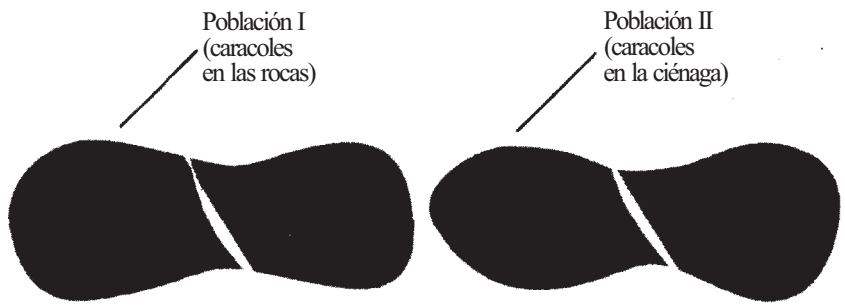


Figura 8.12. ¿Es $p_1 < p_2$?

Obsérvese que $\hat{p}$ viene dado también por

$$\hat{p} = \frac{377 + 296}{863 + 560} = \frac{673}{1423} \cong 0.47$$

como era de esperar. De este modo, el valor que toma en la muestra el estadístico del contraste es

$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.44 - 0.53)}{\sqrt{0.47(1-0.47)\left(\frac{1}{863} + \frac{1}{560}\right)}}$$

$$= \frac{-0.09}{0.0271}$$

$$= -3.321$$

No es sorprendente que este valor sea negativo. Esta prueba es un contraste con cola a (a izquierda que pide que se rechace H_0, para valores negativos grandes del estadístico. ¿Es «grande» -3.321? Para contestar a esta pregunta, obsérvese en la Tabla III del Apéndice B que $P[Z \leq -3.321] = 0.0005$. De este modo, el valor P para el contraste es 0.0005. Hay dos posibles explicaciones para un valor P tan pequeño:

1. Los caracoles no listados no están realmente protegidos por su coloración. Simplemente *por azar* observamos un suceso que se presenta solamente alrededor de 5 veces en cada 10 000 pruebas.
2. Los caracoles no listados están protegidos en la ciénaga por su habilidad para armonizar con el fondo uniforme.

¡Es preferible la última explicación!

Es necesario hacer una nueva puntualización. Aunque nos hemos referido a p_1 y p_2 como proporciones, y en ocasiones las hemos pasado a porcentajes, también pueden ser interpretadas como probabilidades. Esto es admisible porque siempre son números comprendidos entre 0 y 1 que, en efecto, representan la probabilidad de seleccionar en una extracción aleatoria un objeto de la población que tenga el rasgo que está siendo estudiado.

EJERCICIOS 8.6

1. Considérense los resultados del Ejemplo 8.6.2. ¿A qué tipo de error está expuesto el experimentador? Comentar las consecuencias prácticas de cometer tal error. ¿Es correcto afirmar que se ha comprobado que la vitamina C no ayuda en el tratamiento del cáncer al nivel que se pretendía? Explicarlo.
2. Se efectúa un estudio sobre el color de los escarabajos tigre para conseguir pruebas que apoyen el argumento de que la proporción de escarabajos negros puede variar de un lugar a otro. En una muestra de 500 escarabajos capturados en una extensión próxima a Providence, Rhode Island, 95 eran negros. Una captura de 112 escarabajos en Aqueduct, Nueva York, contenía 17 individuos negros.
   a) Construir el contraste de hipótesis adecuado con dos colas.
   b) Hallar una estimación puntual para la diferencia entre las proporciones de escarabajos negros en las dos regiones. ¿Piensa, basándose en esta estimación, que hay diferencia entre las dos proporciones?
   c) Realizar el contraste del apartado *a.* ¿Cuál es su valor *P?* Recuerde que el contraste es de dos colas.

3. Se realiza un estudio para detectar la eficacia de las mamografías. De 31 casos de cáncer de mama detectados en mujeres del grupo de edad de entre los cuarenta y cuarenta y nueve años, 6 lo fueron a través, exclusivamente, de la mamografía. En mujeres de más edad, 38 de 101 cánceres detectados lo fueron sólo por mamografía. ¿Es esto evidencia, al nivel $\alpha = 0.5$, de que la probabilidad de detectar el cáncer mediante la mamografía es más alta en las mujeres mayores que en las más jóvenes? Explicar la respuesta construyendo las hipótesis estadísticas apropiadas y realizando el contraste.
4. En un estudio realizado en 1970, se hicieron pruebas de sangre en 759 pacientes afectados por diversas infecciones de la sangre. En 46 de estos casos, se aislaron al menos dos microorganismos diferentes en la misma muestra de sangre. En 1975, en otro estudio semejante con 838 pacientes, se obtuvo que 109 de ellos presentaban dos o más microorganismos en la sangre. Basándose en estas muestras, ¿puede el lector pretender con seguridad que la proporción de tales casos se ha incrementado en más de 6 puntos porcentuales, en un período de cinco años? Explicar la respuesta construyendo las hipótesis apropiadas y hallando su valor *P*.
5. En un reciente estudio de lesiones de rodilla entre los jugadores de fútbol que juegan sobre césped natural, se compararon dos tipos de zapatos. En 266 jugadores que calzaban zapatos de fútbol con abrazaderas múltiples, se presentaron 14 lesiones de rodilla. De 2055 jugadores que calzaban zapatos de fútbol convencionales de siete puntos, se encontraron 162 de tales lesiones. ¿Es ésta una prueba, para un nivel $\alpha = 0.1$, de que la probabilidad de sufrir una lesión de rodilla cuando se calzan zapatos de fútbol convencionales es más alta que cuando se utiliza el otro tipo de zapatos? ¿Puede hacerse la misma consideración al nivel $\alpha = 0.5$?
6. Cuando un laboratorio farmacéutico anuncia un nuevo medicamento, se incluye un perfil del mismo. En él aparece una comparación de los efectos secundarios del medicamento con los efectos secundarios del de su competidor más cercano. El medicamento en cuestión se utiliza para tratar la acidez. Se obtuvieron los siguientes datos sobre el porcentaje de personas que presentaban efectos secundarios:

Efecto secundario	Propio (n = 465), %	Competidor (n = 195), %
Dolor de cabeza	2.4	2.6
Diarrea	1.9	0.5

   a) Realizar un contraste para las diferencias entre los porcentajes de pacientes con dolor de cabeza. (*Sugerencia:* El contraste es de dos colas.)
   b) Realizar un contraste para las diferencias entre los porcentajes de pacientes que sufren diarrea.

   (Información obtenida de un anuncio que apareció en *Emergency Medicine,* noviembre de 1990, págs. 27-30.)

HERRAMIENTAS COMPUTACIONALES

TI83

XVIII. Intervalos de confianza de proporciones

La calculadora TI83 puede hallar intervalos de confianza de p o de $p_1 - p_2$, a partir del número de individuos que poseen la característica de interés y los tamaños muestrales. Se ilustra el

procedimiento construyendo el intervalo de confianza hallado en el Ejemplo 8.5.3. Recuérdese que la calculadora emplea más cifras decimales en sus cálculos que los que se han hecho en el texto. Por ello habrá ligeras discrepancias en los resultados finales.

Tecla/Comando de la TI83	**Propósito**
1. STAT ◁ ALPHA B	1. Accede a la pantalla necesaria para construir el intervalo de confianza dep_1-p_2.
2. 1 ENTER	2. Introduce 1 como valor de x_1.
3. 34 ENTER	3. Introduce 34 como valor de n_1.
4. 10 ENTER	4. Introduce 10 como valor de x_2.
5. 38 ENTER	5. Introduce 38 como valor de n_2.
6. 0.95 ENTER	6. Pide que el intervalo de confianza sea del 95 %.
7. ENTER	7. Calcula y muestra en pantalla el intervalo de confianza de p_1 - p_2, del 95%.

Los intervalos de confianza de p se hallan sustituyendo el paso 1 por:

STAT
◁
ALPHA
A

XIX. Contrastes de hipótesis de proporciones

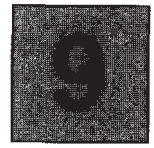
La calculadora TI83 puede realizar contrastes de hipótesis para el valor de p o para comparar p_1 *y* p_2, mediante el contraste de proporciones «conjuntas» descrito en la Sección 8.6. Se ilustra el procedimiento usando los datos del Ejemplo 8.6.3.

Tecla/Comando de la TI83	**Propósito**
1. STAT ◁ 6	1. Accede a la pantalla necesaria para realizar el contraste de proporciones «conjuntas».
2. 377 ENTER	2. Introduce 377 como valor de x_1.
3. 863 ENTER	3. Introduce 863 como valor de n_2.
4. 296 ENTER	4. Introduce 296 como valor de x_2.
5. 560 ENTER	5. Introduce 560 como valor de n_2.

6. ▷ ENTER	6. Indica que el contraste es con cola a la izquierda.
7. ▽ ENTER	7. Realiza el contraste; proporciona el valor del estadístico del contraste (-3.385657423), el valor *P* (0.000355) y la estimación conjunta de *p* (0.4729444835).

El contraste para *p* se puede hacer sustituyendo el paso 1 por:

STAT
◁
5



Comparación de dos medias y dos varianzas

En este capítulo, continuamos con el estudio de problemas relativos a dos muestras considerando métodos para comparar las medias de dos poblaciones. Se hace en condiciones experimentales diferentes, es decir, cuando las muestras extraídas son independientes y cuando los datos están emparejados. Se explicarán a fondo estos términos en sucesivas secciones.

9.1. ESTIMACIÓN PUNTUAL: MUESTRAS INDEPENDIENTES

La situación general se puede describir del modo siguiente: hay dos poblaciones de interés, cada una con media desconocida; se extrae una muestra aleatoria de la primera población y otra de la segunda, de forma tal que los objetos seleccionados de la población I no tengan relación con los objetos de la población II (tales muestras se dice que son independientes); se comparan las medias poblacionales por medio de una estimación puntual (Véase Fig. 9.1.)

Ejemplo 9.1.1. Hasta hace poco tiempo, los granjeros suecos fumigaban el 80 % de todos los cereales sembrados con un fungicida que contenía metilo de mercurio. Se llevó a cabo un estudio para comparar el nivel medio de mercurio en los huevos producidos en Suecia con el de los huevos producidos en Alemania, donde no se utiliza el metilo de mercurio. Se seleccionó una muestra aleatoria de huevos producidos en Suecia y otra muestra aleatoria de huevos producidos en Alemania. Estas muestras son independientes en el sentido de que los huevos seleccionados en un país de ningún modo afectan a aquellos seleccionados en el otro. El estudio puede ser visualizado como se indica en la Figura 9.2.

La forma lógica de estimar la diferencia entre las medias poblacionales, $\mu_1 - \mu_2$, consiste en estimar cada media individualmente, utilizando los métodos descritos en el Capítulo 6, y después estimar $\mu_1 - \mu_2$ como la diferencia entre estas estimaciones individuales. Es decir,

$$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2$$

donde $\bar{X}_1$ es la media muestral basada en la muestra de la población I y $\bar{X}_2$ es la media muestral basada en la muestra independiente extraída de la población II.

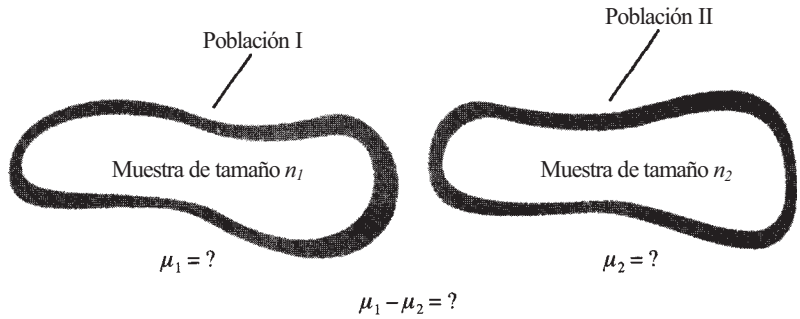


Figura 9.1. Muestras aleatorias independientes extraídas de dos poblaciones.

Ejemplo 9.1.2. Cuando se realizó el estudio del Ejemplo 9.1.1, se obtuvieron los siguientes datos sobre los niveles de mercurio en los huevos:

Suecia	Alemania
$n_1 = 2000$	$n_2 = 2500$
$\bar{x}_1 = 0.026$ ppm	$\bar{x}_2 = 0.007$ ppm
$s_1 = 0.01$	$s_2 = 0.004$

Basándose en esta información, una estimación para la diferencia en los niveles medios de mercurio en los huevos de los dos países es:

$$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$$
$$= 0.026 - 0.007$$
$$= 0.019 \text{ ppm}$$

El Teorema 9.1.1, relativo a la distribución del estimador $\bar{X}_1 - \bar{X}_2$, proporciona la *base* teórica para una estimación por intervalos de confianza y contraste de hipótesis para $\mu_1 - \mu_2$. El teorema puede verificarse parcialmente utilizando las propiedades de la esperanza y la varianza expuestas en el Apéndice A.

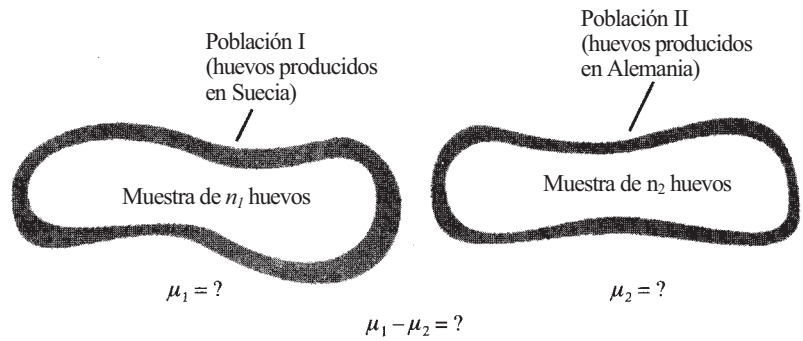


Figura 9.2. Dos muestras aleatorias independientes de huevos de Alemania y de Suecia.

Teorema 9.1.1. Distribución de $\bar{X}_1 - \bar{X}_2$. Sean $\bar{X}_1$ y $\bar{X}_2$ las medias muestrales basadas en muestras independientes de tamaños n_1 y n_2, extraídas de distribuciones normales con media μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2, respectivamente. Entonces, la variable aleatoria $\bar{X}_1 - \bar{X}_2$ es normal con media $\mu_1 - \mu_2$ y varianza $\sigma_1^2/n_1 + \sigma_2^2/n_2$.

Como en el caso de una muestra, por el Teorema central del límite, se puede suponer que, para muestras de tamaño grande, $\bar{X}_1 - \bar{X}_2$ es al menos aproximadamente normal, incluso si las muestras son extraídas de poblaciones que no están normalmente distribuidas.

EJERCICIOS 9.1

1. Está generalmente aceptado que existen diferencias ligadas al sexo relacionadas con la respuesta al estrés producido por el calor. Se sometió a un grupo de 10 varones y 8 mujeres a un programa de ejercicios enérgicos que implicaba el empleo de una «cinta sin fin». El medio era caluroso, y se disponía de una cantidad mínima de agua para los individuos. La variable de interés fue el porcentaje de peso corporal perdido. Se obtuvieron los datos siguientes:

Varones		Mujeres	
2.9	3.7	3.0	3.8
3.5	3.8	2.5	4.1
3.9	4.0	3.7	3.6
3.8	3.6	3.3	4.0
3.6	3.7		

Establecer una estimación puntual para la diferencia en porcentajes de pérdida de peso corporal entre varones y mujeres que hacen ejercicio en estas condiciones.

2. En un estudio llevado a cabo para comparar algunas de las características físicas de las nadadoras olímpicas con los de las corredoras olímpicas, la variable de interés era la grasa corporal total, en kilogramos. Se obtuvieron muestras de 12 corredoras y 10 nadadoras con los siguientes resultados:

Corredoras		Nadadoras	
11.2	7.6	14.1	12.7
10.1	7.3	15.1	13.7
9.4	6.9	11.4	11.9
9.2	5.5	14.3	10.7
8.3	5.0	9.2	8.7
8.2	3.7		

Establecer una estimación puntual para la diferencia en la grasa corporal total media entre las corredoras y las nadadoras olímpicas.

3. Ha de elegirse una muestra de tamaño 20 de una distribución normal con una media 15 y varianza 16. Se selecciona una muestra independiente de tamaño 25 de una distribución

normal con media 10 y varianza 18. ¿Cuál es la media de la variable aleatoria $\bar{X}_1 - \bar{X}_2$? ¿Cuál es la varianza? ¿Cuál es su desviación típica? ¿Qué tipo de variable (Z, T, X^2, ...) es la variable siguiente?

$$\frac{(\bar{X}_1 - \bar{X}_2) - 5}{\sqrt{\frac{16}{20} + \frac{18}{25}}}$$

4. Se llevó a cabo un estudio para investigar el efecto producido por el desagüe de una zona de aparcamiento en la densidad de la vegetación circundante. Se estudiaron dos áreas. Una era objeto del desagüe de una gran zona de aparcamiento; la otra no estaba cerca de ningún aparcamiento y se utilizó de control. Cada área se subdividió en una serie de paneles de 2 metros por 20 metros y se contó el número de plantas encontradas en cada uno, obteniéndose estos datos:

Área de drenaje del aparcamiento		Área de control	
62	64	72	59
76	74	77	64
58	71	60	62
57	59	59	75
79	54	61	69
82	49	64	64
72	53	69	71
77		65	

a) Estimar el número medio de plantas por panel encontradas en cada área.
b) Estimar la diferencia en el número medio de plantas encontradas por panel. Reste en este orden: área de control menos área de drenaje.
c) Se cree que los contaminantes procedentes del aparcamiento harán disminuir el número de plantas encontradas en el área de drenaje. ¿Confirmaría esta idea la estimación puntual encontrada en el apartado *b?* ¿Puede estar muy seguro de que esta idea es correcta a partir de esta estimación puntual? ¿Si quisiese reforzar esta idea de modo que pudiese aportar un porcentaje de error, qué haría?

(Basado en un estudio realizado por Thomas Edward Wilkerson IV, Departamento de Biología, Radford University, 1993.)

5. Se realizó un estudio para comparar el nivel medio de lectura de los pacientes atendidos en clínicas públicas frente a los atendidos en clínicas universitarias. Se obtuvieron los siguientes datos:

Clínica pública	Clínica universitaria
n = 30 $\bar{x}$ = 5.4 (5.° grado, 4.° mes)	« = 90 $\bar{x}$ = 6.8 (6.° grado, 8.° mes)

Estimar la diferencia del nivel medio de lectura entre los pacientes de estos dos tipos de clínicas. (Basado en la información hallada en Terry Davis et al., «The Gap Between Patient Reading Comprehension and the Readability of Patient Education Materials», *The Journal of Family Practice,* noviembre de 1990, págs. 533-537.)

6. Considérese la situación descrita en el Ejercicio 5. Si suponemos que los niveles de lectura en cada clínica varían de 1 a 12, podemos utilizar la idea explicada en la Sección 6.6 para aproximar cada una de las desviaciones típicas muestrales. En este caso, $s_1 = s_2 \cong \text{rango}/4 = 2.75$.
   a) Utilizar la información del Ejercicio 5 para hallar un intervalo de confianza del 95 % del nivel medio de lectura entre los pacientes atendidos en la clínica pública y en la clínica universitaria.
   b) Se analizó la legibilidad del material escrito entregado a los pacientes en estas clínicas, anotándose los niveles medios de lectura de los diferentes materiales:

Formulario de consentimiento del paciente	16.1
Dieta para un corazón sano	14.8
Hipertensión	8.6
Los 12 pasos de Alcohólicos Anónimos	11.3
Problemas del embarazo	12.0

   ¿Indican estos datos que los materiales de lectura entregados a los pacientes están escritos a un nivel demasiado alto para el público al que iba dirigido? Explicarlo basándose en los intervalos de confianza hallados en el punto *a.*

9.2. COMPARACIÓN DE VARIANZAS: LA DISTRIBUCIÓN *F*

Hay dos razones por las que querer comparar dos varianzas poblacionales. Primero, porque muchos estudios tienen como principal propósito la comparación de dos medias. No obstante, hay dos estadísticos para hacer esta comparación. Uno se usa cuando las varianzas poblacionales parecen ser iguales; el otro es apropiado cuando estas varianzas parecen ser diferentes. Necesitamos un instrumento para comparar varianzas, de manera que pueda elegirse el método adecuado para comparar las medias. Desarrollaremos una regla especial o método práctico para este propósito. La segunda razón para querer comparar varianzas es que dicha comparación nos interesa especialmente. Queremos ser capaces de sacar conclusiones acerca de la relación entre dos varianzas porque esta relación afecta a nuestro estudio. Desarrollaremos un «contraste de la *F*» para hacer dichas comparaciones.

Regla práctica para la comparación de varianzas

Considérese una situación en la que se dispone de dos muestras independientes. El propósito principal del estudio es comparar las medias de dos poblaciones de las que se han estraído las dos muestras.

Hay dos estadísticos usados para comparar las medias de dos poblaciones normales. Ello se debe al hecho de que hay dos posibilidades distintas.

Éstas son:

1. σ_1^2 y σ_2^2 son desconocidas, pero se supone que son iguales.
2. σ_1^2 y σ_2^2 son desconocidas y no se supone que sean iguales.

La primera tarea del investigador es determinar cuál de las dos situaciones se da en su estudio. Esto significa que necesitamos desarrollar un procedimiento por el que podamos determinar rápidamente si la evidencia tiende a señalar el hecho de que σ_1^2 y σ_2^2 son diferentes.

Ejemplo 9.2.1. Se realiza un estudio de prácticas de prescripción. El propósito es analizar la prescripción de digoxina, un fármaco importante y comúnmente utilizado que es potencial-

mente tóxico. Se sabe que, generalmente, el nivel de dosificación para los que están por encima de los sesenta y cuatro años de edad debería ser menor que el de personas más jóvenes. Para llevar a cabo este estudio, se extraen muestras independientes de cada grupo y se obtiene el nivel de dosificación de digoxina para cada paciente seleccionado.

Se proponen dos preguntas. Cada una debe ser respondida estadísticamente, basándose en la información obtenida de las muestras. La pregunta principal es ésta: ¿es $\mu_1 < \mu_2$? En todo caso, antes de poder responderla, debemos considerar otra pregunta, ¿es $\sigma_1^2 = \sigma_2^2$?

Es fácil hallar un estadístico lógico para comparar varianzas. Recuérdese que las varianzas muestrales S_1^2 y S_2^2 son estimadores para las varianzas poblacionales σ_1^2 y σ_2^2, respectivamente. De este modo, para comparar σ_1^2 con σ_2^2, compararemos simplemente S_1^2 con S_2^2. Esto se hace, no considerando la diferencia entre las dos, sino considerando su cociente S_1^2/S_2^2. Si las dos poblaciones desconocidas resultan ser de hecho iguales, estiman ambas la misma cosa. En este caso, esperaríamos que S_1^2 y S_2^2 tuviesen valores semejantes, forzando al cociente S_1^2/S_2^2 a ser cercano a 1. Esto es, valores cercanos a 1 confirman la idea de que $\sigma_1^2 = \sigma_2^2$.

Para usar la regla práctica para comparar varianzas poblacionales, designamos a la *mayor* de las varianzas muestrales como s_1^2 y a la menor como s_2^2. Si el cociente s_1^2/s_2^2 está próximo a 1, no hay apenas evidencia de que σ_1^2 y σ_2^2 sean diferentes. Sin embargo, si s_1^2/s_2^2 es mucho mayor que 1, tenemos evidencias suficientes de que las varianzas poblacionales son diferentes. ¿Qué tamaño debe tener el cociente para estar convencidos de que $\sigma_1^2 \neq \sigma_2^2$? Esta cuestión queda resuelta por la regla práctica.

Regla práctica para comparar σ_1^2 y σ_2^2. Sean s_1^2 y s_2^2 las varianzas de dos muestras extraídas de distribuciones normales. Supongamos que $s_1^2 \geq s_2^2$. Entonces, si $s_1^2/s_2^2 \geq 2$, suponemos que $\sigma_1^2 \neq \sigma_2^2$.

Esta regla simplemente dice que si la varianza muestral más grande es al menos dos veces mayor que la pequeña, supondremos que σ_1^2 y σ_2^2 son diferentes. Tendremos esto en cuenta cuando elijamos un estadístico para comparar las medias.

Ejemplo 9.2.2. Cuando se realizó el estudio descrito en el Ejemplo 9.2.1, se obtuvieron los siguientes datos:

Pacientes con más de 64 años	Pacientes con 64 años o menos
$n_1 = 41$	$n_2 = 29$
$\bar{x}_1 = 0.265$ mg/día	$\bar{x}_2 = 0.268$ mg/día
$s_1 = 0.102$ mg/día	$s_2 = 0.068$ mg/día
$s_1^2 = 0.010404$	$s_2^2 = 0.004624$

Antes de comparar μ_1 y μ_2, comparamos las varianzas. La varianza muestral mayor se designa como s_1^2. El cociente s_1^2/s_2^2 viene dado por:

$$s_1^2/s_2^2 = 0.010404/0.004624 = 2.25$$

Como el valor de esta relación excede de 2, concluimos que las varianzas poblacionales σ_1^2 y σ_2^2 son diferentes.

La regla es bastante tolerante. No queremos usar un estadístico para comparar medias que suponga que $\sigma_1^2 = \sigma_2^2$, si existe el más ligero indicio de que esa suposición no es cierta. En el Ejercicio 9 se le pedirá que investigue matemáticamente esta regla práctica.

Contraste de la *F* para comparar varianzas: distribución *F* (opcional)

Puede efectuarse un contraste de hipótesis formal sobre el cociente de dos varianzas, bien como método para elegir el estadístico apropiado para comparar medias, bien porque nuestro interés se centre en las varianzas mismas. Los contrastes pueden tomar una cualquiera de las tres formas usuales, dependiendo del propósito del estudio. Estas formas son:

I	II	III
H_0: $\sigma_1^2 \leq \sigma_2^2$	H_0: $\sigma_1^2 \geq \sigma_2^2$	H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 > \sigma_2^2$	H_1: $\sigma_1^2 < \sigma_2^2$	H_1: $\sigma_1^2 \neq \sigma_2^2$
Contraste con cola a la derecha	Contraste con cola a la izquierda	Contraste con dos colas

El estadístico usado para comprobar cualquiera de estas hipótesis es S_1^2/S_2^2, el mismo empleado en la regla práctica. En este caso, S_1^2 no tiene por qué ser la mayor de las dos varianzas. Si la hipótesis nula es cierta y las varianzas poblaciones son realmente iguales, entonces esperaríamos que S_1^2 y S_2^2 tuvieran un valor próximo, forzando a S_1^2/S_2^2 a estar próximo a 1. Si el cociente está cerca de cero, entonces, naturalmente, concluimos que las varianzas poblacionales no son iguales y que, de hecho, $\sigma_1^2 < \sigma_2^2$. Recíprocamente, si S_1^2/S_2^2 es mucho mayor que 1, también concluimos que las varianzas poblacionales son diferentes y, en este caso, que $\sigma_1^2 > \sigma_2^2$.

Cuando utilizamos las frases *está cerca de cero* y *mucho mayor que uno,* estamos hablando en términos de probabilidades. Es decir, un valor observado del estadístico está «cerca de cero» cuando es demasiado *pequeño* para haber ocurrido razonablemente por azar si, de hecho, las varianzas poblacionales son iguales. Análogamente, un valor observado es «mucho mayor que uno» si es demasiado *grande* para haber ocurrido razonablemente al azar. Para determinar la probabilidad de observar distintos valores del estadístico S_1^2/S_2^2, debemos conocer su distribución de probabilidad. Veremos que este estadístico sigue una nueva distribución. Concretamente, si las varianzas poblacionales son iguales, se obtiene lo que se llama una distribución *F.* Esta distribución se define en términos de una distribución previamente estudiada, la distribución ji-cuadrado. En particular, una variable aleatoria *F* puede escribirse como el cociente de dos variables aleatorias ji-cuadrado independientes, afectadas por sus respectivos grados de libertad.

Definición 9.2.1. Distribución *F.* Sean $\mathbf{X}^2_{\gamma_1}$ y $\mathbf{X}^2_{\gamma_2}$ dos variables aleatorias ji-cuadrado independientes, con γ_1 y γ_2 grados de libertad, respectivamente. Entonces, la variable aleatoria:

$$\frac{\mathbf{X}^2_{\gamma_1}/\gamma_1}{\mathbf{X}^2_{\gamma_2}/\gamma_2}$$

sigue lo que se llama una *distribución F* con γ_1 y γ_2 grados de libertad.

Las propiedades más importantes de la familia de variables aleatorias *F* son las siguientes:

1. Hay una infinidad de variables aleatorias *F,* cada una identificada por dos parámetros γ_1 y γ_2, llamados *grados de libertad.* Estos parámetros son siempre enteros positivos: γ_1 está asociado con la variable aleatoria ji-cuadrado del numerador de la variable *F*

y γ_2 está asociado con la variable ji-cuadrado del denominador. La notación F_{γ_1, γ_2} representa una variable aleatoria F con γ_1 y γ_2 grados de libertad.

2. Cada variable aleatoria F es continua.
3. La gráfica de la función de densidad de cada variable F es una curva asimétrica con la forma general mostrada en la Figura 9.3.
4. Las variables F no pueden tomar valores negativos.

En la Tabla IX del Apéndice B se da una tabla parcial de valores de la función de distribución acumulativa para las variables F con distintos grados de libertad. En ella γ_1, grados de libertad del numerador, aparece como encabezamiento de columna; γ_2, grados de libertad del denominador, aparece como encabezamiento de fila.

Los puntos con las áreas 0.01, 0.025, 0.05 y 0.1 a la derecha pueden leerse directamente en la tabla. Los puntos con estas áreas a la izquierda pueden calcularse a partir de la tabla. La técnica se muestra en el Ejemplo 9.2.3.

Ejemplo 9.2.3. Considérese $F_{10,15}$, variable aleatoria F con 10 y 15 grados de libertad.

a) Hallar $P[F_{10,15} \leq 2.54]$. Buscar en las tablas F, en la columna 10 y fila 15, hasta que se encuentre el valor 2.54. Éste se encuentra en la tabla llamada $P[F_{\gamma_1, \gamma_2} \leq f] = 0.95$. Por lo tanto, $P[F_{10,15} \leq 2.54] = 0.95$.

b) Hallar $P[F_{10,15} \geq 3.06]$. El valor 3.06 se encuentra en la columna 10, fila 15, de la tabla llamada $P[F_{\gamma_1, \gamma_2} \leq f] = 0.975$. Así $P[F_{10,15} \leq 3.06] = 0.975$. o , a su vez, implica que $P[F_{10,15} \geq 3.06] = 1 - 0.975 = 0.025$.

c) El punto con el área 0.025 a su derecha (0.975 a la izquierda) es 3.06.

d) El punto con el área 0.05 a su derecha (0.95 a la izquierda) es 2.54.

e) El punto con el área 0.01 a su derecha (0.99 a la izquierda) es 3.80.

f) ¿Qué punto tiene el área 0.975 a su derecha (0.025 a la izquierda)? Es el punto con cola a la izquierda que aparece en la Figura 9.4. Ya que no puede leerse directamente en la Tabla IX, se calcula tomando el recíproco del punto correspondiente con cola a la derecha para una variable aleatoria F con los grados de libertad en orden inverso. Considérese la variable aleatoria $F_{15,10}$. Obsérvese que hemos invertido los grados de libertad. El punto que para esta curva deja un área de 0.025 a la derecha es 3.52. De manera que, el punto buscado, con la cola a la izquierda, tiene un valor de $1/3.52 = 0.28$.

g) El punto, con cola a la izquierda, con un área de 0.95 a la derecha y 0.05 a la izquierda es:

$$\frac{1}{2.85} = 0.35$$

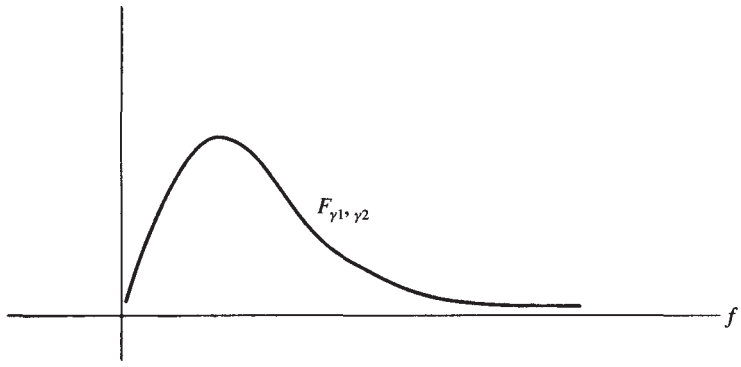


Figura 9.3. Una típica función de densidad F.

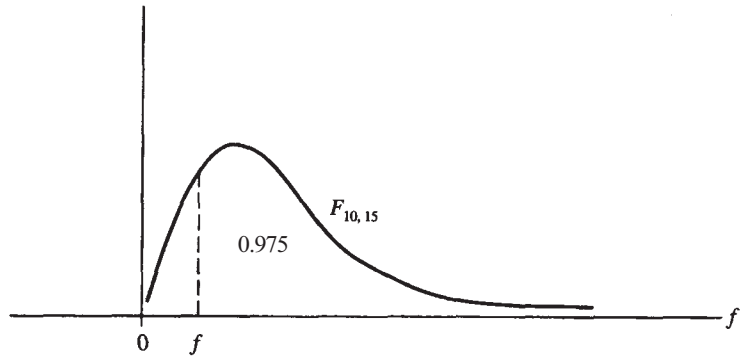


Figura 9.4

En ocasiones, las tablas no permiten la lectura de algún punto F con cola a la derecha. Por ejemplo, si el numerador tiene 50 grados de libertad, el punto F no está en la lista. Podemos, sin embargo, leer puntos para 40 y 60 grados de libertad. Cada uno de ellos da una buena aproximación del punto deseado. Haremos una aproximación conservadora y elegiremos el punto que ligeramente sobreestime el punto de interés. En la mayoría de los casos, esto significará que elijamos el menor de los posibles grados de libertad para realizar la estimación.

El Teorema 9.2.1 proporcionará el estadístico necesario para realizar el contraste de hipótesis sobre el cociente de dos varianzas poblacionales.

Teorema 9.2.1. Sean S_1^2 y S_2^2 las varianzas muestrales basadas en muestras aleatorias independientes de tamaños n_1 y n_2 extraídas de dos poblaciones normales, con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2, respectivamente. Si $\sigma_1^2 = \sigma_2^2$, el estadístiS_1^2/S_2^2 sigue una distribución F con $n_1 - 1$ y $n_2 - 1$ grados de libertad, respectivamente.

Obsérvese que los grados de libertad asociados con el estadístico S_1^2/S_2^2 son $n_1 - 1$ y $n_2 - 1$. Es decir, el número de grados de libertad del numerador es 1 menos que el tamaño de la muestra extraída de la población I; el del denominador es 1 menos que el tamaño de la muestra extraída de la población II. En el Ejemplo 9.2.4 se expüca la utilización del estadístico S_1^2/S_2^2 en un contraste de hipótesis.

Ejemplo 9.2.4. En el estudio de la digoxina del Ejemplo 9.2.2, queremos contrastar:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Los datos obtenidos son los siguientes:

Pacientes con más de 64 años	Pacientes de 64 años o menos
$n_1 = 41$	$n_2 = 29$
$\bar{x}_1 = 0.265$ mg/día	$\bar{x}_2 = 0.268$ mg/día
$s_1 = 0.102$ mg/día	$s_2 = 0.068$ mg/día
$s_1^2 = 0.010404$	$s_2^2 = 0.004624$

El contraste es de dos colas. La hipótesis H_0 se rechazará en favor de H_1 si el valor observado del estadístico es demasiado grande o demasiado pequeño para que haya sucedido por azar cuando las varianzas poblacionales son iguales. El número de grados de libertad asociados con el estadístico es $n_1 - 1 = 41 - 1 = 40$ y $n_2 - 1 = 29 - 1 = 28$. El valor observado del estadístico es

$$s_1^2/s_1^2 = \frac{0.010404}{0.004624} = 2.25$$

En la tabla *F* con 40 y 28 grados de libertad vemos que $P[F_{40,28} \geq 2.05] = 0.025$ y que $P[F_{40,28} \geq 2.35] = 0.01$. Puesto que el valor observado del estadístico cae entre 2.05 y 2.35, el valor *P* para el contraste con cola a la derecha está entre 0.01 y 0.025. Sin embargo, como él contraste que se está efectuando es de dos colas, es el doble del contraste de una cola. Es decir, para el contraste, tal como se ha establecido,

$$0.02 < P < 0.05$$

Como esta probabilidad es pequeña, rechazamos H_0 y concluimos que las dos varianzas poblacionales son diferentes.

Debe resaltarse un punto importante. Estamos suponiendo, una vez más, que las poblaciones en estudio son normales. Este supuesto es necesario para que el estadístico S_1^2/S_2^2 tenga una distribución *F*. La consecuencia de no cumplir este supuesto es que el valor *P* o el nivel α obtenido, según el caso, pueden no ser exactos. En todo caso, se ha encontrado que este problema se minimiza si las muestras son de *igual tamaño.*

La ventaja de la regla práctica sobre el contraste de la *F* es obvia. Es un medio fácil y rápido para contestar a la pregunta ¿es $\sigma_1^2 = \sigma_2^2$? Las ventajas del contraste de la *F* sobre la regla son que se pueden realizar hipótesis direccionales y añadirse un valor *P* al contraste.

Si el único propósito de comparar varianzas es determinar un estadístico apropiado para comparar medias, sugerimos que se use la regla práctica. Si el interés se centra en la relación entre dos varianzas poblacionales, deberá usarse el contraste de la *F*.

Tanto el SAS como la TI83 están programados para realizar el contraste de la *F* para comparar varianzas.

EJERCICIOS 9.2

1. Se ha realizado un estudio sobre la velocidad en vuelo de diversas especies de pájaros. El propósito era comparar las velocidades del pelícano pardo y el ostrero americano. Se cronometró a los pájaros volando con el viento de costado con una velocidad de viento de 5 a 8 millas/h, y se obtuvo la siguiente información (supóngase normalidad):

Pelícano pardo	**Ostrero**
$n_1 = 9$	$n_2 = 12$
$\bar{x}_1 = 26.05$ millas/h	$\bar{x}_2 = 30.19$ millas/h
$s_1 = 6.34$ millas/h	$s_2 = 3.20$ millas/h

Úsese la regla práctica para ver si hay evidencias de que $\sigma_1^2 \neq \sigma_2^2$.

2. Se ha llevado a cabo un estudio para comparar el diámetro medio de los anillos de los abetos de Fraser que se encuentran a una altitud de 5000 pies en dos años diferentes, 1983 y 1988. Se obtuvieron estos datos:

1983	**1988**
$n_1 = 10$	$n_2 = 10$
$\bar{x}_1 = 0.535$ mm	$\bar{x}_2 = 0.439$ mm
$s_1 = 0.049$	$s_2 = 0.055$

Úsese la regla práctica para ver si hay evidencia de que $\sigma_1^2 \neq \sigma_2^2$.
(Basado en un estudio realizado por Christopher Cook, Departamento de Biología, Universidad de Radford, 1993.)

3. Se realiza un estudio para considerar el efecto que el tabaquismo de las madres tiene sobre los fetos. El estudio implica una muestra aleatoria de 3461 no fumadoras y 2238 fumadoras. Todos los sujetos en estudio son mujeres de raza blanca. La variable de interés es el peso en gramos del niño al nacer. Supongamos que esta variable está normalmente distribuida. Se dispone de la siguiente información:

No fumadoras	**Fumadoras**
$n_1 = 3461$	$n_2 = 2238$
$\bar{x}_1 = 3480.1$ g	$\bar{x}_2 = 3256.5$ g
$s_1 = 8.68$ g	$s_2 = 11.02$ g

Basándose en la regla práctica, ¿hay evidencia suficiente de que $\sigma_1^2 \neq \sigma_2^2$?

4. Utilizar la Tabla IX del Apéndice B para hallar cada uno de los valores siguientes: [En los apartados *d* a *k,* el subíndice del punto / indica el área a la derecha del punto. Por ejemplo, $f_{0.05}(15,12\ \text{GL})$ indica el punto asociado con la curva $\mathbf{F}_{15,12}$ que deja un área de 0.05 a la derecha y 0.95 a la izquierda].
   a) $P[F_{24,15} \leq 2.29]$
   b) $P[F_{20,3} \leq 14.17]$
   c) $P[F_{\infty,29} \leq 2.03]$
   d) $f_{0.05}(15, 12\ \text{GL})$
   e) $f_{0.01}(30, 5\ \text{GL})$
   d) $f_{0.1}(40, 9\ \text{GL})$
   g) $f_{0.1}(50, 9\ \text{GL})$
   h) $f_{0.9}(24, 15\ \text{GL})$
   i) $f_{0.95}(40, 30\ \text{GL})$
   j) $f_{0.975}(20, 20\ \text{GL})$
   k) $f_{0.99}(20, 35\ \text{GL})$
5. En cada ejemplo de la Figura 9.5, hallar los puntos *ayb* indicados.
6. En un estudio del metabolismo de los hidratos de carbono, se compara el crecimiento de la raíz en guisantes cultivados en agua a 6 °C con el de las plantas cultivadas en una solución de fructosa a la misma temperatura. Se piensa que la varianza es mayor entre las plantas cultivadas en agua y se dispone de la siguiente información (supóngase normalidad):

Cultivo en agua	**Cultivo en fructosa**
$n_1 = 16$	$n_2 = 25$
$\bar{x}_1 = 9.48$ mm/120 h	$\bar{x}_2 = 9.46$ mm/120 h
$s_1 = 0.53$	$s_2 = 0.25$

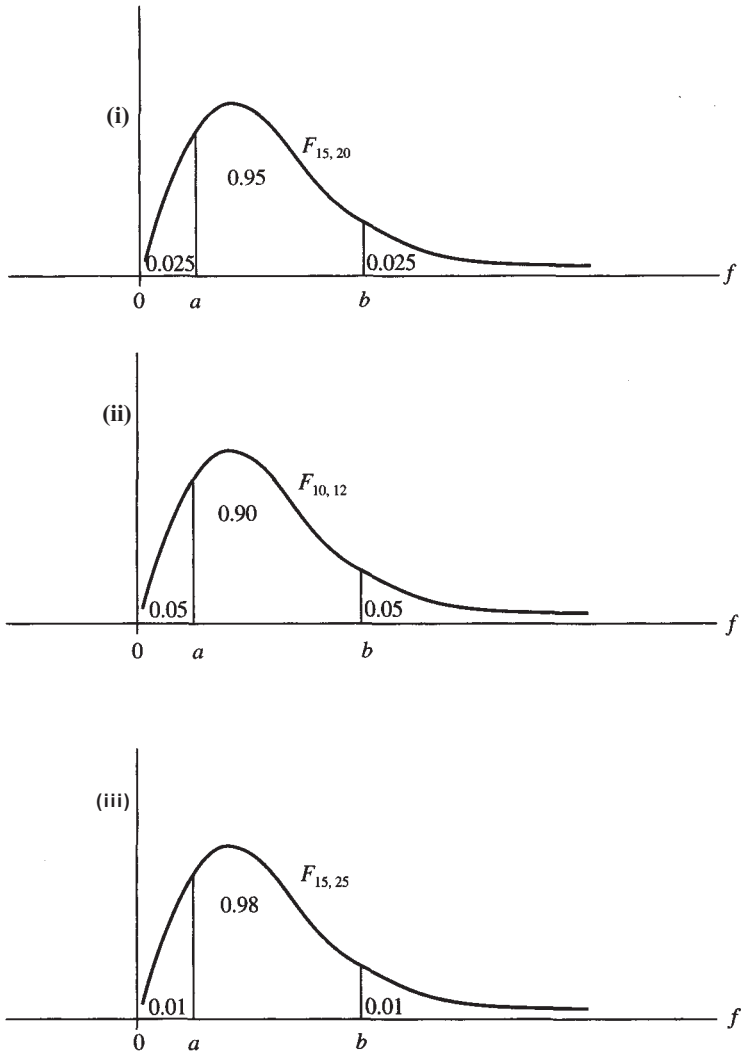


Figura 9.5

Basándose en el contraste de la *F,* ¿se puede concluir que $\sigma_1^2 > \sigma_2^2$? ¿Cuál es el valor P aproximado para el contraste?

7. Considérese el estudio descrito en el Ejercicio 1 de la Sección 9.1. Se piensa que a varianza para las mujeres es mayor que para los varones. Hallar s_H^2 y s_M^2, las varianzas muestrales para el porcentaje de peso corporal perdido por los varones y las mujeres, respectivamente. Basándose en el contraste de la *F,* ¿se puede concluir que, al nivel $\alpha = 0.05$, $\sigma_M^2 > \sigma_H^2$? (Supóngase normalidad.)
8. Considérese el estudio descrito en el Ejercicio 2 de la Sección 9.1. Se piensa que la varianza para las nadadoras es mayor que para las corredoras. Obtener s_N^2 y s_C^2, las varianzas muestrales para el total de partes grasas del cuerpo para las corredoras y nadadoras olímpicas, respectivamente. Hallar s_N^2/s_C^2. Basándose en el contraste de la *F,* ¿se puede concluir al nivel $\alpha = 0.05$ que $\sigma_N^2 > \sigma_C^2$? (Supóngase normalidad.)

9. La regla práctica está diseñada de modo que sus resultados sean consecuentes con los obtenidos por el contraste de la *F* para muestras de tamaño grande y mediano con un nivel $\alpha = 0.20$; en el caso de muestras de tamaños menores, la regla práctica es más tolerante que el contraste de la *F* en el sentido de que establecerá como desiguales a las varianzas en casos en que el contraste de la *F* no lo haría. Considere la distribución *F* para *F(f)* = 0.9. En el contraste:

$$H_0: \sigma_1^2 = \sigma_2^2$$

esta tabla da valores sobre los cuales s_1^2/s_2^2 debe mentir al rechazar H_0 con un nivel a = 0.20.

a) Considere $F_{15,15}$. Si $s_1^2/s_2^2 = 3.00$, ¿rechazará H_0 la regla práctica? ¿Rechazará H_0 el contraste de la F?

b) Considere $F_{10\ 10}$. Si $s_1^2/s_2^2 = 2.0$, ¿rechazará H_0 la regla práctica? ¿Rechazará H_0 el contraste de la F?

c) Dé otro ejemplo en el que la regla práctica y el contraste de la F coincidan.

d) Dé otro ejemplo en el que la regla práctica rechace H_0 pero el contraste de la *F* no lo haga.

e) ¿Hay algún caso en el que la regla práctica no rechace H_0 pero el contraste de la *F* sí? Explíquelo.

9.3. INFERENCIAS SOBRE $\mu_1 - \mu_2$: *T* CONJUNTA

Recordemos que hay dos maneras de comparar las medias: proceder como si las varianzas poblacionales, aunque desconocidas, fuesen iguales, o suponer que difieren entre sí. Haremos que los datos sean los que guíen nuestra decisión. De este modo, nuestra aproximación para comparar medias será:

1. Contrastamos H_0: $\sigma_1^2 = \sigma_2^2$ frente a H_1: $\sigma_1^2 \neq \sigma_2^2$ mediante la regla práctica discutida en la Sección 9.2.
2. Si no rechazamos H_0, procederemos como si las varianzas poblacionales fueran iguales. Se compararán las medias poblacionales utilizando el procedimiento *T* de «varianza conjunta».
3. Si rechazamos H_0, es evidente que las varianzas poblacionales no son iguales. Se compararán las medias poblacionales utilizando el procedimiento T de Smith-Satterthwaite, que se comentará en la Sección 9.4.

Como en anteriores ocasiones, presentaremos técnicas para estimación por intervalos de confianza y por contraste de hipótesis. Empezaremos considerando estos procedimientos en el caso de que se suponga que las varianzas poblacionales son iguales.

Estimación por intervalo de $\mu_1 - \mu_2$

Comenzamos desarrollando los límites para un intervalo de confianza de la diferencia de las medias poblacionales.

> **Ejemplo 9.3.1**. En un estudio sobre angina de pecho en ratas, se dividió aleatoriamente a 18 animales afectados en dos grupos de 9 individuos cada uno. A un grupo se le suministró un placebo y al otro un fármaco experimental FL113. Después de un ejercicio controlado sobré una «cinta sin fin», se determinó el tiempo de recuperación de cada rata. Se

piensa que el FL113 reducirá el tiempo medio de recuperación. Se dispone de la siguiente información:

Placebo	FL113
$n_1 = 9$	$n_2 = 9$
$\bar{x}_1 = 329$ segundos	$\bar{x}_2 = 283$ segundos
$s_1 = 45$ segundos	$s_2 = 43$ segundos

El cociente $s_1^2/s_2^2 = 45^2/43^2 = 1.09$ se utiliza para comparar varianzas. Como este cociente es inferior a 2, por la regla práctica hay pocas evidencias de que σ_1^2 y σ_2^2 sean diferentes. Por tanto, comparando medias, suponemos que las varianzas poblacionales, aunque desconocidas, son iguales. Una estimación puntual para la diferencia en el tiempo medio de recuperación es $\bar{x}_1 - \bar{x}_2 = 329 - 283 = 46$ segundos.

Se sabe que $\bar{X}_1 - \bar{X}_2$ es el estimador puntual lógico para $\mu_1 - \mu_2$. Para extender esta estimación puntual a un intervalo de confianza, debemos hallar una vez más una variable aleatoria cuya expresión contenga el parámetro de interés, en este caso $\mu_1 - \mu_2$, y cuya distribución sea conocida. El Teorema 9.1.1 proporciona tal variable. Este teorema establece que, cuando se muestrean poblaciones normales, la variable aleatoria $\bar{X}_1 - \bar{X}_2$ es normal con media $\mu_1 - \mu_2$ y varia$\sigma_1^2/n_1 + \sigma_2^2/n_2$. ficando esta variable, puede concluirse que la variable aleatoria

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

es normal tipificada.

Si se han comparado las varianzas poblacionales y no se ha detectado diferencia entre ellas, no hay otra alternativa que suponer que son iguales. Sea σ^2 su varianza poblacional común. Es decir, sea $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Sustituyendo en la expresión superior, concluimos que la variable aleatoria:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}$$

es normal tipificada. Puesto que σ^2 no se conoce, ha de ser estimada a partir de los datos, lo que se hace mediante una varianza muestral *conjunta.* Obsérvese que ya teníamos dos estimadores para σ^2, a saber, S_1^2 y S_2^2. La idea es agrupar, o combinar, estos estimadores para que formen un único estimador para σ^2 de tal forma que se tengan en cuenta los tamaños de las muestras. Es natural atribuir mayor importancia, o «peso», a la varianza muestral asociada con la muestra más grande. La varianza conjunta hace exactamente eso. La definimos del siguiente modo:

Definición 9.3.1. Varianza conjunta. Sean S_1^2 y S_2^2, varianzas muestrales basadas en muestras independientes de tamaños n_1 y n_2, respectivamente. La *varianza conjunta,* que se designa por S_p^2, viene dada por

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Obsérvese que ponderamos S_1^2 y S_2^2 multiplicando por $n_1 - 1$ y $n_2 - 1$, respectivamente. La forma más natural para ponderar es multiplicar por los correspondientes tamaños muestra-

les n_1 y n_2. Elegimos esta forma tan heterodoxa para que la variable aleatoria $(n_1 + n_2 - 2)$ S_p^2/σ^2 siga una distribución ji-cuadrado. Es necesario que el estadístico que utilizamos para contrastar la igualdad de medias siga una distribución *T*.

Ejemplo 9.3.2. Consideremos una varianza muestral $s_1^2 = 24$ de una muestra de tamaño 16 y una segunda varianza muestral $s_2^2 = 20$ de una muestra de tamaño 121. El valor del cociente s_1^2/s_2^2 es 24/20 = 1.20. Como la regla práctica no nos hace pensar que σ_1^2 y σ_2^2 sean diferentes, los valores de s_1^2 y s_2^2 son estimadores de la misma varianza, σ^2.

La estimación conjunta para la varianza poblacional común es

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{15(24) + 120(20)}{16 + 121 - 2}$$

$$= \frac{2760}{135} = 20.44$$

Obsérvese que esta estimación es completamente diferente del valor 22, que es el que se obtendría ignorando los tamaños de las muestras y promediando aritméticamente s_1^2 y s_2^2.

Para obtener una variable aleatoria que pueda ser utilizada para construir un intervalo de confianza de $\mu_1 - \mu_2$, reemplazamos la varianza poblacional desconocida σ^2 en la variable aleatoria Z

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}$$

por el estimador conjunto S_p^2 para obtener la variable aleatoria

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$

Como en el caso de una muestra, reemplazar la varianza poblacional por su estimador afecta a la distribución. La primera variable aleatoria es una variable Z; la última tiene una distribución Tcon $n_1 + n_2 - 2$ grados de libertad. La estructura algebraica de esta variable es la misma que la hallada previamente, es decir:

$$\frac{\text{Estimador - parámetro}}{D}$$

donde *D* representa la desviación típica del estimador del numerador o el estimador de esta desviación típica. Por tanto, el intervalo de confianza para $\mu_1 - \mu_2$ toma la misma forma general que la mayor parte de los intervalos encontrados previamente.

Teorema 9.3.1. Intervalo de confianza de $\mu_1 - \mu_2$: varianza conjunta. Sean $\bar{X}_1$ y $\bar{X}_2$ medias muestrales basadas en muestras aleatorias simples extraídas de distribuciones normales con medias μ_1 y μ_2, respectivamente, y varianza común σ^2. Designemos por S_p^2 la varianza muestral conjunta. Los límites para un intervalo de confianza de $\mu_1 - \mu_2$ son

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

donde el punto *t* se calcula con respecto a la distribución $T_{n_1 + n_2 - 2}$.

Ejemplo 9.3.3. En un estudio sobre hábitos de alimentación en murciélagos, se marcan *25* hembras y 11 machos y se les rastrea por radio. Una variable de interés es *X,* la distancia que recorren volando en una pasada en busca de alimento. El experimento proporcionó la siguiente información (suponiendo normalidad):

Hembras	Machos
$n_1 = 25$	$n_2 = 11$
$\bar{x}_1 = 205$ metros	$\bar{x}_2 = 135$ metros
$s_1 = 100$ metros	$s_2 = 95$ metros

Obsérvese que $s_1^2/s_2^2 = 100^2/95^2 = 1.11$. Como el cociente es inferior a 2, la regla práctica np indica que existan diferencias entre las varianzas poblacionales. Puesto que no se pueden detectar diferencias en las varianzas poblacionales, promediamos s_1^2 y s_2^2 para obtener:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{24(100^2) + 10(95^2)}{25 + 11 - 2} = 9713.24$$

Hallemos un intervalo de confianza, del 90 %, de $\mu_1 - \mu_2$ a fin de comparar las medias. El número de grados de libertad necesarios es $n_1 + n_2 - 2 = 25 + 11 - 2 = 34$. En la Figura 9.6 aparece la partición de la curva T_{34}. Los límites para el intervalo de confianza son:

$$(\bar{x}_1 - \bar{x}_2) \pm t\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = (205 - 135) \pm 1.697\sqrt{9713.24\left(\frac{1}{25} + \frac{1}{22}\right)}$$

$$= 70 \pm 60.51$$

Podemos tener un 90 % de confianza en que la diferencia en las distancias medias recorridas en busca de alimento entre murciélagos hembras y machos está entre 9.49 y 130.51 metros. El intervalo no contiene al número 0 y es positivo, lo que indica que la distancia media recorrida por las hembras es mayor que la recorrida por los machos. Algunos biólogos han interpretado que esto significaría que las hembras están siendo expulsadas de las áreas de alimentación más cercanas por los machos más agresivos. Sin embargo, esta teoría no ha sido confirmada.

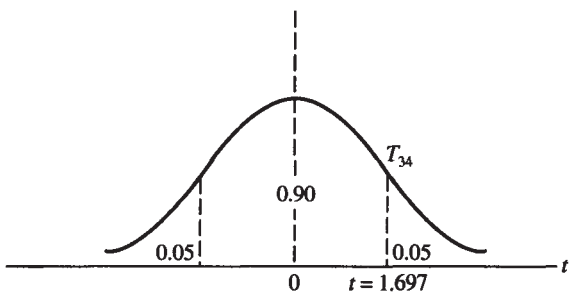


Figura 9.6. Partición de la curva T_{34} que se necesita para obtener un intervalo de confianza de $\mu_1 - \mu_2$, del 90%.

Contraste Tde varianza conjunta

Como en ejemplos previos, la variable aleatoria utilizada para obtener límites de confianza para un parámetro, también sirve como contraste estadístico para contrastar hipótesis relativas al parámetro. En este caso, la variable aleatoria:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} = T_{n_1+n_2-2}$$

(Contraste T sobre $\mu_1 - \mu_2$: varianza conjunta)

sirve como estadístico para cualquiera de los contrastes de hipótesis usuales, en donde $(\mu_1 - \mu_2)_0$ designa el valor asignado por hipótesis a la diferencia de las medias poblacionales. Esta diferencia puede tomar cualquier valor, pero el más común es el cero. En este caso, lo que se pretende es determinar si las medias poblacionales difieren y, si es así, cuál es la mayor. Las hipótesis adoptan las formas siguientes:

I	II	III
H_0: $\mu_1 \leq \mu_2$	H_0: $\mu_1 \geq \mu_2$	H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$	H_1: $\mu_1 < \mu_2$	H_1: $\mu_1 \neq \mu_2$
Contraste con cola a la derecha	Contraste con cola a la izquierda	Contraste con dos colas

Puesto que el estadístico del contraste utilizado para distinguir entre H_0 y H_1 emplea el estimador conjunto de σ^2, el contraste T mostrado en esta sección se denomina contraste T de varianza conjunta.

Ejemplo 9.3.4. La recogida de datos para el estudio de la angina de pecho en ratas, del Ejemplo 9.3.1, da los siguientes resultados:

Placebo	FL113
$n_1 = 9$	$n_2 = 9$
$\bar{x}_1 = 329$ segundos	$\bar{x}_2 = 283$ segundos
$s_1 = 45$ segundos	$s_2 = 43$ segundos

La estimación puntual para la diferencia en tiempo medio de recuperación entre los que reciben un placebo y los que reciben el fármaco experimental es $\bar{x}_1 - \bar{x}_2 = 46$ segundos. ¿Es esta diferencia lo suficientemente grande para concluir que el fármaco experimental tiende a reducir el tiempo de recuperación? Para responder a esta pregunta, debemos hallar la estimación conjunta para la varianza poblacional común. Esta estimación es:

$$s_p^2 = \frac{(n-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{8(45)^2 + 8(43)^2}{9+9-2}$$

$$= \frac{45^2 + 43^2}{2} = 1937$$

Obsérvese que, puesto que los tamaños de las muestras son los *mismos,* en este caso la estimación conjunta es la media aritmética de las estimaciones individuales s_1^2 y s_2^2. Ahora evaluamos el estadístico:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2(1/n_1 + 1/n_2)}} = \frac{46 - 0}{\sqrt{1937(\frac{1}{9} + \frac{1}{9})}}$$
$$= 2.22$$

El número de grados de libertad asociados con el estadístico T de varianza conjunta es $n_l + n_2 - 2 = 16$. La hipótesis de investigación es que el FL113 reducirá el tiempo medio de recuperación (H_1: $\mu_1 > \mu_2$). Ya que el contraste es con cola a la derecha:

$$P = P[T_{16} > 2.22]$$

Y dado que el valor observado del estadístico (2.22) esté situado entre 2.120 y 2.583 y el valor P entre 0.01 y 0.025, estamos en condiciones de mantener el argumento de que el fármaco experimental es eficaz en la reducción del tiempo de recuperación en ratas con angina de pecho.

EJERCICIOS 9.3

1. *a)* Sea $s_1^2 = 42$, $s_2^2 = 37$, $n_l = 10$, $n_2 = 14$. Hallar s_p^2.
   b) Sea $s_1^2 = 28$, $s_2^2 = 30$, $n_l = 20$, $n_2 = 20$. Hallar s_p^2. (¡No utilizar la calculadora!)
   c) Sea $s_1^2 = 20$, $s_2^2 = 40$, $n_l = 10$, $n_2 = 50$. Hallar s_p^2. ¿Por qué está s_p^2 más próximo al valor de s_2^2 que al de s_1^2?
2. Se estudian los hábitos de alimentación de dos especies de arañas. Estas especies, *Dinopis* y *Menneus,* coexisten en Australia oriental. Una variable de interés es el tamaño de las presas de cada especie. La araña *Menneus* adulta tiene aproximadamente el mismo tamaño que la araña *Dinopis* joven. Se sabe que existe una diferencia entre el tamaño de las presas en *Dinopis* adulta y joven a causa de la diferencia entre sus tamaños. ¿Hay una diferencia en el tamaño medio de las presas entre *Dinopis* adulta y *Menneus* adulta? Si es así, ¿cuál es la causa? Para responder a estas preguntas, se obtuvieron las siguientes observaciones sobre el tamaño, en milímetros, de las presas de las dos especies:

Dinopis adulta		*Menneus* adulta	
12.9	11.9	10.2	5.3
10.2	7.1	6.9	7.5
7.4	9.9	10.9	10.3
7.0	14.4	11.0	9.2
10.5	11.3	10.1	8.8

   a) Usar la regla práctica para comparar las varianzas poblacionales.
   b) Si no se detectan diferencias en las varianzas de las poblaciones, hallar un intervalo de confianza del 90 % para $\mu_1 - \mu_2$.
   c) Basándose en el intervalo de la parte *b,* ¿hay constancia de una diferencia en el tamaño medio de las presas entre las dos especies? Razonar la respuesta. (Los

biólogos piensan que cualquier diferencia detectada se explica por diferencias en situación y tamaño de las telarañas tejidas por cada una.)

Considérense los datos del Ejercicio 1 de la Sección 9.1. Úsese la regla práctica para comprobar si existen diferencias entre las varianzas poblacionales. Si es apropiado, hallar un intervalo de confianza del 95 % para la diferencia entre las medias poblacionales.

Considérense los datos del Ejercicio 2 de la Sección 9.1. Supuesta la normalidad, úsese la regla práctica para comprobar si existen diferencias entre las varianzas poblacionales. Si es apropiado, hallar s_p^2 y un intervalo de confianza del 98 % para la diferencia de medias poblacionales. Basándose en este intervalo, ¿piensa el lector que existe diferencia entre la media total de partes grasas del cuerpo de las corredoras y nadadoras olímpicas? Explíquelo.

Se lleva a cabo un estudio de dos tipos de tratamiento por fármacos para su utilización potencial en trasplantes de corazón. El fin de los fármacos es actuar como inraunosupresores: reprimir la tendencia natural del cuerpo a rechazar el trasplante. Las ratas ACI machos sirven como donantes, y las ratas noruegas Lewis-Brown machos como receptoras. Estas ratas son conocidas por su escasa propensión a la compatibilidad. La variable de interés es *X,* tiempo de supervivencia en días. Se obtuvieron los siguientes resultados estadísticos:

Salicilato sódico solamente	Salicilato sódico y azatioprina
$n_1 = 9$	$n_2 = 9$
$\bar{x}_1 = 16$ días	$\bar{x}_2 = 15$ días
$s_1 = 10.1$ días	$s_2 = 10$ días

Utilizar esta información para comparar las varianzas poblacionales. Hallar un intervalo de confianza del 90 % para la diferencia en los tiempos medios de supervivencia entre los dos tratamientos. Interpretar este intervalo en función de sus implicaciones prácticas.

6. Puesto que un nivel de colesterol elevado es un factor de alto riesgo en el desarrollo de la aterosclerosis cardíaca y coronaria, es importante determinar los niveles a esperar en los diferentes grupos de edad. Se realizó un estudio para comparar el nivel de colesterol en varones de entre 20 y 29 años, frente a mujeres del mismo grupo de edad. Se obtuvieron los datos siguientes:

Varones	Mujeres
$n_1 = 96$	$n_2 = 85$
$\bar{x}_1 = 167.16$ mg/dL	$\bar{x}_2 = 178.12$ mg/dL
$s_1 = 30$ mg/dL	$s_2 = 32$ mg/dL

(a) Comprobar si hay diferencias en las varianzas poblacionales.

b) Hallar un intervalo de confianza del 95 % para la diferencia entre los niveles medios de colesterol entre estos dos grupos.

(Basado en los estudios descritos en Marvin Bell y Sharon Joseph, «Community Screening for Hypercholesterolemia», *Journal of Family Practice,* octubre de 1990, págs. 365-368.)

7. Se utilizaron 23 vegetarianos en un estudio de enfermedad diverticular y dieta. Una variable de interés fue el total de fibra de la dieta. Se obtuvo la siguiente información para dos grupos, los que no tenían la enfermedad y los que la tenían (supóngase normalidad):

Sin	Con
$n_1 = 18$	$n_2 = 5$
$\bar{x}_1 = 42.7$ g	$\bar{x}_2 = 27.7$ g
$s_1 = 9.9$ g	$s_2 = 9.5$ g

Comprobar si hay diferencias en las varianzas poblacionales. ¿Hay razón suficiente para pretender que la media total de contenido de fibra en las dietas de los que no tienen la enfermedad es más alta que en la de aquellos que la tienen? Explicar la respuesta a partir del valor *P* del contraste.

8. Otra variable de interés en el estudio de la angina de pecho en las ratas (véase el Ejemplo 9.3.1) es el consumo de oxígeno, medido en mililitros por minuto. El experimento proporcionó la siguiente información:

Placebo	FL113
$n_1 = 9$	$n_2 = 9$
$\bar{x}_1 = 1509$ mL/min	$\bar{x}_2 = 1702$ mL/min
$s_1 = 169$ mL/min	$s_2 = 181$ mL/min

Utilizar la información para comparar las varianzas poblacionales. Basándose en el experimento, ¿hay razón suficiente para pretender que el consumo de oxígeno de las ratas que toman FL113 es más elevado que el de las que toman el placebo? Explicar la respuesta tomando como base el valor *P* del contraste.

9. En un estudio de características corporales de las gaviotas de pico anillado, la variable considerada es la longitud del pico. Se dispone de los siguientes datos:

Hembras	Machos
$n_1 = 51$	$n_2 = 41$
$\bar{x}_1 = 59.1$ mm	$\bar{x}_2 = 65.2$ mm
$s_1 = 1.9$ mm	$s_2 = 2.0$ mm

Utilizar esta información para detectar diferencias en las varianzas de la población ¿Hay evidencia para sostener el argumento de que la longitud media del pico en los machos es mayor que en las hembras? Explicar la respuesta basándose en el valor *P* del contraste.

10. Una variable que se utiliza para comparar las características físicas de las nadadoras olímpicas con los de las corredoras es la circunferencia de la parte superior del brazo, en centímetros, mientras están relajadas. Se dispone de los siguientes datos:

Nadadoras	Corredoras
$n_1 = 10$	$n_2 = 12$
$x_1 = 27.3$ cm	$x_2 = 23.5$ cm
$s_1 = 1.9$ cm	$s_2 = 1.7$ cm

Supuesta la normalidad, contrastar la igualdad de varianzas. ¿Hay razón suficiente para pretender que la media de la circunferencia de la parte superior del brazo es mayor en las nadadoras que en las corredoras? Explicar la respuesta basándose en el valor *P* del contraste.

11. Se cree que los jóvenes adolescentes que fuman comienzan a hacerlo a una edad más temprana que las chicas adolescentes fumadoras. ¿Apoyan los datos esta suposición?

Chicos	Chicas
$n_1 = 33$ Edad media a la que empiezan a fumar = 11.3 años $s_1^2 = 4$ g	$n_2 = 14$ Edad media a la que empiezan a fumar = 12.6 años $s_2^2 = 3.5$ g

(Medias halladas en Nadu Tuakli, Mindy Smith y Caryl Heaton, «Smoking in Adolescence: Methods for Health Education and Smoking Cessation», *Journal of Family Practice,* octubre de 1990, págs. 369-373.)

9.4. INFERENCIAS SOBRE $\mu_1 - \mu_2$: VARIANZAS DISTINTAS

Si, cuando se comparan las varianzas de la población, se detecta una diferencia, resulta inapropiado promediarlas. Pero todavía es posible comparar medias utilizando un estadístico *T* aproximado. Nuevamente, el estadístico se halla modificando la variable Z:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

de forma lógica. Puesto que ahora hay evidencia de que $\sigma_1^2 \neq \sigma_2^2$, cada varianza poblacional se estima por separado; no se hace una estimación conjunta. Las varianzas poblacionales en la variable aleatoria Z anterior son reemplazadas por sus estimadores respectivos, S_1^2 y S_2^2, para obtener la variable aleatoria

$$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}}$$

Como ocurría antes, este cambio conduce a una variación en la distribución de Z que se aproxima a una *T*. El número de grados puede estimarse a partir de los datos. Se han sugerido varios métodos para hacerlo. Aquí emplearemos el procedimiento de Smith-Satterthwaite. De acuerdo con este procedimiento, γ, el número de grados de libertad, viene dado por

$$\boxed{\gamma \cong \frac{[S_1^2/n_1 + S_2^2/n_2]^2}{\dfrac{[S_1^2/n_1]^2}{n_1 - 1} + \dfrac{[S_2^2/n_2]^2}{n_2 - 1}}}$$

No es necesario que el valor y sea entero. Si no lo es, redondeamos por *defecto* al entero más próximo. Lo redondeamos por defecto en lugar de por exceso para tener una aproximación conservadora. Recuérdese que, cuando el número de grados de libertad asociado a la variable aleatoria T aumenta, las curvas acampanadas correspondientes se hacen más compactas. En la práctica, esto significa que, por ejemplo, el punto asociado con la curva T_{10} con un 5 % de área a la derecha (1.812) es un poco mayor que el punto asociado con la curva T_{11} con un 5 % de área a la derecha (1.796). Como resultado de este planteamiento conservador, un intervalo de confianza basado en la distribución T_{10} es un poco más grande que uno basado en la curva T_{11}. Además, si podemos rechazar la hipótesis nula basándose en la distribución T_{10}, también lo será basándose en la distribución T_{10}. El recíproco no se cumple necesariamente.

Los límites de confianza para un intervalo de confianza de $\mu_1 - \mu_2$ cuando las varianzas son desiguales son similares a los hallados anteriormente. Éstos son:

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

(Intervalo de confianza para $\mu_1 - \mu_2$: varianzas desiguales)

donde t es un punto basado en la distribución T con y grados de libertad. Los grados de libertad se calculan utilizando la fórmula de Smith-Satterthwaite.

El estadístico para contrastar H_0: $(\mu_1 - \mu_2) = (\mu_1 - \mu_2)_0$ frente a cualquiera de las tres alternativas habituales es:

$$T_\gamma = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_1^2/n_1 + S_1^2/n_2}}$$

(Contraste T para $\mu_1 - \mu_2$. varianzas desiguales)

donde γ se calcula utilizando la fórmula de Smith-Satterthwaite. Este procedimiento se explica en el Ejemplo 9.4.1.

Ejemplo 9.4.1. Se realizó un estudio sobre las necesidades energéticas para el crecimiento y mantenimiento de un nido de «aviones» en Perthshire, Escocia. Se obtuvieron los siguientes resultados estadísticos para la variable normal X, número de kilocalorías por gramo y hora que se requieren por pájaro:

Adultos incubando	**Adultos precriando**
$n_1 = 57$	$n_2 = 12$
$\bar{x}_1 = 0.0167$ kcal/(g)(h)	$\bar{x}_2 = 0.0144$ kcal/(g)(h)
$s_1 = 0.0042$ kcal/(g)(h)	$s_2 = 0.0024$ kcal/(g)(h)

¿Indican estos datos que el número medio de kilocalorías requerido por los adultos que están incubando es mayor que el requerido por los adultos que están precriando?

El valor observado del estadístico S_1^2/S_2^2 es $(0.0042)^2/(0.0024)^2 = 3.0625$. Como este valor excede de 2, hay evidencia de que $\sigma_1^2 \neq \sigma_2^2$. No promediaremos las varianzas muestrales puesto que no estiman una varianza poblacional común. El valor del estadístico T_γ es:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{0.0167 - 0.0144}{\sqrt{(0.0042)^2/57 + (0.0024)^2/12}} = 2.59$$

El número de grados de libertad asociado con este estadístico es:

$$\gamma \cong \frac{[s_1^2/n_1 + s_2^2/n_2]^2}{\dfrac{[s_1^2/n_1]^2}{n_1 - 1} + \dfrac{[s_2^2/n_2]^2}{n_2 - 1}}$$

$$= \frac{[(0.0042)^2/57 + (0.0024)^2/12]^2}{[(0.0042)^2/57]^2/56 + [(0.0024)^2/12]^2/11} = 27.5$$

Ya que el número de grados de libertad debe ser un entero positivo, redondeamos por defecto a 27. Basándose en la distribución T_{27}, la probabilidad de obtener un valor de 2.59 ó mayor está entre 0.01 y 0.005. Es decir, el valor P para el contraste de $\mathbf{H_0}$: $\mu_1 \leq \mu_2$ está entre 0.01 y 0.005. Dado que este valor es muy pequeño, concluimos que la energía media requerida por los pájaros que están incubando es mayor que para los que están precriando.

Tanto la calculadora TI83 como el programa SAS están programados para realizar el contraste de la T conjunta y el contraste de la T de Smith-Satterthwaite. Cuando se utilizan estas herramientas, es tarea del investigador decidir cuál de los dos contrastes es el apropiado. Las herramientas computacionales hacen el trabajo de cálculo por usted, no interpretan los resultados.

EJERCICIOS 9.4

1. Considérese el Ejemplo 9.4.1. Si uno falla al comprobar la igualdad de las varianzas y equivocadamente las promedia, ¿cuál sería el valor observado del estadístico? ¿Cuántos grados de libertad se utilizarían? ¿Estaría afectado el valor P?
2. El estroncio 90, un elemento radiactivo que se produce en las pruebas nucleares, está estrechamente relacionado con el calcio. En las tierras especializadas en producción de leche, el estroncio 90 se localiza en la leche, a través de los pastos ingeridos por las vacas lecheras. Después se concentra en los huesos de los consumidores de leche. En 1959, se llevó a cabo un estudio para comparar la concentración media de estroncio 90 en los huesos de los niños con la de los adultos. Se pensó que el nivel en los niños era mayor a causa de que la sustancia estaba presente durante sus años de formación. ¿Está este argumento sostenido por los siguientes datos? Expliqúese (supóngase normalidad).

Niños	Adultos
$n_1 = 121$	$n_2 = 61$
$\bar{x}_1 = 2.6$ picocurios por gramo	$\bar{x}_2 = 0.4$ picocurios por gramo
$s_1 = 1.2$ picocurios por gramo	$s_2 = 0.11$ picocurios por gramo

3. Se realiza un estudio para comparar las tortugas de Malabar con las dé Grande-Terre, islas del atolón Aldabra, en él Océano Índico. Una variable de interés es X, peso de un huevo en el momento de la puesta. Muestras aleatoriamente seleccionadas de las dos islas proporcionaron los siguientes resultados (supóngase normalidad):

Grande-Terre	Malabar
$n_1 = 31$	$n_2 = 148$
$\bar{x}_1 = 64.0$ gramos	$\bar{x}_2 = 82.7$ gramos
$s_1 = 6.5$ gramos	$s_2 = 3.6$ gramos

¿Hay razones para mantener que el peso medio de un huevo en el momento de la puesta en Malabar es mayor que en Grande-Terre? Expliquese.

4. Se realiza un estudio del tejedor piquirrojo. Con el propósito de comparar una colonia situada en las proximidades del Lago Chad, con otra situada en Botswana. La colonia del Lago Chad se malogró porque los adultos abandonaron los nidos; la colonia de Botswana se mantuvo. Una variable que se cree tiene influencia sobre la capacidad de la hembra para permanecer en el nido es *X,* su nivel proteínico muscular. Se pretenden comparar los niveles proteínicos musculares de las hembras de las dos colonias.
   a) Al principio del ciclo de puesta se recogieron los siguientes datos:

Lago Chad	Botswana
$n_1 = 100$	$n_2 = 100$
$\bar{x}_1 = 0.99$ g	$\bar{x}_2 = 1.00$ g
$s_1 = 0.01$ g	$s_2 = 0.01$ g

   Comparar las varianzas poblacionales y contrastar la igualdad de las medias, utilizando el estadístico *T* apropiado. Respecto a esta variable, ¿parecen ser idénticas las dos poblaciones al inicio del ciclo de puesta?
   b) Al final del ciclo de puesta, se obtuvieron los siguientes datos:

Lago Chad	Bostwana
$n_1 = 100$	$n_2 = 100$
$\bar{x}_1 = 0.87$ g	$\bar{x}_2 = 0.90$ g
$s_1 = 0.02$ g	$s_2 = 0.01$ g

   Comparar las varianzas poblacionales y contrastar la igualdad de medias, utilizando el estadístico *T* apropiado. Respecto a esta variable, ¿parecen ser idénticas ahora las dos poblaciones?

5. Considérense los datos del Ejercicio 3 de la Sección 9.2. El propósito del estudio consiste en confirmar el argumento de que el peso medio de los recién nacidos es inferior entre las madres fumadoras que entre las no fumadoras. ¿Mantienen los datos el argumento? Dar razones que apoyen la elección del contraste.
6. Se lleva a cabo un estudio para investigar la capacidad de los monocitos para destruir ciertas células de levadura halladas en pacientes con cirrosis hepática. Estas células son perjudiciales en el sentido de que hacen al paciente susceptible frente a infecciones recurrentes de diversos tipos. Se toman muestras de sangre de 16 pacientes con cirrosis y de 9 pacientes control. Se obtuvieron los siguientes datos sobre el porcentaje de células de ese tipo eliminadas por los monocitos del cultivo (supóngase normalidad):

Controles	Pacientes
$n_1 = 9$	$n_2 = 16$
$\bar{x}_1 = 44.22\%$	$\bar{x}_2 = 28.22\%$
$s_1 = 6.17\%$	$s_2 = 4.11\%$

Comparar las varianzas poblacionales. Basándose en los resultados de este contraste, comparar las medias poblacionales utilizando el estadístico T apropiado. ¿Hay razón suficiente para pretender que el porcentaje medio de células eliminadas por los monocitos entre los controles es mayor que entre los pacientes? Expliquese.

7. Se ha realizado un estudio para comparar la concentración de plomo en el agua de dos casas. En una casa se utilizó una soldadura con el 50 % de plomo y el 50 % de estaño en las tuberías de agua que llegaban a ella. En la otra casa no se utilizó esta soldadura. Los datos son de las muestras de agua tomadas de la cocina y de los baños de las casas en el momento en que se abrieron los grifos.

Lugar 1 (soldadura con el 50 % de plomo y el 50 % de estaño)	Lugar 2
$n_1 = 25$	$n_2 = 25$
$\bar{x}_1 = 390$ ppb[1]	$\bar{x}_2 = 10$ ppb
$s_1 = 217.5$ ppb	$s_2 = .5$ ppb

Se cree que la concentración media de plomo en el agua en el lugar 1 supera a la del lugar 2. Contrastar esta hipótesis. (Información hallada en E. Cosgrove et al., «Childhood Lead Poisoning», *Journal of Environmental Health,* julio de 1989, págs. 346-349.)

8. Se ha realizado un estudio de pruebas cruzadas preoperatorias en cirugía electiva. La operación estudiada es la histerectomía abdominal electiva. La variable de interés es X, número de unidades sanguíneas compatibles inmediatamente disponibles. El objeto es comparar el número medio de unidades disponibles en 1990 con el disponible actualmente. Disponemos del siguiente resumen de datos:

1990	Actualmente
$n_1 = 25$	$n_2 = 25$
$\bar{x}_1 = 2.73$	$\bar{x}_2 = 1.27$
$s_1 = 0.65$	$s_2 = 1.0$

Tras comparar las varianzas poblacionales, hallar un intervalo de confianza del 95 % de $\mu_1 - \mu_2$. ¿Hay alguna evidencia de que se haya producido un descenso del número medio de unidades disponibles desde 1990 hasta la actualidad? Explicarlo.

9. Considérese los datos del Ejercicio 1 de la Sección 9.2. Hallar un intervalo de confianza del 90% de la diferencia en la velocidad media de vuelo de los pelícanos pardos y los ostreros americanos cuando vuelan con el viento de costado. Basándose en este intervalo, ¿hay evidencias de que exista una diferencia en las medias poblacionales? Explicarlo.

10. Considérese los datos del Ejercicio 6 de la Sección 9.2. Hallar un intervalo de confianza del 95 % para la diferencia en el crecimiento medio de los guisantes cultivados en agua y los cultivados en una solución de fructosa. Basándose en este intervalo, ¿hay evidencias de que exista alguna diferencia en los índices de crecimiento? Explicarlo.

[1] *N del T:* La unidad es partes por billón (ppb), pero recuérdese que en Estados Unidos un billón son mil millones.

11. Se han realizado estudios sobre desechos peligrosos en dos lugares distintos. Los desechos de interés son los generados por las casas familiares y los pequeños negocios. Ejemplos de dichos desechos son el aceite usado de los automóviles, las pilas, los anticongelantes, la pintura y el diluyente de pintura, y los disolventes. Se han obtenido estos datos:

Albuquerque	Anchorage
$n_1 = 96$ 320	$n_2 = 81609$
$\bar{x}_1 = 16.59$ libras al año	$\bar{x}_2 = 22.06$ libras al año
$s_1^2 = 25$	$s_2^2 = 36$

Hallar un intervalo de confianza del 95 % de la diferencia en la cantidad media de libras de desechos peligrosos producidas por unidad de muestreo al año. (Basado en los estudios registrados en David Wigglesworth, «Hazardous Waste Management at the Local Level», *Journal of Environmental Health,* agosto de 1989, págs. 323-326.)

12. La competencia entre especies desempeña un importante papel en el crecimiento de las plantas. Se llevó a cabo un estudio para investigar el efecto de la competencia en el crecimiento de la lechuga. Se plantaron sesenta y cuatro semillas de lechuga, y se dejaron crecer en un medio en el que sólo había lechugas. También se plantaron treinta y cinco semillas de lechuga en un medio en el que las plantas estaban en competencia con otras de espinacas. Al finalizar el estudio, se midió el peso seco en gramos de cada lechuga, obteniéndose los siguientes datos:

Lechugas solas	Lechugas en competencia con espinacas
$n_1 = 64$	$n_2 = 35$
$\bar{x}_1 = 0.030$ g	$\bar{x}_2 = 0.023$ g
$s_1 = 0.018$ g	$s_2 = 0.012$ g

¿Apoyan estos datos la teoría de que el peso seco medio de las plantas crecidas en competencia es menor que el de las plantas crecidas en ausencia de competencia? Dar razones que apoyen la elección del contraste. (Basado en un estudio realizado por Melissa M. Stone, Departamento de Biología, Universidad de Radford, 1996.)

9.5. INFERENCIAS SOBRE $\mu_1 - \mu_2$: *T* PARA DATOS EMPAREJADOS

En muchos ejemplos surgen problemas en los que se dispone de dos muestras aleatorias no independientes; más bien, cada observación en una muestra está naturalmente, o a propósito, emparejada con una observación de la otra. Ejemplos de esta clase de emparejamientos ocurren en estudios realizados con gemelos en psicología, medicina y enseñanza; en estudios que incluyen contraste a priori y a posteriori en enseñanza y educación física; en estudios en los que se administran dos tratamientos al mismo sujeto; y en muchas otras ocasiones. Considérese el Ejemplo 9.5.1.

Ejemplo 9.5.1. Se ha realizado un estudio para investigar el efecto del ejercicio físico en el nivel de colesterol en plasma, en el que participaron once sujetos. Ante del ejercicio, se

tomaron muestras de sangre para determinar el nivel del colesterol de cada participante. Después, los individuos fueron sometidos a un programa de ejercicios que se centraba en carreras y marchas diarias. Al final del periodo de ejercicios, se tomaron nuevamente muestras de sangre y se obtuvo una segunda lectura del nivel de colesterol en plasma. De este modo, se dispone de dos conjuntos de observaciones del nivel de colesterol en plasma de los sujetos. Los conjuntos de datos no son independientes; se basan en los mismos sujetos tomados en diferentes momentos y, por tanto, están naturalmente emparejados para cada uno. Se recogieron los siguientes datos:

Sujeto	Nivel previo *x*, mg/dL	Nivel posterior y, mg/dL
1	182	198
2	232	210
3	191	194
4	200	220
5	148	138
6	249	220
7	276	219
8	213	161
9	241	210
10	480	313
11	262	226

Se pretende estimar la diferencia entre el nivel medio de colesterol antes y después del ejercicio.

Cuando se tiene un emparejamiento tal como el que se acaba de exponer, no son aplicables los métodos de las Secciones 9.3 y 9.4. Cualquier procedimiento utilizado para comparar medias deberá tener en cuenta ahora el hecho de que las observaciones están emparejadas. Esto es fácil de hacer. Considérese la generalización del problema indicado en la Figura 9.7. Obsérvese que hay, asociada con esta situación, una diferencia de poblaciones $D = X - Y$ y una muestra aleatoria de diferencias que se seleccionan de esta población, $D_i = X_i - Y_i$, $i = 1, 2, 3, \ldots, n$. (Véase Fig. 9.8.)

El valor medio de la diferencia D es la diferencia de los valores medios de X e Y. Esto es,

$$\mu_D = \mu_X - \mu_Y$$

Por lo tanto, la pregunta original, ¿qué es $\mu_X - \mu_Y$?, es equivalente a la pregunta, ¿qué es μ_D? Hemos reducido el problema original de dos muestras al problema de *una* muestra que consiste en hacer una inferencia sobre la media de la población de diferencias. Este problema no es nuevo, y puede abordarse utilizando los métodos del Capítulo 6. En particular, la fórmula para los límites de confianza de $\mu_X - \mu_Y = \mu_D$ es

$$\overline{D} \pm \frac{tS_d}{\sqrt{n}}$$

(Intervalo de confianza para la media de la diferencia)

donde $\overline{D}$ y S_d son la media muestral y la desviación típica muestral de la muestra de diferencias, respectivamente, y t el punto apropiado relativo a la distribución T_{n-1}. En el Ejemplo 9.5.2 se desarrolla la utilización de esta fórmula.

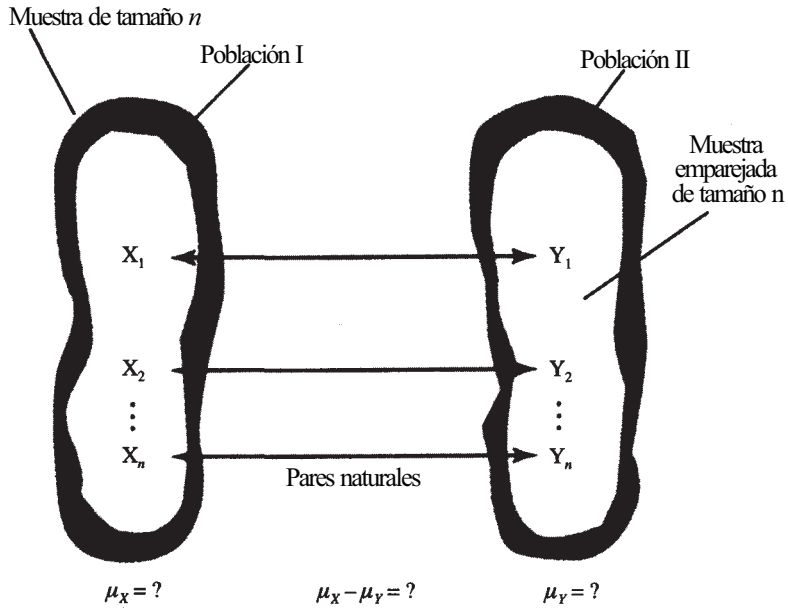


Figura 9.7. Datos emparejados.

Ejemplo 9.5.2. Considérense los datos del Ejemplo 9.5.1, y fórmese la muestra de diferencias sustrayendo la segunda lectura de colesterol de la primera.

Sujeto	Previo *x*	Posterior y	Diferencia d = *x* -*y*
1	182	198	-16
2	232	**210**	22
3	191	194	-3
4	200	220	-20
5	**148**	138	10
6	249	220	29
7	276	219	57
8	213	161	52
9	241	210	31
10	480	313	167
11	262	226	36

Para construir un intervalo de confianza de μ_D del 90 %, necesitamos calcular la media muestral y la desviación típica muestral para el conjunto de diferencias:

Para estos datos

$$\bar{d} = 33.2 \qquad s_d = 51.1$$

La partición de la c u $_1T_{n-1} = T_{10}$ e necesitamos es la de la Figura 9.9. Los límites de confianza son:

$$\bar{d} \pm t\frac{s_d}{\sqrt{n}} = 33.2 \pm 1.812\,\frac{51.1}{\sqrt{11}}$$

$$= 33.2 \pm 27.9$$

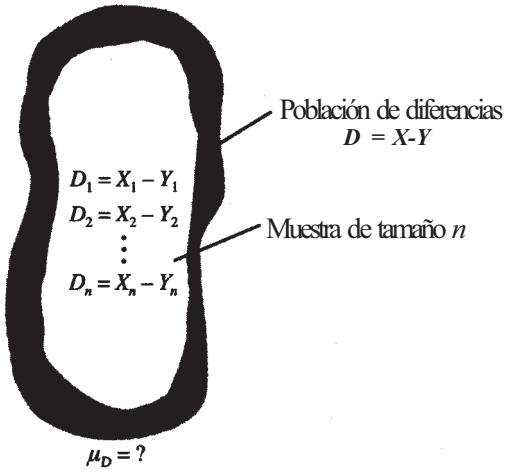


Figura 9.8. Datos emparejados generadores de una población de diferencias.

Podemos tener un 90 % de confianza en que la diferencia media de niveles de colesterol en plasma está entre 5.3 y 61.1 mg/dL. Es decir, podemos tener un 90% de confianza de que el nivel medio de colesterol se reducirá, como mínimo, en 5.3 mg/dL.

Contraste T para datos emparejados

Las medias pueden ser comparadas también utilizando la aproximación por contraste de hipótesis. La hipótesis nula $\mu_X = \mu_Y$ es equivalente a la hipótesis $\mu_D = 0$. El estadístico para contrastar esta hipótesis basado en la muestra de diferencias es:

$$\frac{\bar{D} - 0}{S_d/\sqrt{n}} \quad \text{(Contraste } T \text{ para datos emparejados)}$$

que sigue una distribución *T con n* - 1 grados de libertad, si H_0 es cierta. La utilización de este estadístico se explica en el ejemplo siguiente.

Ejemplo 9.5.3. Se lleva a cabo un estudio sobre la aparición de los dientes en los aborígenes australianos. El propósito es detectar diferencias, si existen, en el tiempo de aparición de los dientes permanentes de los lados izquierdo y derecho. Uno de los dientes estudiados es el

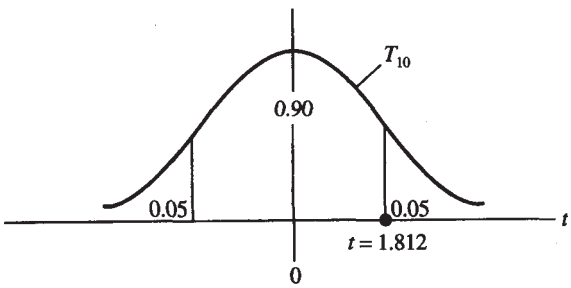


Figura 9.9. Partición de la curva T_{10} necesaria para obtener un intervalo de confianza de μ_D del 90%.

incisivo. Todos los individuos son varones. Se determinan la edad de cada uno en el momento de aparición del incisivo izquierdo y la edad en el momento de aparición del incisivo derecho. De este modo, cada individuo produce un par de observaciones. Los resultados estadísticas del estudio son los que se indican a continuación, donde el orden de sustracción es: edad del lado izquierdo menos edad del lado derecho:

$$n = 17 \qquad \bar{d} = 1.5 \text{ años} \qquad s_d = 4.7$$

El valor observado del estadístico es:

$$\frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{1.5}{4.7/\sqrt{17}} = 1.31$$

Obsérvese que $P[T_{16} \geq 1.31] > 0.10$. Puesto que no se indica ninguna preferencia direccional, el contraste es con dos colas. El valor P para el contraste con dos colas es superior a 0.20. Se concluye que no hay evidencia para pretender que hay diferencia en el tiempo medio de aparición de los incisivos izquierdo y derecho en los aborígenes varones de Australia.

En la utilización de estos procedimientos, se ha supuesto que la variable $D = X - Y$ está, al menos aproximadamente, normalmente distribuida.

EJERCICIOS 9.5

1. Se estudió también el efecto del ejercicio físico sobre el nivel de triglicéridos utilizando los 11 individuos del Ejemplo 9.5.1, obteniéndose las siguientes lecturas (en miligramos de triglicéridos por 100 mililitros de sangre) previas y posteriores al ejercicio:

Sujeto	Previo	Posterior
1	68	95
2	77	90
3	94	86
4	73	58
5	37	47
6	131	121
7	77	136
8	24	65
9	99	131
10	629	630
11	116	104

Hallar un intervalo de confianza del 90 % para el cambio medio en el nivel de triglicéridos. ¿Hay pruebas de que exista alguna diferencia? Si es así, ¿cuál es la dirección del cambio?

2. Se realizó un estudio para comparar el contenido de sodio en el plasma de las focas *peleteras* australes jóvenes, con el nivel de sodio en la leche de las focas. Se obtuvieron

las siguientes observaciones sobre el contenido de sodio [en milimoles por litro de leche (o plasma)] en 10 focas aleatoriamente seleccionadas:

Sujeto	Leche	Plasma
1	**93**	**147**
2	104	157
3	**95**	**142**
4	81.5	**141**
5	**95**	142
6	**95**	147
7	76.5	148
8	80.5	**144**
9	79.5	144
10	87.0	146

Hallar un intervalo de confianza del 95 % de la diferencia media de niveles de sodio en los dos líquidos corporales. ¿Hay pruebas de que exista alguna diferencia? Si es así, ¿cuál es la dirección de la diferencia?

3. Se realiza un estudio para determinar el efecto de un medidor casero para ayudar a los diabéticos a controlar sus niveles de glucosa en sangre (glucemia). En el estudio participa una muestra aleatoria de 36 diabéticos. Se obtuvieron los niveles de glucemia para cada paciente antes de que se les enseñara a utilizar el medidor, y después de que hubieron utilizado el medidor durante varias semanas. Se registró una diferencia muestral media de 2.78 mmol/litro, con una desviación típica muestral de 6.05 mmol/litro (la sustracción viene dada en el orden «antes» menos «después»). ¿Hay prueba suficiente para pretender que el medidor es efectivo para ayudar a los pacientes a reducir sus niveles de glucosa? Dar la respuesta a partir del valor *P*.
4. Se pensó que un programa de ejercicios regulares, moderadamente activos, podría beneficiar a los pacientes que habían sufrido previamente un infarto de miocardio. Once individuos participaron en un estudio para comprobar este argumento. Antes de que empezara el programa, se determinó la capacidad de trabajo de cada persona midiendo el tiempo que tardó en alcanzar una frecuencia de 160 latidos por minuto mientras caminaba sobre una cinta sin fin. Después de 25 semanas de ejercicio controlado, se repitieron las medidas en la cinta sin fin y se registró la diferencia en tiempo para cada sujeto. Resultaron los siguientes datos:

Sujeto	Antes	Después
1	**7.6**	14.7
2	**9.9**	14.1
3	**8.6**	11.8
4	**9.5**	16.1
5	**8.4**	**14.7**
6	**9.2**	14.1
7	**6.4**	13.2
8	**9.9**	14.9
9	**8.7**	12.2
10	10.3	13.4
11	**8.3**	14.0

¿Sostienen estos datos el argumento de los investigadores? Dar la respuesta tomando como base el valor *P*.

5. Los datos de temperatura recogidos en 1000 estaciones meteorológicas terrestres y marítimas de todo el mundo dieron una temperatura media de 57 °F en 1950. En 1988, la temperatura media en estas estaciones fue de 57.6 °F. Emparejando las lecturas de 1988 y 1950 por estación, se estima que la desviación típica de la diferencia de las lecturas es $s_d = 4.1$ °F. ¿Sostienen estos datos el argumento de que la temperatura media en 1988 fue superior que en 1950? Explicarlo a partir del valor *P* del contraste apropiado. (Basado en las temperaturas descritas en U.S. Senator Timothy Wirth, «Conservation Is the Key», *Journal of Environmental Health,* primavera de 1990, pág. 25.)

HERRAMIENTAS COMPUTACIONALES

TI83

La calculadora TI83 realizará el contraste de *F* descrito en la Sección 9.2. También hallará intervalos de confianza y contrastes de hipótesis sobre la relación entre μ_1 y μ_2 usando el contraste de *T,* tanto el de variación conjunta como el de Smith-Satterthwaite. En el cálculo de intervalos de confianza, al ejecutar los contrastes de *T* se le preguntará si quiere o no promediar las varianzas. Puede usar tanto datos introducidos como resúmenes estadísticos en cada uno de los casos.

XX. Contraste de *F* para comparar varianzas

Para realizar el contraste de *F* utilizaremos los datos del Ejemplo 9.2.4.

Tecla/Comando de la TI83	Propósito
1. STAT ◁ ALPHA D	1. Accede a la pantalla usada para realizar el contraste de la *F*.
2. cursor a STATS ENTER	2. Indica que se usará un resumen estadístico, en lugar de introducir los datos sin procesar, para realizar el contraste.
3. ▽ 0.102	3. Introduce 0.102 como valor para s_1
4. ▽ 41	4. Introduce 41 como el valor de n_1
5. ▽ 0.068	5. Introduce 0.068 como valor para s_2.
6. ▽ 29	6. Introduce 29 como el valor de n_2.
7. ▽ ENTER	7. Indica que el contraste es de dos colas.
8. ▽ ENTER	8. Calcula y muestra el valor del estadístico *F* (2.25); muestra el valor *P* para el contraste con dos colas (0.0271927299).

XXI. Intervalo de confianza para $\mu_1 - \mu_2$

Este procedimiento se realizará usando los datos del Ejemplo 9.3.3. En este ejemplo, promediaremos.

Tecla/Comando de la TI83	Propósito
1. STAT ◁ 0	1. Accede a la pantalla necesaria para formular el intervalo de confianza para $\mu_1 - \mu_2$.
2. cursor a STATS ENTER	2. Indica que se usará un resumen estadístico en lugar de introducir los datos sin procesar, para realizar el test.
3. ▽ 205	3. Introduce 205 como valor para $\bar{x}_1$.
4. ▽ 100	4. Introduce 100 como el valor de s_1.
5. ▽ 25	5. Introduce 25 como valor para n_1.
6. ▽ 135	6. Introduce 135 como el valor de $\bar{x}_2$.
7. ▽ 95	7. Introduce 95 como el valor de s_2.
8. ▽ 11	8. Introduce 11 como el valor de n_2.
9. ▽ 0.90	9. Introduce 0.90 como el nivel de confianza deseado.
10. ▽ ▷ ENTER	10. Indica que las varianzas han de promediarse.
11. ▽ ENTER	11. Calcula y muestra el intervalo de confianza al 90 % para $\mu_1 - \mu_2$, da el valor de s .

XXII. Contraste de la *T* para dos muestras

Los datos del Ejemplo 9.4.1 se usan para realizar el contraste de la *T* para dos muestras. En este ejemplo, no es apropiado promediar.

Tecla/Comando de la TI83	Propósito
1. STAT ◁ 4	1. Accede a la pantalla necesaria para realizar el contraste de la *T* para dos muestras.
2. cursor a STATS ENTER	2. Indica que se usará un resumen estadístico, en lugar de introducir los datos sin procesar, para realizar el test.

3.	▽ 0.0167	3. Introduce 0.0167 como valor para $\bar{x}_1$.
4.	▽ 0.0042	4. Introduce 0.0042 como el valor de s_1.
5.	▽ 57	5. Introduce 57 como valor para n_1
6.	▽ 0.0144	6. Introduce 0.0144 como el valor de $\bar{x}_2$.
7.	▽ 0.0024	7. Introduce 0.0024 como el valor de s_2.
8.	▽ 12	8. Introduce 12 como el valor de n_2.
9.	▽ ▷ ▷ ENTER	9. Indica que el contraste es con cola a la derecha.
10.	▽ ENTER	10. Indica que las varianzas no han de promediarse.
11.	▽ ENTER	11. Calcula los grados de libertad (27.51045397); calcula el valor observado para el estadístico de contraste (2.588564596); muestra el valor *P* para el contraste con cola a la derecha (0.0076098754).

XXIII. Contraste *T* para datos emparejados

La calculadora TI83 puede hallar intervalos de confianza para la diferencia de medias y contraste de hipótesis de μ_D. La técnica se ilustra con los datos del Ejemplo 9.5.2.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos STAT.
2. 182 ENTER 232 ENTER ⋮ 262 ENTER	2. Introduce las observaciones de *x* (previos) en la columna L,.
3. ▷ 198 ENTER 210 ENTER ⋮ 226 ENTER	3. Introduce las observaciones de y (posteriores) en la columna L_2.

4. STAT 1 ▷ △ 2ND L1 - (restar) 2ND L2 ENTER	4. Define L_3 como L_1 - L_2 para formar una columna con los valores de la diferencia.
5. STAT ◁ 8	5. Accede a la pantalla necesaria para encontrar el intervalo de confianza sobre μ_D.
6. cursor a DATA ENTER	6. Indica que se utilizarán los datos sin procesar, para realizar el contraste.
7. ▽ 2ND L_3	7. Indica que el intervalo de confianza se tiene que hallar usando los datos de L_3.
8. ▽ ▽ ENTER	8. Indica que el nivel de confianza es 0.90.
9. ENTER	9. Halla y muestra el intervalo de confianza al 90% para μ_D.

Paquete estadístico SAS

VIII. Contraste *T* para dos muestras

Una ventaja de efectuar el contraste Tpara dos muestras con el ordenador es que se calculan y registran los valores *P* de todos los contrastes pertinentes. El trabajo del investigador es interpretar correctamente el *output.* Ilustraremos la idea considerando los datos del Ejercicio 2, Sección 9.3. El código SAS necesario para analizar los datos es el siguiente:

Código SAS	**Propósito**
OPTIONS LS=80 PS=60 NODATE;	Establece las especificaciones de impresión.
DATA SPIDER;	Nombra el conjunto de datos.
INPUT GROUP SIZE;	Nombra las variables
LINES;	Indica que los datos vienen a continuación.
1 12.9 1 10.2 ⋮ 1 11.3 2 10.2 2 **6.9** ⋮ **2 8.8**	Líneas de datos; el grupo 1 es la *Dinopis* adulta y el grupo 2 es la *Menneus* adulta.
;	Señala el final de los datos.

PROC TTEST; CLASS GROUP;	Pide que un contreste Tpara dos muestras se ejecute; nombra la variable GROUP, que identifica a las dos poblaciones bajó estudio.
TITLE 'Contraste de igualdad de medias y varianzas'	Titula el *output.*

Obsérvese que en el Ejercicio 2 de la Sección 9.3 comparamos el tamaño medio de jas presas de *Dinopis* adulta con el de *Menneus* adulta. Para decidir si promediar o no las varianzas, se puede realizar un contraste *F* preliminar o usar la regla práctica. El valor de s_1^2/s_2^2 siendo s_1^2 la mayor de las varianzas, se muestra en ①. Como este valor es menor que 2, es adecuado promediar. Si las varianzas han sido comparadas mediante el contraste de la *F,* entonces el valor *P* del contraste con dos colas de H_0: $\sigma_1^2 = \sigma_2^2$ viene dado por ②. Puesto que el valor *P* es grande, tenemos poca evidencia de que $\sigma_1^2 \neq \sigma_2^2$. El contraste de la *F* también indica que promediar es apropiado. Así pues, al comparar las medias se utiliza el procedimiento conjunto. El valor observado del estadístico *T* de varianza conjunta, sus grados de libertad y el valor *P* del contraste con dos colas de H_0: $\mu_1 = \mu_2$ vienen dados p o③, ④, ⑤, respectivamente. Si se realizara un contraste con una cola, el investigador debería darse cuenta de este hecho y dividir el valor *P* por 2, para obtener el valor *P* correcto para el contraste. El estadístico *Tde* Smith-Satterthwaite, sus grados de libertad y el valor *P* del contraste con dos colas para igualdad de las medias, vienen dados por ⑥, ⑦ y ⑧, respectivamente. Sería apropiado usar estos valores para comparar medias si la regla práctica o el contraste de la *F* nos hubiesen mostrado que las varianzas poblacionales son distintas

```
                    TESTING FOR EQUALITY OF MEANS AND VARIANCES
                                TTEST PROCEDURE

VARIABLE: SIZE

GROUP   N          MEAN        STD DEV    STD ERROR      MÍNIMUM        MÁXIMUM

    1  10   10.26000000    2.51360741  0.79487246  7.00000000    14.40000000
    2  10    9.02000000    1.89666374  0.59977774  5.30000000    11.00000000
VARIANCES    T         DF     PROB > |T|
                ⑥          ⑦         ⑧
UNEQUAL     1.2453     16.7     0.2302
EQUAL       1.2453     18.0     0.2290
                ③          ④         ⑤

FOR HO: VARIANCES ARE EQUAL, F' =   1.76 WITH 9 AND 9 DF      PROB > F'= 0.4142
                                     ①                            ②
```

IX. Contraste *T* para datos emparejados

PROC MEANS es un procedimiento de SAS que puede utilizarse para ejecutar un contraste *T* para datos emparejados. Hacemos que SAS forme los valores de las diferencias y entonces utilizamos PROC MEANS para analizar esas diferencias. Los datos del Ejercicio 4 de la Sección 9.5 se usan para ilustrar el procedimiento.

Código SAS	**Propósito**
OPTIONS LS=80 PS=60 NODATE;	Especificar las opciones de impresión.
DATA EXERCISE;	Nombra el conjunto de datos.
INPUT PRETEST POSTTEST;	Nombra las variables.

```
D=POSTTEST - PRETEST;
LINES;

7.6     14.7
9.9     14.1
:
8.3     14.0
;
PROC MEANS MEAN
STD T PRT;

TITLE 'Contraste T para datos
emparejados';
```

Forma los valores de las diferencias.
Indica que los datos vienen a continuación.

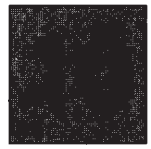
Líneas de datos.

Señala el final de los datos.
Hace que $\bar{d}$ y s_d sean calculadas; hace que se calcule el estadístico *T;* halla el valor *P* para el contraste H_0: $\mu_D = 0$ frente a H_1: $\mu_D \neq 0$.
Titula el *output.*

La salida de este programa se muestra más abajo. Obsérvese que $\bar{d}$ se muestra en ①. El valor de s_d se muestra en ②. El valor observado para el estadístico de contraste se da en ③, y el valor *P* del contraste con dos colas se muestra en ④. Como el contraste que nosotros queremos es con cola a la derecha, el valor *P* para el contraste se halla dividiendo el valor mostrado en ④ entre 2. Es decir, $P = 0.0001/2 = 0.00005$.

PAIRED T TEST

Variable	Mean	Std Dev	T	Prob > |T|
PRETEST	8.8000000	1.1392980	25.6177904	0.0001
POSTTEST	13.9272727	1.2345776	37.4148522	0.0001
D	① 5.1272727	② 1.4819520	③ 11.4748922	④ 0.0001



Procesos para k-muestras: introducción al diseño

En los Capítulos 6 y 7, se describieron métodos para hacer inferencias sobre la media y la varianza de una única población. El Capítulo 9 aborda métodos para comparar medias y varianzas de dos poblaciones. Ahora ampliamos los métodos del Capítulo 9 a más de dos poblaciones. También introducimos algunos de los aspectos elementales del diseño experimental y del *análisis de la varianza* (ANOVA). El término *diseño experimental* se refiere a un vasto campo de la estadística aplicada, relacionado con el estudio de métodos para recoger y analizar datos, cuya intención es aumentar al máximo la cantidad y mejorar la exactitud de la información proporcionada por un determinado experimento. El término *análisis de la varianza* hace referencia a un procedimiento analítico por el que se subdivide la variación total en la magnitud de una determinada respuesta en componentes que pueden atribuirse a algún origen reconocible y utilizarse para contrastar hipótesis de interés.

10.1. CLASIFICACIÓN SIMPLE O DE UNA VÍA, DISEÑO COMPLETAMENTE ALEATORIO CON EFECTOS FIJOS

El primer diseño que presentamos es el de *clasificación simple o de una vía, diseño completamente aleatorio con efectos fijos*. Mientras lo describimos, introduciremos algunos conceptos generales, así como la terminología subyacente al diseño experimental y al análisis de la varianza. Empezamos considerando dos estudios que pueden ser analizados por la técnica del análisis de la varianza.

Ejemplo 10.1.1. Se realiza un estudio para comparar la eficacia de tres programas terapéuticos para el tratamiento del acné de tipo medio a moderado. Se emplean tres métodos:

I. Este método, el más antiguo, supone el lavado, dos veces al día, con un cepillo de polietileno y un jabón abrasivo, junto con el uso diario de 250 mg de tetraciclina.

II. Este método, el utilizado actualmente, consiste en la aplicación de crema de tretinoína, evitar el sol, lavado dos veces al día con un jabón emulsionante y agua, y utilización, dos veces al día, de 250 mg de tetraciclina.

III. Éste es un nuevo método que consiste en evitar el agua, lavado dos veces al día con un limpiador sin lípidos, y uso de crema de tretinoína y de peróxido de benzoilo.

Se comparan estos tres tratamientos en cuanto a su eficacia en la reducción del número de lesiones de acné en los pacientes. En el estudio participaban treinta y cinco pacientes. Se les separó aleatoriamente en tres subgrupos de tamaños 10, 12 y 13. A uno de los grupos se le asignó el tratamiento I; a otro el tratamiento II; y al tercero, el tratamiento III. Después de 16 semanas, se anotó para cada paciente el porcentaje de mejoría en el número de lesiones.

Ejemplo 10.1.2. Uno de los focos de contaminación del agua lo constituyen los vertidos industriales y agrícolas ricos en fósforo. Demasiado fósforo puede causar una explosión en el crecimiento de plantas y microorganismos, a lo que se denomina *afloramiento.* Se determina el nivel de fósforo en los cuatro lagos principales de una determinada región, por extracción y análisis de muestras de agua. Se piensa que uno de los lagos se está viendo excesivamente contaminado por los vertidos de una planta industrial próxima y se espera que, comparando el nivel de fósforo de este lago con el de los otros, esto se demuestre.

Los Ejemplos 10.1.1 y 10.1.2 reflejan dos situaciones experimentales en las que *la clasificación simple, diseño completamente aleatorio con efectos fijos* aparece con frecuencia. En general, se describen del modo siguiente:

1. Tenemos una colección de N unidades experimentales y queremos estudiar el efecto de k tratamientos diferentes. Estas unidades son divididas aleatoriamente en k grupos de tamaños $n_1, n_2, ..., n_k$, y cada subgrupo recibe un tratamiento diferente. Se anota la respuesta. A los k subgrupos se les considera muestras aleatorias independientes de tamaños $n_1, n_2, ..., n_k$ extraídas de poblaciones con respuestas medias $\mu_1, \mu_2, ..., \mu_k$, respectivamente. Queremos contrastar la hipótesis nula de que los tratamientos tienen el mismo efecto medio:

$$\mathbf{H_0}: \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{(no hay diferencia en las medias de los } k \text{ tratamientos)}$$

$$\mathbf{H_1}: \mu_i \neq \mu_j \text{ para algún } i \text{ y } j \qquad \text{(al menos una media difiere de las otras)}$$

2. Tenemos k poblaciones, cada una identificable por alguna característica común que será estudiada en el experimento. Se seleccionan, de cada una de las k poblaciones, muestras aleatorias independientes de tamaños $n_1, n_2, ..., n_k$, respectivamente. Cada muestra recibe el mismo tratamiento, y cualquier diferencia observada en las respuestas medidas se atribuye a diferencias básicas entre las k poblaciones. La hipótesis es:

$$\mathbf{H_0}: \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{(no hay diferencia en las medias poblacionales)}$$

$$\mathbf{H_1}: \mu_i \neq \mu_j \text{ para algún i y } j \qquad \text{(al menos una media difiere de las otras)}$$

donde μ_i denota la respuesta media de la i-ésima población.

Si bien estas situaciones son experimentalmente algo diferentes, se asemejan en que cada una produce k muestras con medias $\mu_1, \mu_2, ..., \mu_k$, respectivamente. El propósito del experimento en cada caso es comparar medias poblacionales. Este diseño representa la extensión natural del problema de dos muestras no emparejadas del Capítulo 9 a más de dos muestras. La situación puede visualizarse como se refleja en la Figura 10.1.

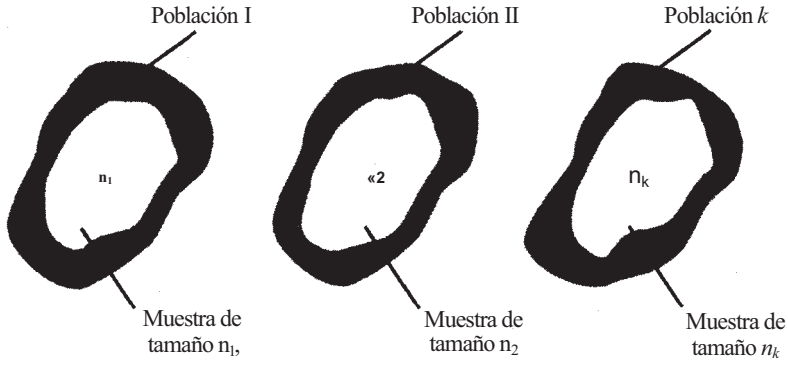


Figura H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ (¿Son iguales las medias poblacionales?)

Ejemplo 10.1.3. El Ejemplo 10.1.1 responde a la primera descripción general. Se dispone de una colección de $N = 35$ pacientes que sufren acné de tipo medio a moderado. Hay $k = 3$ tratamientos a comparar. Los pacientes se dividen aleatoriamente en tres subgrupos de tamaños n, = 10, $n_2 = 12$ y $n_3 = 13$. Cada subgrupo recibe un tratamiento diferente y la respuesta anotada es el porcentaje de mejoría en el número de lesiones observado al final de 16 semanas de tratamiento. Los tres subgrupos se consideran muestras aleatorias independientes extraídas de las poblaciones de pacientes que están recibiendo los tratamientos I, II y III, respectivamente. Basándose en los datos obtenidos, queremos contrastar:

H_0: $\mu_1 = \mu_2 = \mu_3$ (no hay diferencia en la respuesta media entre los tres tratamientos)

H_1: $\mu_i \neq \mu_j$ para algún i y j (al menos un tratamiento es diferente de los otros)

Ejemplo 10.1.4. El Ejemplo 10.1.2 satisface la segunda descripción general. Estamos estudiando k = 4 lagos. Cada lago constituye una población. De cada lago se seleccionarán muestras independientes. Cada muestra recibe el mismo tratamiento en cuanto a que se analiza el contenido en fósforo de cada uno de ellos por el mismo método. Cualquier diferencia en los niveles medios de fósforo que aparezcan se atribuye al hecho de que las muestras se extrajeron de diferentes lagos con distinta composición de sus aguas.

Cada ejemplo es una *clasificación simple, de una vía, diseño completamente aleatorio con efectos fijos.* El término *clasificación simple o de una vía* se refiere a que solamente se estudia un factor en cada experimento. El experimento implica k niveles de dicho factor. En el Ejemplo 10.1.1, el factor que interesa es el tipo de tratamiento recibido. Ningún otro factor, tal como la edad, tipo de piel, hábitos dietéticos o sexo del paciente, se tiene en cuenta. Están estudiándose tres tratamientos; de este modo, el factor está investigándose a tres niveles. En el Ejemplo 10.1.2, el único factor es el lago implicado. Ningún otro, tal como temperatura, estación o profundidad del lago, tienen interés en el estudio. Puesto que están implicados cuatro lagos, se consideran cuatro niveles. El término *diseño completamente aleatorio* se refiere a que no se ha realizado ningún intento de emparejar unidades experimentales de las distintas muestras. Las k muestras son independientes unas de otras. El término *efectos fijos* expresa que el experimentador selecciona específicamente los niveles del factor implicados, porque considera que éstos tienen un interés especial. No se seleccionan aleatoriamente de un grupo más grande de niveles posibles. En el Ejemplo 10.1.1, el propósito del experimento es comparar los tres tratamientos específicos. Los tratamientos no han sido seleccionados aleatoriamente de entre un gran grupo de tratamientos disponibles contra el acné. En el Ejem-

plo 10.1.2, los cuatro lagos seleccionados para el estudio se eligen específicamente por ser los lagos principales de la región. No han sido seleccionados aleatoriamente de entile todos los lagos de la región.

Formato de los datos y notación

Los datos recogidos en un experimento de un único factor se registran convenientemente en el siguiente formato:

Disposición de datos para la clasificación simple

Nivel de factor				
1	2	3	⋯	*k*
X_{11}	X_{21}	X_{31}	⋯	X_{k1}
X_{12}	X_{22}	X_{32}	⋯	X_{k2}
X_{13}	X_{23}	X_{33}	⋯	X_{k3}
⋮	⋮	⋮	⋮	⋮
X_{1n_1}	X_{2n_2}	X_{3n_3}	⋯	X_{kn_k}

Obsérvese que n_i es el tamaño de la muestra extraída de la i-ésima población y que $N = \sum_{i=1}^{k} n_i$ designa el número total de respuestas. Además, X_{ij}, $i = 1, 2, ..., k$, $j = 1, 2, ... n_i$, es una variable aleatoria que indica la respuesta de laj-ésima unidad experimental al i-ésimo nivel del factor. Al utilizar datos muestrales para comparar medias poblacionales, se requieren ciertos estadísticos:

$$T_{i\cdot} = \sum_{j=1}^{n_i} X_{ij} = \text{suma total de las respuestas en el nivel i-ésimo, } i = 1, 2, ..., k$$

$$\bar{X}_{i\cdot} = \frac{T_{i\cdot}}{n_i} = \text{media muestral para el nivel i-ésimo, i} = 1, 2, ..., k$$

$$T_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \sum_{j=1}^{k} T_{i\cdot} = \text{suma total de las respuestas}$$

$$\bar{X}_{\cdot\cdot} = \frac{T_{\cdot\cdot}}{N} = \text{media muestral de todas las respuestas}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 = \text{suma de los cuadrados de cada respuesta}$$

Obsérvese en esta notación que el punto indica el subíndice sobre el que se aplica el sumatorio. Por ejemplo, $T_{1\cdot}$ es la suma de todas las observaciones del nivel 1. Es decir,

$$T_{1\cdot} = X_{11} + X_{12} + X_{13} + \cdots + X_{1n_1}$$

El primer subíndice es siempre 1; el segundo subíndice varía desde 1 hasta n_1 El punto en $T_{1\cdot}$ indica que el segundo subíndice es el que varía. En el Ejemplo 10.1.5, se describe el cálculo de estos estadísticos.

Ejemplo 10.1.5. Cuando se realizó el experimento del Ejemplo 10.1.1, se obtuvieron los siguientes datos. (Recuérdese que la respuesta observada es el porcentaje de mejoría registrado por paciente, en el número de lesiones de acné al final de las 16 semanas de tratamiento.)

Nivel del factor (tratamiento recibido)

I		II		III	
48.6	50.8	68.0	71.9	67.5	61.4
49.4	47.1	67.0	71.5	62.5	67.4
50.1	52.5	70.1	69.9	64.2	65.4
49.8	49.0	64.5	68.9	62.5	63.2
50.6	46.7	68.0	67.8	63.9	61.2
		68.3	68.9	64.8	60.5
				62.3	

Obsérvese que $n_1 = 10$, $n_2 = 12$, $n_3 = 13$ y $N = \sum_{i=1}^{3} n_i = 10 + 12 + 13 = 35$. La observación 70.1 corresponde a la respuesta de la tercera unidad experimental del nivel 2 del factor y, por lo tanto, es el valor observado de la variable aleatoria X_{23}. Los valores observados de los otros estadísticos son:

$$T_{1\cdot} = \sum_{j=1}^{10} X_{1j} = \text{suma de las respuestas al tratamiento I}$$

$$= 48.6 + 49.4 + 50.1 + \cdots + 46.7 = 494.6$$

$$T_{2\cdot} = \sum_{j=1}^{12} X_{2j} = \text{suma de las respuestas al tratamiento II}$$

$$= 68.0 + 67.0 + 70.1 + \cdots + 68.9 = 824.8$$

$$T_{3\cdot} = \sum_{j=1}^{13} X_{3j} = \text{suma de las respuestas al tratamiento III}$$

$$= 67.5 + 62.5 + 64.2 + \cdots + 60.5 = 826.8$$

$$\bar{X}_{1\cdot} = \frac{T_{1\cdot}}{n_1} = \text{media muestral de las respuestas al tratamiento I}$$

$$= \frac{494.6}{10} = 49.46$$

$$\bar{X}_{2\cdot} = \frac{T_{2\cdot}}{n_2} = \text{media muestral de las respuestas al tratamiento II}$$

$$= \frac{824.8}{12} = 68.73$$

$$\bar{X}_{3\cdot} = \frac{T_{3\cdot}}{n_3} = \text{media muestral de las respuestas al tratamiento IIII}$$

$$= \frac{826.8}{13} = 63.60$$

$$T_{..} = \sum_{i=1}^{3} T_{i\cdot} = T_{1\cdot} + T_{2\cdot} + T_{3\cdot} = \text{suma total de todas las respuestas}$$

$$= 494.6 + 824.8 + 826.8 = 2146.2$$

$$\bar{X}_{..} = \frac{T_{..}}{N} = \text{media muestral de todas las respuestas}$$

$$= \frac{2146.2}{35} = 61.32$$

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i} X_{ij}^2 = (48.6)^2 + (49.4)^2 + (50.1)^2 + \cdots + (60.5)^2$$

$$= 133\,868.94$$

Contraste de H_0: $\mu_1 = \mu_2 = \bullet\bullet\bullet = \mu_k$

Para ver cómo pueden utilizarse estos estadísticos para contrastar la hipótesis de que las medias poblacionales son iguales, podemos introducir un modelo para la *clasificación simple, diseño completamente aleatorio con efectos fijos.* Un modelo es una representación matemática de una respuesta típica expresada en componentes atribuibles a diversos orígenes idenlificables. Para escribir el modelo se utiliza la siguiente notación:

μ_i = promedio teórico o respuesta esperada
al i-ésimo nivel, i = 1, 2, ..., *k*
= media de la i-ésima población (constante desconocida).

μ = promedio teórico o respuesta esperada, ignorando
los niveles del factor (constante desconocida)
= media de la población que resulta de la combinación
de las *k* poblaciones en una.

Obsérvese que si los niveles del factor no tienen efecto sobre la respuesta, entonces las medias μ_1, μ_2,..., μ_k serán la misma y se igualarán a la media global μ; no será así si los niveles del factor afectan a la respuesta. Por lo tanto, la diferencia entre la media del i-ésimo nivel y la media global $\mu_i - \mu$ indica el efecto, si lo hay, del i-ésimo nivel del factor. Obsérvese también que, a pesar del hecho de que cada miembro de la i-ésima población recibe el mismo tratamiento, las respuestas obtenidas variarán algo a causa de influencias aleatorias. Es decir, dentro de cada población hay cierta variabilidad natural en torno a la media poblacional. Para una determinada respuesta, X_{ij} esta variabilidad viene dada por la diferencia $X_{ij} - \mu_j$. Esta diferencia se conoce como *error aleatorio.* Teniendo en cuenta todo lo expuesto, el modelo para la *clasificación simple, diseño completamente aleatorio con efectos fijos,* puede expresarse del siguiente modo:

Modelo

$$X_{ij} \equiv \mu + (\mu_i - \mu) + (X_{ij} - \mu_i) \qquad \begin{matrix} i = 1, 2, \ldots. k \\ j = 1, 2, \ldots, n_i \end{matrix}$$

Este modelo expresa matemáticamente la idea de que cada respuesta puede dividirse en tres componentes, del siguiente modo:

Respuesta de la j-ésima unidad experimental al i-ésimo tratamiento (X_{ij})	=	respuesta media global (μ)	+	desviación de la media global debida al hecho de que la unidad reciba el tratamiento i-ésimo ($\mu_i - \mu$)	+	desviación aleatoria de la i-ésima media poblacional debida a influencias aleatorias ($X_{ij} - \mu_i$)

La Figura 10.2 ilustra esto de manera gráfica.

Como en situaciones anteriores, para contrastar la hipótesis nula, debe obtenerse un estadístico. El estadístico deberá ser lógico, pero, lo que es más importante aún, su distribución de probabilidad ha de ser conocida bajo el supuesto de que la hipótesis nula es cierta y las medias de las *k* poblaciones son iguales. Para que esto ocurra, deben hacerse ciertas suposiciones acerca de las poblaciones de las que han sido extraídas las muestras. En particular, suponemos lo siguiente:

Supuestos del modelo

1. Las *k* muestras representan muestras aleatorias independientes extraídas de *k* poblaciones específicas con medias $\mu_1, \mu_2, ..., \mu_k$, donde $\mu_1, \mu_2, ..., \mu_k$ son constantes desconocidas.
2. Cada una de las *k* poblaciones es normal.
3. Cada una de las k poblaciones tiene la misma varianza, σ^2.

Obsérvese que estos supuestos son paralelos a los del Capítulo 9 relativos al proceso *T* conjunto para comparar dos medias.

El análisis de la varianza se ha definido como un procedimiento por medio del cual la variación total en algunas respuestas medidas se subdivide en componentes que pueden atri-

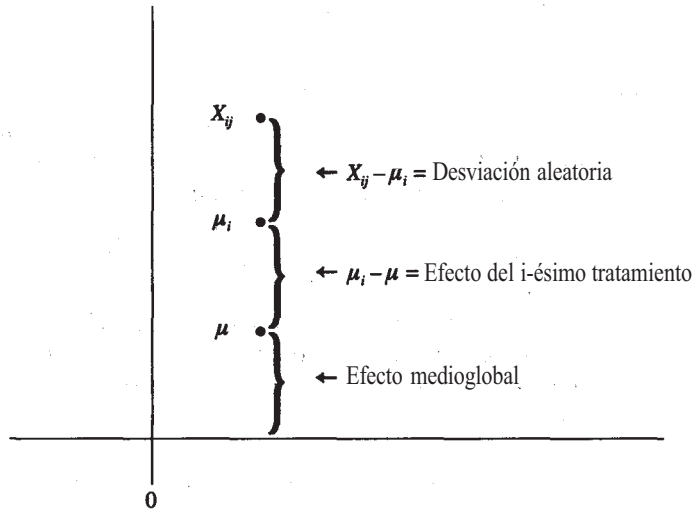


Figura 10.2. El valor de la ij-ésima observación puede dividirse en tres componentes: el efecto medio global μ, el efecto del i-ésimo tratamiento $\mu_i - \mu$, y la desviación aleatoria o no explicada $X_{ij} - \mu_i$.

buirse a orígenes reconocibles. Puesto que $\mu_1, \mu_2, ..., \mu_k$ son medias poblacionales teóricas, el modelo hace esto solamente desde el punto de vista teórico. En la práctica, para subdividir una observación, estas medias teóricas han de reemplazarse por sus estimadores $\bar{X}_{..}, \bar{X}_{1.}, \bar{X}_{2.}, ..., \bar{X}_{k.}$, respectivamente. Sustituyendo en el modelo las medias teóricas por sus estimadores, se obtiene la siguiente identidad:

$$X_{ij} \equiv \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.})$$

Obsérvese que $\bar{X}_{..}$ es un estimador para μ, media global de la población combinada; $\bar{X}_{i.} - \bar{X}_{..}$ es un estimador para $\mu_i - \mu$, el efecto del i-ésimo tratamiento; $\bar{X}_{ij.} - \bar{X}_{i.}$ s un estimador para $X_{ij} - \mu_i$, el error aleatorio. Al término $X_{ij} - \bar{X}_{i.}$ se le llama, usualmente, *residuo.* Esta identidad es equivalente a:

$$X_{ij} - \bar{X}_{..} \equiv (\bar{X}_{i.} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.})$$

Si se eleva al cuadrado cada miembro de la identidad y se suma después sobre todos los posibles valores de *i y j,* se obtiene la llamada *identidad de la suma de cuadrados para el diseño de clasificación simple.*

Identidad de la suma de cuadrados

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2 \equiv \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$$

Se obtiene por medio de una sencilla aplicación de las reglas de la suma. Obsérvese que en esta identidad hay tres componentes. Cada una tiene una interpretación práctica no desdeñable. En particular,

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2$ = suma de los cuadrados de las desviaciones de las observaciones respecto a la media global
= medida de la variabilidad total en los datos
= suma total de cuadrados = SS_{Total}

$\sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$ = suma ponderada de los cuadrados de las desviaciones del nivel o de las medias de los tratamientos respecto a la media global
= medida de la variabilidad en los datos atribuida al hecho de que se utilicen diferentes niveles o tratamientos
= suma de cuadrados de los tratamientos = SS_{Tr}

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$ = suma de los cuadrados de las desviaciones de las observaciones respecto a la media de los tratamientos asociada con la observación.
= medida de la variabilidad en los datos atribuida a las fluctuaciones aleatorias entre sujetos dentro del mismo nivel del factor
= suma de los cuadrados de los residuos, o error = SS_E

Utilizando esta notación taquigráfica, la identidad de la suma de cuadrados puede escribirse de la forma:

$$SS_{\text{Total}} = SS_{\text{Tr}} + SS_{\text{E}}$$

Para contrastar:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

necesitamos definir dos estadísticos que sean funciones de SS_{Tr} y SS_E. El primero, llamado *cuadrado medio de los tratamientos* MS_{Tr}, se halla dividiendo SS_{Tr} por $k - 1$; el segundo, llamado *cuadrado medio del error* MS_E, se halla dividiendo SS_E por $N - k$. Es decir, definimos:

$$MS_{Tr} = \frac{SS_{Tr}}{k-1} = \text{cuadrado medio de los tratamientos}$$

$$MS_E = \frac{SS_E}{N-k} = \text{cuadrado medio residual}$$

Puesto que cada uno de estos cuadrados medios es un estadístico, es también una variable aleatoria. Como tal, cada uno tiene una distribución de probabilidad y un valor medio o esperado. Estos valores esperados son particularmente importantes, porque proporcionan la base lógica para contrastar la igualdad de las medias por el procedimiento ANOVA. Si bien las demostraciones están más allá del ámbito de este texto, se puede probar que

$$E[MS_{Tr}] = \sigma^2 + \sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1}$$

y

$$E[MS_E] = \sigma^2$$

¿Cómo pueden utilizarse MS_{Tr} y MS_E para contrastar H_0? Para responder a esta pregunta, necesitamos sólo observar que si H_0 es cierta, entonces $\mu_1 = \mu_2 = \cdots = \mu_k = \mu$, y por tanto,

$$\sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1} = 0$$

Si H_0 no es cierta, este término será positivo. De ese modo, si H_0 es cierta, esperaríamos que MS_{Tr} y MS_E tuvieran valores próximos, ya que ambos estiman él mismo parámetro, es decir, σ^2; si H_0 no es cierta, esperaríamos que MS_{Tr} fuera algo mayor que MS_E, lo que sugiere el cociente:

$$\frac{MS_{Tr}}{MS_E}$$

como estadístico lógico. Si H_0 es cierta, su valor será próximo a 1; si no, deberá tomar un valor superior a 1. El cociente se puede utilizar como un estadístico ya que, si H_0 es cierta, se sabe que tiene una distribución F con $k - 1$ y $N - k$ grados de libertad. El contraste es siempre con cola a la derecha, rechazándose H_0 para los valores del estadístico

$$\boxed{F_{k-1,\,N-k} = \frac{MS_{Tr}}{MS_E}} \qquad \text{(Estadístico del contraste)}$$

que se estiman como *demasiado grandes* para deberse al azar.

Es difícil calcular SS_{Total}, SS_{Tr} y SS_E directamente a partir de las definiciones. Por ello se han desarrollado fórmulas de cálculo más fáciles de manejar y especialmente adecuadas para el empleo de calculadoras electrónicas. Las fórmulas son consecuencia directa de las reglas de la suma.

Fórmulas de cálculo

$$SS_{\text{Total}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{N}$$

$$SS_{\text{Tr}} = \sum_{i=1}^{k} \frac{T_{i\cdot}^2}{n_i} - \frac{T_{\cdot\cdot}^2}{N}$$

$$SS_E = SS_{\text{Total}} - SS_{\text{Tr}}$$

Todo lo que aquí se ha dicho puede recopilarse en una tabla. Teniendo en cuenta que k indica el número de tratamientos o de niveles del factor que se están investigando; n_i, el número de observaciones seleccionadas de la i-ésima población; $T_{i\cdot}$, la suma de las observaciones para el i-ésimo nivel; N, número total de observaciones, y T., la suma de todas las observaciones, él resumen de datos quedaría como en la Tabla 10.1. Obsérvese que las columnas «grados de libertad» y «suma de cuadrados» son aditivas. Es decir, los grados de libertad para los tratamientos y para el error suman N - 1, número total de grados de libertad. Además, $SS_{Tr} + SS_E = SS_{Total}$.

El procedimiento ANOVA se describe en el Ejemplo 10.1.6.

Ejemplo 10.1.6. En el Ejemplo 10.1.5, comenzamos los cálculos necesarios para contrastar la hipótesis nula de que los tres tratamientos tienen el mismo efecto medio. En particular, se obtuvieron los siguientes valores:

$T_{1\cdot} = 494.6$ $\quad \sum_{i=1}^{3}\sum_{j=1}^{n_i} X_{ij}^2 = 133\,868.94$ $\quad n_1 = 10$

$T_{2\cdot} = 824.8$ $\quad n_2 = 12$

$T_{3\cdot} = 826.8$ $\quad n_3 = 13$

$T_{\cdot\cdot} = 2146.2$ $\quad N = 35$

Tabla 10.1. Análisis de la varianza: clasificación simple, diseño completamente aleatorio con efectos fijos

Origen de la variación	Grados de libertad *DF*	Suma de cuadrados *SS*	Cuadrado medio *MS*	Cuadrado medio esperado	Cociente *F*
Tratamiento o nivel	$k - 1$	$\sum_{i=1}^{k} \frac{T_{i\cdot}^2}{n_i} - \frac{T_{\cdot\cdot}^2}{N}$ (SS_{Tr})	$\frac{SS_{\text{Tr}}}{k-1}$	$\sigma^2 + \sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1}$	$F_{k-1,\,N-k} = \frac{MS_{\text{Tr}}}{MS_E}$
Residuo o error	$N - k$	$SS_{\text{Total}} - SS_{\text{Tr}}$ (SS_E)	$\frac{SS_E}{N-k}$	σ^2	
Total	$N - 1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{N}$ (SS_{Total})			

A partir de estos estadísticos básicos, podemos evaluar SS_{Total}, SS_{Tr}, y SS_E:

$$SS_{\text{Total}} = \sum_{i=1}^{3}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{N}$$

$$= 133\,868.94 - \frac{(2146.2)^2}{35} = 2263.96$$

$$SS_{\text{Tr}} = \sum_{i=1}^{3} \frac{T_{i\cdot}^2}{n_i} - \frac{T_{\cdot\cdot}^2}{N}$$

$$= \frac{(494.6)^2}{10} + \frac{(824.8)^2}{12} + \frac{(826.8)^2}{13} - \frac{(2146.2)^2}{35} = 2133.66$$

$$SS_E = SS_{\text{Total}} - SS_{\text{Tr}}$$

$$= 2263.96 - 2133.66 = 130.30$$

Los cuadrados medios correspondientes vienen dados por

$$MS_{\text{Tr}} = \frac{SS_{\text{Tr}}}{k-1} = \frac{SS_{\text{Tr}}}{2} = \frac{2133.66}{2} = 1066.83$$

$$MS_E = \frac{SS_E}{N-k} = \frac{SS_E}{32} = \frac{130.30}{32} = 4.07$$

El valor observado del estadístico es

$$F_{k-1,\,N-k} = F_{2,32} = \frac{MS_{\text{Tr}}}{MS_E}$$

$$= \frac{1066.83}{4.07} = 262.12$$

Recuérdese que H_0 debe ser rechazada si su valor es significativamente mayor que 1. ¡Éste parece ser ciertamente el caso! Para estar seguros, consultamos la tabla de *F*. En ella vemos que,

$$P = P[F_{2,32} > 262.12] < 0.01$$

Ya que este valor *P* es pequeño, tenemos evidencia estadística de que los tres tratamientos difieren en el efecto medio. Se suelen disponer los resultados en una tabla de análisis de la varianza (ANOVA). La tabla ANOVA para estos datos se muestra en la Tabla 10.2.

Conviene hacer algunos comentarios de naturaleza práctica. Suponemos que las poblaciones muestreadas están normalmente distribuidas y que las varianzas poblacionales son iguales. Existen diversos contrastes para comprobar estas suposiciones. En el Capítulo 13 presentamos el contraste de Lilliefors, un procedimiento gráfico para contrastar la normalidad, y el contraste de Bartlett, un procedimiento analítico utilizado para comparar varianzas. Debe señalarse igualmente que, en el modelo de clasificación simple, los tamaños de las muestras pueden no ser iguales. Hay, desde luego, ciertas ventajas cuando se dispone de muestras de igual tamaño. En primer lugar, las consecuencias de no cumplir el supuesto de varianzas iguales no son graves si las muestras son del mismo tamaño. Entonces, el empleo de un contraste de Bartlett o de otro para varianzas iguales, aunque sería preferible, no es esencial. En segundo lugar, si se rechaza H_0 y se declaran distintas las medias poblacionales, generalmente se necesitan nuevos contrastes. Muchos de éstos están concebidos pensando en muestras de igual tamaño.

Tabla 10.2. ANOVA para los datos del Ejemplo 10.1.6

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	***2***	2133.66	1066.83	262.12
Error	32	130.30	4.07	
Total	34	2263.96		

En la práctica, los estudios de una vía se hacen normalmente con la ayuda de un ordenador en vez de manualmente. En la sección de Herramientas Computacionales del final de este capítulo, presentaremos el análisis realizado por la TI83 y el programa SAS necesario para realizar estos estudios.

EJERCICIOS 10.1

1. Se sabe que el dióxido de carbono tiene un efecto crítico en el crecimiento microbiológico. Cantidades pequeñas de CO_2 estimulan el crecimiento de muchos microorganismos, mientras que altas concentraciones inhiben el crecimiento de la mayor parte de ellos. Este último efecto se utiliza comercialmente cuando se almacenan productos alimenticios perecederos. Se realizó un estudio para investigar el efecto de CO_2 sobre la tasa de crecimiento de *Pseudomonas fragi,* un corruptor de alimentos. Se administró dióxido de carbono a cinco presiones atmosféricas diferentes. La respuesta anotada fue el cambio porcentual en la masa celular después de un tiempo de crecimiento de una hora. Se utilizaron diez cultivos en cada nivel. Se obtuvieron los siguientes datos:

Nivel del factor (presión de CO_2 en atmósferas)

0.0	**0.083**	**0.29**	**0.50**	**0.86**
62.6	50.9	45.5	29.5	24.9
59.6	44.3	41.1	22.8	17.2
64.5	47.5	29.8	19.2	7.8
59.3	49.5	38.3	20.6	10.5
58.6	48.5	40.2	29.2	17.8
64.6	50.4	38.5	24.1	22.1
50.9	35.2	30.2	22.6	22.6
56.2	49.9	27.0	32.7	16.8
52.3	42.6	40.0	24.4	15.9
62.8	41.6	33.9	29.6	8.8

a) Se suponen efectos fijos: ¿qué implica esto respecto a los niveles atmosféricos elegidos?
b) Plantear la hipótesis nula a contrastar.
c) Hallar los valores de $T_{i\cdot}$, $\overline{X}_{i\cdot}$, $T_{\cdot\cdot}$, $\overline{X}_{\cdot\cdot}$ y $\sum_{i=1}^{5}\sum_{j=1}^{10} X_{ij}^2$.
d) Hallar SS_{Total}, SS_{Tr} y SS_E.
e) Hallar MS_{Tr} y MS_E.
f) Evaluar el estadístico F utilizado para contrastar H_0.
g) ¿Puede rechazarse H_0? Explicarlo basándose en el valor P del contraste.
h) ¿Qué suposiciones se están haciendo sobre las cinco poblaciones muestreadas?

2. El enrojecimiento por clorpropamida/alcohol (CAPF) es un enrojecimiento facial experimentado por los pacientes diabéticos tratados con clorpropamida después del consumo de alcohol. Se realiza un experimento para estudiar la capacidad de la indometacina para bloquear esta reacción. Participaron en el estudio tres grupos de diabéticos: I, diabéticos sin complicaciones; II, diabéticos con retinopatía grave, y III, diabéticos con enfermedades de los grandes vasos. Al principio del experimento, se toma la temperatura facial de cada paciente y se le suministran después 250 mg de clorpropamida. Después de doce horas, se le dan 40 mL de jerez y se anota la temperatura facial. Se repite el experimento con cada paciente que recibe 100 mg de indometacina 75 minutos antes de tomar el jerez. Nuevamente, se anota el cambio en la temperatura facial. Se obtuvieron las siguientes observaciones para *X,* diferencia de temperaturas (temperatura después de utilizarse la indometacina menos temperatura antes de que se tomara):

Nivel del factor (tipo de diabético)

I	II	III
-0.23	0.32	-0.35
-0.76	0.25	-0.13
-0.15	0.29	0.16
-0.34	0.07	0.12
-0.54	0.10	-0.43
-1.90	0.18	0.49
-2.07	0.16	-0.30
-1.21	0.23	0.44

a) Plantear la hipótesis nula a contrastar.
b) Contrastar la hipótesis nula completando la tabla ANOVA. Supóngase normalidad.
c) En la práctica, ¿cuál es el significado de los signos negativos asociados a algunas de las observaciones?

3. Se realiza un estudio de contenido de azufre en cinco de los principales yacimientos de carbón en Texas. Se toman muestras aleatoriamente de cada uno de los yacimientos y se analizan. Los datos del porcentaje de azufre por muestra se indican en la tabla de la parte inferior. Supuestas normalidad y varianzas iguales, contrastar la igualdad de medias. ¿Qué conclusiones pueden extraerse de estos datos?

Nivel del factor (yacimiento)

1	2	3	4	5
1.51	1.69	1.56	1.30	0.73
1.92	0.64	1.22	0.75	0.80
1.08	0.90	1.32	1.26	0.90
2.04	1.41	1.39	0.69	1.24
2.14	1.01	1.33	0.62	0.82
1.76	0.84	1.54	0.90	0.72
1.17	1.28	1.04	1.20	0.57
	1.59	2.25	0.32	1.18
		1.49		0.54
				1.30

4. Se realiza un estudio para determinar los efectos de una planta cloralcalina sobre los peces que viven en el río que fluye junto a la planta. La variable de interés es el nivel total de mercurio, en microgramos por gramo de peso corporal, por pez en el área. Las muestras de peces se toman en cuatro puntos a lo largo del río:

   I. 5.5 km por encima de la planta.
   II. 3.7 km por debajo de la planta.
   III. 21 km por debajo de la planta.
   IV. 133 km por debajo de la planta.

   Se obtuvieron los siguientes datos:

   Nivel del factor (localización en el río)

I	II	III	IV
0.45	1.64	1.56	0.65
0.35	1.67	1.55	0.59
0.32	1.85	1.69	0.69
0.68	1.57	1.67	0.62
0.53	1.59	1.60	0.70
0.34	1.61	1.68	0.64
0.61	1.53	1.65	0.81
0.41	1.40	1.59	0.58
0.51	1.70	1.75	0.53
0.71	1.48	1.49	0.75

   Supuesta la normalidad, contrastar la igualdad de medias.

5. Se realiza un estudio de diversas especies de pájaros que son de similar naturaleza y comparten un medio común. El canto de cada especie tiene un conjunto de rasgos distintivos que permite reconocerla. Una característica investigada es la duración del canto en segundos. Se estudian tres especies; los *towhee,* el cuelliamarillo común, y el malviz pardo. Se obtuvieron los siguientes datos:

   Nivel del factor (especies)

Towhee	Cuelliamarillo común	Malviz pardo
1.11	2.17	0.42
1.23	1.85	0.93
0.91	1.99	0.77
0.95	1.74	0.37
0.99	1.54	0.50
1.08	1.86	0.48
1.18	1.87	0.68
1.29	2.04	0.62
1.12	1.69	0.39
0.88		0.67
1.34		1.03
		0.79

Supuesta la normalidad y varianzas iguales, contrastar la hipótesis nula de que las duraciones medias de los cantos son las mismas para las tres especies.

6. Se lleva a cabo un estudio para investigar el crecimiento de los robles rojos americanos a tres alturas diferentes (975 metros, 825 metros y 675 metros). La variable es la medida en centímetros de la médula durante un período de 10 años. Se obtuvieron los siguientes resultados:

975 m	825 m	675 m	975 m	825 m	675 m
3.8	5.0	1.8	2.8	3.0	2.3
1.3	2.0	2.3	3.8	1.6	1.1
2.6	2.9	2.0	1.5	1.4	1.1
2.2	3.4	2.2	4.0	3.0	2.6
2.0	3.2	2.4	1.7	1.3	2.1

Basándose en estos datos, ¿puede deducirse que existen diferencias en el crecimiento medio en 10 años de estos árboles, entre las tres altitudes? Explíquese la respuesta a partir del valor *P* de su ANOVA. (Basado en un estudio realizado por Allison Field, Departamento de Biología, Universidad de Radford, 1996).

10.2. COMPARACIONES MÚLTIPLES Y POR PAREJAS

Una vez que se ha realizado un análisis de la varianza mediante una clasificación simple para comparar k medias poblacionales, nos encontraremos en una de las dos siguientes situaciones:

1. No hemos podido rechazar H_0. Basándonos en los datos disponibles, no nos ha sido posible detectar ninguna diferencia entre las k medias poblacionales. En este caso, el análisis de los datos es completo.
2. Hemos tenido razones para rechazar H_0 y, por tanto, llegamos a la conclusión de que hay algunas diferencias entre las k medias poblacionales. En este caso, el análisis de los datos sólo ha empezado, ya que es natural continuar la investigación para tratar de localizar con precisión dónde está la diferencia.

Hay varios métodos para detectar diferencias entre medias poblacionales una vez que ha sido rechazada la hipótesis de igualdad. Lentner y Bishop [9] presentan una muy buena reseña sobre la mayoría de ellos. Aquí presentamos dos posibilidades: los *contrastes T de Bonferroni y* el *contraste de Duncan de rango múltiple.*

Contraste *Tde* Bonferroni: comparaciones por parejas

Considérese un conjunto de k medias poblacionales. Pueden formarse $\binom{k}{2} = k(k-1)/2$ posibles pares de medias. Así, hay $k(k - 1)/2$ contrastes posibles de la forma

$$H_0: \mu_i = \mu_j$$
$$H_1: \mu_i \neq \mu_j$$

que pueden realizarse. Por ejemplo, si estamos comparando cinco medias, entonces $k = 5$ y hay $k(k - 1)/2 = 10$ contrastes posibles que pueden realizarse:

$$H_0: \mu_1 = \mu_2 \quad H_0: \mu_1 = \mu_3 \quad H_0: \mu_1 = \mu_4 \quad H_0: \mu_1 = \mu_5$$
$$H_1: \mu_1 \neq \mu_2 \quad H_1: \mu_1 \neq \mu_3 \quad H_1: \mu_1 \neq \mu_4 \quad H_1: \mu_1 \neq \mu_5$$
$$H_0: \mu_2 = \mu_3 \quad H_0: \mu_2 = \mu_4 \quad H_0: \mu_2 = \mu_5$$
$$H_1: \mu_2 \neq \mu_3 \quad H_1: \mu_2 \neq \mu_4 \quad H_1: \mu_2 \neq \mu_5$$
$$H_0: \mu_3 = \mu_4 \quad H_0: \mu_3 = \mu_5$$
$$H_1: \mu_3 \neq \mu_4 \quad H_1: \mu_3 \neq \mu_5$$
$$H_0: \mu_4 = \mu_5$$
$$H_1: \mu_4 \neq \mu_5$$

Puesto que una de las suposiciones del modelo es que las varianzas poblacionales son iguales, puede contrastarse cada una de estas hipótesis utilizando un contraste T de varianza conjunta de dos colas (bilateral), tal como se ha descrito en el Capítulo 9. En este contraste, el estadístico del contraste es:

$$T_{n_i+n_j-2} = \frac{|\bar{X}_{i\cdot} - \bar{X}_{j\cdot}|}{\sqrt{S_p^2(1/n_i + 1/n_j)}}$$

donde S_p^2 es el estimador conjunto de σ^2, la varianza poblacional común, basado en muestras extraídas de las poblaciones *i* y j. En el caso que ahora estamos considerando, disponemos de otro estimador de σ^2, es decir, MS_E. Puesto que este estimador está basado en todos los datos disponibles, el contraste *T* puede mejorarse utilizando

$$T_{N-k} = \frac{|\bar{X}_{i\cdot} - \bar{X}_{j\cdot}|}{\sqrt{MS_E(1/n_i + 1/n_j)}}$$

como estadístico del contraste. La ejecución manual de $k(k - 1)/2$ contrastes *T* individuales es laboriosa, pero se puede hacer. Muchos paquetes estadísticos comerciales incluyen un procedimiento para ejecutar los contrastes *T* de Bonferroni basándose en el estadístico anterior.

Este procedimiento implica un serio problema, que debe ser tratado con cuidado. Es decir, si todos los contrastes se ejecutan a un nivel de significación α, la probabilidad general de hacer al menos un rechazo incorrecto, indicada por α', es mayor que α, y su valor generalmente es desconocido. Sin embargo, se puede demostrar que siempre que se realice un conjunto de *c* contrastes, cada uno de ellos al nivel de significación α, α' es, como mucho, $1 - (1 - \alpha)^c$\ Por ejemplo, si $k = 5$, entonces, como se ha demostrado previamente, hay 10 pares posibles de) medias que pueden compararse. Si cada contraste se realiza al nivel $\alpha = 0.05$, la probabilidad de efectuar al menos un rechazo incorrecto es, como mucho, de $1 - (1 - 0.05)^{10} = 0.40$. Es fácil apreciar que, a medida que *k* aumenta, la probabilidad general de error puede ser inaceptablemente alta. Para compensar este problema, se sugiere que sólo *se realicen los contrastes de interés real para el investigador.* De forma pararela, considérese un médico que solicita que se ejecuten una serie de pruebas de diagnóstico a un paciente concreto. Cada prueba tiene un índice de falso-positivo predeterminado; es decir, existen posibilidades de que cada prueba

detecte una afección que, de hecho, no está presente. Si se realizan suficientes pruebas, se obtendrá eventualmente un resultado falso-positivo. Si se realizan las pruebas suficientes, eventualmente se observará que el paciente tiene una afección que, de hecho, no está presente y se cometerá un error. El médico no deberá comprobar sistemáticamente cada trastorno o enfermedad conocidos, sino más bien sólo contrastará aquellos estados que sospeche que puedan ser la causa del problema del paciente.

Para realizar los contrastes *T* de Bonferroni de forma responsable, elegimos algún límite superior razonablemente pequeño, *b,* para la probabilidad de cometer al menos un rechazo incorrecto. A continuación podemos realizar cada contraste *T* al nivel *b/c* de significación donde *c* indica el número real de contrastes a realizar. Por ejemplo, si queremos que α' sea como máximo 0.10 y ejecutamos todos los contrastes posibles para $k = 5$ grupos, realizaremos cada una de nuestras comparaciones por parejas al nivel $\alpha = 0.10/10 = 0.01$ de significación. El Ejemplo 10.2.1 ilustra este procedimiento.

Ejemplo 10.2.1. Un ingeniero químico está estudiando un polímero recientemente desarrollado para que sea utilizado en la eliminación de los residuos tóxicos del agua. Los experimentos se realizan a cinco temperaturas diferentes. La respuesta observada es el porcentaje de impurezas eliminadas por el tratamiento. Se han obtenido los datos siguientes:

Temperatura				
I	II	III	IV	V
40	36	49'	47	55
45	42	51	49	60
42	38	53	51	62
48	39	53	52	63
50	37	52	50	59
51	40	50	51	61

En la Tabla 10.3 se presenta la tabla ANOVA para estos datos. Obsérvese que la hipótesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ puede rechazarse con $P < 0.01$. Supóngase que deseamos realizar las 10 comparaciones por parejas y que queremos que la probabilidad de cometer al menos un rechazo incorrecto sea como máximo de 0.10. Para ello, cada contraste Tdebe realizarse al nivel $b/c = 0.10/10 = 0.01$. El contraste estadístico para el contraste con dos colas es:

$$T_{N-k} = \frac{|\bar{X}_{i\cdot} - \bar{X}_{j\cdot}|}{\sqrt{MS_E(1/n_i + 1/n_j)}}$$

Puesto que los tamaños de las muestras son todos iguales a 6, en este caso el estadístico *T* es de la forma:

$$T_{25} = \frac{|\bar{X}_{i\cdot} - \bar{X}_{j\cdot}|}{\sqrt{7.63\left(\frac{1}{6} + \frac{1}{6}\right)}}$$

donde $MS_E = 7.63$ se obtiene del ANOVA de la Tabla 10.3. En la tabla de T puede observarse que:

$$P[T_{25} \geq 2.787] = 0.005$$

Por lo cual, para que el valor *P* del contraste *T* bilateral sea, a lo sumo, de 0.01, el valor observado del estadístico debe ser, al menos, de 2.787. Este punto se denomina *punto crítico*

Tabla 10.3. ANOVA para los datos del Ejemplo 10.2.1

Origen	*DF*	*SS*	*MS*	*F*
Tratamiento	4	1458.13	364.53	47.78
Error	25	190.67	7.63	
Total	29	1648.80		

para un contraste de nivel $\alpha = 0.01$. Es crítico en el sentido de que los puntos en o por encima de este valor llevan a rechazar H_0, siempre que α haya sido prefijada en 0.01. Obsérvese que para que:

$$\frac{|\bar{X}_{i\cdot} - \bar{X}_{j\cdot}|}{\sqrt{7.63\left(\frac{1}{6} + \frac{1}{6}\right)}}$$

sea igual o mayor que 2.787, la diferencia en valor $|\bar{X}_{i\cdot} - \bar{X}_{j\cdot}|$ debe ser igual o mayor que

$$2.787\sqrt{7.63\left(\tfrac{1}{6} + \tfrac{1}{6}\right)} = 4.44$$

Para realizar cada contraste, sólo necesitamos comparar las diferencias en valor absoluto entre las medias muestrales respectivas con este valor. Si $|\bar{X}_{i\cdot} - \bar{X}_{j\cdot}| \geq 4.44$, entonces rechazamos H_0: $\mu_i = \mu_j$ y concluimos que las dos medias poblacionales μ_i y μ_j son diferentes. Por ejemplo, para contrastar H_0: $\mu_1 = \mu_2$ hallamos $\bar{X}_1$ y $\bar{X}_2$.. En este caso $\bar{X}_{1\cdot} = 46.0$ $\bar{X}_{2\cdot} = 38.7$. La diferencia $|\bar{X}_{1\cdot} - \bar{X}_{2\cdot}|$ es 7.3. Dado que $7.3 > 4.44$, podemos concluir que $\mu_1 \neq \mu_2$. Los nueve pares restantes de medias pueden compararse de forma similar.

Obsérvese que la magnitud de la diferencia entre $\bar{X}_{i\cdot}$ y $\bar{X}_{j\cdot}$, necesaria para declarar que μ_i y μ_j son diferentes, depende de n_i y n_j. Si todos los tamaños muestrales son iguales, como en el Ejemplo 10.2.1, puede utilizarse una única diferencia crítica para realizar todas las comparaciones por parejas. Sin embargo, si los tamaños muestrales no son los mismos, se debe hallar una diferencia crítica por separado para cada contraste.

Contraste de Duncan de rango múltiple

El contraste de Duncan de rango múltiple fue desarrollado por D. B. Duncan. Es uno de los métodos más antiguos, de los actualmente en uso, para comparar medias, y es citado a menudo en la bibliografía estadística. Por esta razón, el lector debe familiarizarse con su uso. A diferencia de los contrastes *T* de Bonferroni tratados anteriormente, se hace un intento de explicar la colocación del par de medias dentro de la lista de medias muestrales ordenadas. Por ejemplo, supongamos que tenemos una colección de cinco medias muestrales $\bar{X}_{1\cdot}$, $\bar{X}_{2\cdot}$, $\bar{X}_{3\cdot}$, $\bar{X}_{4\cdot}$ y $\bar{X}_{5\cdot}$, ordenadas de menor a mayor. Se utiliza el estadístico $|\bar{X}_{1\cdot} - \bar{X}_{5\cdot}|$ para contrastar H_0: $\mu_1 = \mu_5$. El método de Bonferroni compara estas medias poblacionales mediante un contraste T que no tiene en cuenta el hecho de que las medias muestrales $\bar{X}_{1\cdot}$ y $\bar{X}_{5\cdot}$ son los extremos de un conjunto de cinco medias; el procedimiento de Duncan reconoce este hecho y hace un ajuste con ello. El ajuste implica el cambio de los puntos críticos, de forma que las medias muestrales que se sitúen una al lado de la otra no requieran presentar tanta diferencia como las que están separadas, con el fin de concluir que las medias poblacionales correspondientes son «significativamente diferentes». El contraste fue originalmente diseñado para ser utilizado con muestras del mismo tamaño. Sin embargo, C. Y. Kramer lo amplió para incluir muestras de tamaño distinto. Describimos en primer lugar el proceso para muestras de *igual* tamaño.

Ejemplo 10.2.2. Se analizan muestras de agua de cuatro lagos en cuanto a su contenido en fósforo. El nivel de fósforo viene dado en partes por millón (ppm). Se obtienen los siguientes datos estadísticos, a partir de 20 muestras aleatoriamente seleccionadas de cada lago:

$$T_{1.} = 0.40 \qquad T_{4.} = 1.00$$
$$T_{2.} = 0.20 \qquad T_{..} = 1.62$$
$$T_{3.} = 0.02 \qquad \sum\sum X_{ij}^2 = 0.2880$$

El ANOVA para el experimento se da en la Tabla 10.4. Utilizaremos la distribución $F_{3,60}$ para aproximar la distribución $F_{3,76}$. En la tabla de F, observamos que

$$P[F_{3,60} \geq 3.34] = 0.025$$

y

$$P[F_{3,60} \geq 2.76] = 0.05$$

Puesto que 3.0333, valor observado para nuestro estadístico F, se sitúa entre 2.76 y 3.34, y el valor P para el contraste de igualdad de medias está entre 0.025 y 0.05, es decir, es un valor P pequeño, rechazamos H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ y concluimos que hay diferencias en los niveles medios de fósforo en los cuatro lagos. Queda por ver, precisamente, cuáles son esas diferencias.

El contraste de Duncan de rango múltiple se ha diseñado para detectar diferencias en medias poblacionales comparando medias muestrales. Se hace dividiendo en subgrupos las k medias muestrales y, por tanto, las k medias poblacionales, de forma que no se considera que las medias, dentro de los subgrupos, sean significativamente diferentes. El contraste se lleva a efecto del modo siguiente:

1. Se ordenan en orden ascendente las k medias muestrales.
2. Se considera cualquier subconjunto de p medias muestrales $2 < p < k$. Para que las medias de cualquiera de las poblaciones correspondientes se consideren diferentes, el rango de las medias en el subgrupo (de la mayor a la más pequeña) debe sobrepasar un determinado valor, denominado el *menor rango significativo SSRp.*
3. El menor rango significativo se calcula mediante la Tabla X del Apéndice B y la siguiente fórmula:

$$SSR_p = r_p \sqrt{\frac{MS_E}{n}}$$

Tabla 10.4. ANOVA para los datos del Ejemplo 10.2.2. La hipótesis nula de igualdad de medias puede rechazarse con $0.025 < P < 0.05$

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	3	0.0272	0.0091	3.0333
Error	76	0.2280	0.0030	
Total	79	0.2552		

donde r_p = menor rango significativo *«studentizado»* obtenido de la Tabla X
MS_E = cuadrado medio del error del ANOVA
n = tamaño muestral común
y = grados de libertad para MS_E

4. Los resultados se resumen subrayando cualquier subconjunto de medias adyacentes que no se consideran significativamente diferentes al nivel α que se haya seleccionado.

El procedimiento se describe continuando el análisis de los datos del Ejemplo 10.2.2.

Ejemplo 10.2.3. Los niveles medios de fósforo estimados para los cuatro lagos son:

$$\bar{x}_{1\cdot} = \frac{T_{1\cdot}}{20} = \frac{0.40}{20} = 0.02 \text{ ppm}$$

$$\bar{x}_{2\cdot} = \frac{T_{2\cdot}}{20} = \frac{0.20}{20} = 0.01 \text{ ppm}$$

$$\bar{x}_{3\cdot} = \frac{T_{3\cdot}}{20} = \frac{0.02}{20} = 0.001 \text{ ppm}$$

$$\bar{x}_{4\cdot} = \frac{T_{4\cdot}}{20} = \frac{1.00}{20} = 0.05 \text{ ppm}$$

En orden ascendente:

$\bar{x}_{3\cdot}$	$\bar{x}_{2\cdot}$	$\bar{x}_{1\cdot}$	$\bar{x}_{4\cdot}$
0.001	0.01	0.02	0.05

A continuación, construimos una tabla que dé los valores de SSR_p para $\alpha = 0.05$. La tabla se basa en $\gamma = 60$ grados de libertad, porque este valor es el más próximo al valor 76 que requiere el contraste de Duncan. Obsérvese que $MS_E = 0.0030$ se obtiene del ANOVA del Ejemplo 10.2.2; el valor de r_p viene dado por la Tabla X del Apéndice B, y $SSR_p = r_p\sqrt{MS_E/20}$.

p	2	3	4
	2.829	2.976	3.073
SSR_p	0.0346	0.0364	0.0376

Las diferencias entre las medias poblacionales se detectan comparando la mayor media muestral con la más pequeña, la mayor con la siguiente más pequeña, y así sucesivamente. De este modo, necesitamos potencialmente considerar pares de medias muestrales en el orden

1. $\bar{x}_{4\cdot} - \bar{x}_{3\cdot}$
2. $\bar{x}_{4\cdot} - \bar{x}_{2\cdot}$
3. $\bar{x}_{4\cdot} - \bar{x}_{1\cdot}$
4. $\bar{x}_{1\cdot} - \bar{x}_{3\cdot}$
5. $\bar{x}_{1\cdot} - \bar{x}_{2\cdot}$
6. $\bar{x}_{2\cdot} - \bar{x}_{3\cdot}$

En todo caso, una vez que se ha visto que un grupo de medias no es significativamente diferente, nuevos contrastes no afirmarán que lo sean. Por lo tanto, en la práctica, puede no ser preciso realizar todas las comparaciones indicadas. Las comparaciones necesarias se recogen en la Tabla 10.5.

Llegados a este punto, no hace falta hacer nuevas comparaciones, ya que no se han detectado diferencias entre los niveles medios de fósforo de los lagos 1, 2 y 3. Resumiendo, podamos concluir que $\mu_4 \neq \mu_3$ y $\mu_4 \neq \mu_2$. No se ha detectado ninguna otra diferencia.

Tabla 10.5.

Diferencia d	Número en el subgrupo p	SSR_p (de la tabla)	¿Es $d > SSR_p$?	Agrupaciones
$\bar{x}_{4.} - \bar{x}_{3.} = 0.0490$	4	0.0376	Sí	$\bar{x}_{3.}\ \bar{x}_{2.}\ \bar{x}_{1.}\ \bar{x}_{4.}$
$\bar{x}_{4.} - \bar{x}_{2.} = 0.04$	3	0.0364	Sí	$\bar{x}_{3.}\ \bar{x}_{2.}\ \bar{x}_{1.}\ \bar{x}_{4.}$
$\bar{x}_{4.} - \bar{x}_{1.} = 0.03$	2	0.0346	No	$\bar{x}_{3.}\ \bar{x}_{2.}\ \underline{\bar{x}_{1.}\ \bar{x}_{4.}}$
$\bar{x}_{1.} - \bar{x}_{3.} = 0.019$	3	0.0364	No	$\bar{x}_{3.}\ \bar{x}_{2.}\ \underline{\bar{x}_{1.}\ \bar{x}_{4.}}$

C. Y. Kramer, en 1956, amplió el contraste de Duncan de rango múltiple a fin de incluir muestras de tamaños distintos. Se realiza de forma similar a la del proceso original de Duncan, con dos variaciones. En particular, el menor rango significativo para el contraste ajustado, representado por SSR'_p viene dado por:

$$SSR'_p = r_p\sqrt{MS_E}$$

donde r_p = menor rango significativo *«studentizado»* obtenido de la Tabla X.
MS_E = cuadrado medio del error del ANOVA

El estadístico para comparar dos medias poblacionales μ_i y μ_j es

$$|\bar{x}_{i.} - \bar{x}_{j.}|\sqrt{\frac{2n_i n_j}{n_i + n_j}}$$

Este estadístico refleja los tamaños de las muestras, y las medias μ_i *y* μ_j son consideradas diferentes si, y sólo si, el valor observado del estadístico excede el de SSR_p. El proceso se describe en el Ejemplo 10.2.4.

Ejemplo 10.2.4. En el Ejemplo 10.1.6 se concluyó que los tres tratamientos del acné diferían en su efecto medio. Para concretar las diferencias utilizamos el contraste de Duncan de rango múltiple. Puesto que los tamaños de las muestras no son iguales, seria apropiado un ajuste de Kramer. De los resultados previos se sabe que:

$n_1 = 10$ $\quad \bar{x}_{1.} = 49.46$ $\quad MS_E = 4.07$
$n_2 = 12$ $\quad \bar{x}_{2.} = 68.73$ $\quad \sqrt{MS_E} \cong 2.02$
$n_3 = 13$ $\quad \bar{x}_{3.} = 63.60$

En orden ascendente, las medias muestrales son:

$\bar{x}_{1.}$	$\bar{x}_{3.}$	$\bar{x}_{2.}$
49.46	63.60	68.73

Utilizando la Tabla X del Apéndice B con $\alpha = 0.01$ y $\gamma = 30$ grados de libertad (como una aproximación a $\gamma = 32$ grados de libertad), y la fórmula $SSR'_p = r_p\sqrt{MS_E}$, obtendremos los siguientes valores ajustados para el menor rango significativo:

p	2	3
r_p	3.889	4.506
SSR'_p	7.85	9.10

Para comparar μ_2 con μ_1, evaluamos el estadístico:

$$|\bar{x}_{2\cdot} - \bar{x}_{1\cdot}|\sqrt{\frac{2n_1n_2}{n_1+n_2}} = |68.73 - 49.46|\sqrt{\frac{2 \cdot 10 \cdot 12}{10+12}}$$

$$= 63.65$$

Puesto que este valor supera a $SSR'_3 = 9.10$, podemos concluir que $\mu_1 \neq \mu_2$. El estadístico

$$|\bar{x}_{2\cdot} - \bar{x}_{3\cdot}|\sqrt{\frac{2n_2n_3}{n_2+n_3}} = |68.73 - 63.60|\sqrt{\frac{2 \cdot 12 \cdot 13}{12+13}}$$

$$= 18.12$$

se utiliza para comparar μ_2 con μ_3. Dado que este valor supera a $SSR'_2 = 7.85$, podemqs concluir que $\mu_2 \neq \mu_3$. La última comparación es la de μ_3 con μ_1. El estadístico necesario es

$$|\bar{x}_{3\cdot} - \bar{x}_{1\cdot}|\sqrt{\frac{2n_3n_1}{n_3+n_1}} = |63.60 - 49.46|\sqrt{\frac{2 \cdot 13 \cdot 10}{13+10}}$$

$$= 47.54$$

Ya que este valor también supera a SSR'_2, concluim $\mu_1 \neq \mu_3$. Resumiendo estos resultados, tenemos

Antiguo	Nuevo	Actual
$\bar{x}_{1\cdot}$	$\bar{x}_{3\cdot}$	$\bar{x}_{2\cdot}$

Es decir, al nivel $\alpha = 0.01$ podemos afirmar que cada una de las medias es significativamente diferente de cada una de las otras.

Recuérdese que si los tamaños muestrales son iguales, quizá no tendremos que hacer todas las comparaciones posibles. Esto se debe al hecho de que, cuando los tamaños muestrales son iguales, siempre que el par de medias más extremo no sea significativamente diferente, se supone que todas las medias del subgrupo son iguales sin que se requiera que se realicen más contrastes. Esto no es así con tamaños muestrales diferentes. En ese caso, deben realizarse *todas* las comparaciones.

Nota sobre los cálculos

Como puede apreciarse, los cálculos necesarios para realizar un análisis de varianza son extensos. Hay paquetes informáticos comerciales disponibles para realizar el análisis inicial, y muchos incluyen la posibilidad de realizar comparaciones por parejas y múltiples. Así, el principal trabajo del investigador es reconocer una situación experimental que necesita el análisis de la varianza de una vía y tener la habilidad de interpretar correctamente los resultados obtenidos con el programa *(printout)*.

En el Ejemplo 10.2.5 presentamos la salida *(output)* de un programa SAS utilizado para analizar los datos del Ejemplo 10.2.1. Se realizan los contrastes *T de* Bonferroni y el contraste de Duncan de rango múltiple con fines ilustrativos. En la práctica, deberá utilizarse uno de los

dos pero no ambos. El código SAS del programa necesario para producir el *output* se incluye en las Herramientas Computacionales, al final del capítulo.

Ejemplo 10.2.5. A continuación, se muestran los resultados del SAS para el análisis de los datos del Ejemplo 10.2.1. Obsérvese que la fuente de variación, a la que nosotros denominamos «Tratamiento», en el SAS se llama «Model». Por otro lado, el *output* del SAS concuerda con el obtenido mediante la calculadora manual y presentado en la Tabla 10.3 (aparte de algunas diferencias de redondeo). Obsérvese que el procedimiento de Duncan no encuentra diferencias entre el porcentaje medio de impurezas eliminado por los polímeros 3 y 4, puesto que estos polímeros están rotulados como B en el *printout.* Todos los demás polímeros difieren de éstos y entre sí. Los resultados del contraste de Bonferroni difieren ligeramente de los obtenidos utilizando el procedimiento de Duncan. El contraste de Bonferroni no detecta diferencias entre las respuestas medias de los polímeros 4 y 1 puesto que ambos están rotulados con una C en el *printout.* La diferencia de interpretación entre los dos contrastes se debe al hecho de que el contraste de Bonferroni es muy conservador. Para obtener una probabilidad general de efectuar al menos un rechazo incorrecto de 0.05, cada una de las 10 comparaciones emparejadas posibles se realiza al nivel 0.05/10 = 0.005.

SAS

Analysis-of-Variance Procedure

Dependent Variable: PERCENT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > *F*
Model	4	1458.133333	364.533333	47.80	0.0001
Error	25	190.666667	7.626667		
Corrected Total	29	1648.800000			

R-Squares	C.V.	Root MSE	PERCENT Mean
0.884360	5.613094	2.761642	49.2000000

Source	DF	Anova *SS*	Mean Square	*F* Value	Pr > *F*
TEMP	4	1458.133333	364.533333	47.80	0.0001

SAS

Analysis-of-Variance Procedure

Duncan's Multiple Range Test for variable: PERCENT

NOTE: This test controls the type I comparisonwise error rate, not the experimentwise error rate.

Alpha = 0.05 DF = 25 MSE = 7.626667

Number of Means 2 3 4 5

Critical Range 3.281 3.447 3.562 3.632

Means with the same letter are not signif icantly different.

Duncan Grouping	Mean	N	TEMP
A	60.000	6	5
B	51.333	6	3
B			
B	50.000	6	4
C	46.000	6	1
D	38.667	6	7

SAS

Analysis-of-Variance Procedure

Bonferroni (Dunn) T tests for variable: PERCENT

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha = 0.05 DF = 25 MSE = 7.62667

Minimum Significant Difference = 4.908

Means with the same letter are not signif icantly diferent.

Bon Grouping	Mean	*N*	TEMP
A	60.000	6	5
B	51.333	6	3
B			
C B	50.000	6	4
C			
C	46.000	6	1
D	38.667	6	2

EJERCICIOS 10.2

1. Considérense los datos del Ejemplo 10.2.1.
   a) Verificar que

$$T_{1\cdot} = 276 \qquad \bar{X}_{1\cdot} = 46.0 \qquad \sum_{i=1}^{5}\sum_{j=1}^{6} X_{ij}^2 = 74\,268$$
$$T_{2\cdot} = 232 \qquad \bar{X}_{2\cdot} = 38.7$$
$$T_{3\cdot} = 308 \qquad \bar{X}_{3\cdot} = 51.3$$
$$T_{4\cdot} = 300 \qquad \bar{X}_{4\cdot} = 50.0$$
$$T_{5\cdot} = 360 \qquad \bar{X}_{5\cdot} = 60.0$$
$$T_{\cdot\cdot} = 1476 \qquad \bar{X}_{\cdot\cdot} = 49.2$$

   b) Verificar el ANOVA de la Tabla 10.3.
   c) Contrastar las otras nueve hipótesis de la forma H_0: $\mu_i = \mu_j$ frente a H_1: $\mu_i \neq \mu_j$. Registrar esquemáticamente los resultados ordenando las medias muestrales de menor a mayor y subrayando los pares que no son significativamente diferentes.
2. En cada caso, determinar el número de comparaciones posibles dos a dos.
   α) k = 3
   b) k = 6
   c) k = 10
3. Sea α = 0.05 y supongamos que deben realizarse todas las comparaciones posibles dos a dos. Para cada valor de *k* dado en el Ejercicio 2, hallar un límite superior para α', probabilidad general de efectuar al menos un rechazo incorrecto. En cada caso, ¿qué nivel α deberá utilizarse para garantizar que $\alpha' \leq 0.10$?
4. Los científicos comprometidos en el tratamiento del agua residual de arenas asfálticas estudiaron tres métodos de tratamiento para la eliminación del carbono orgánico. (Basado en W. R. Pirie, *Statistical Planning and Analysis for Treatments ofTar Sand Wastewater,* Centro de Información Técnica, Oficina de Información Tecnológica y Científica, Departamento de Energía de Estados Unidos.) Los tres métodos de tratamiento utilizados fueron: flotación de aire (FA), separación por espuma (SF) y coagulación ferroclórica (CFC). Las mediciones del material de carbono orgánico para los tres tratamientos arrojaron los siguientes datos:

FA	SE	CFC
34.6	38.8	26.7
35.1	39.0	26.7
35.3	40.1	27.0
35.8	40.9	27.1
36.1	41.0	27.5
36.5	43.2	28.1
36.8	44.9	28.1
37.2	46.9	28.7
37.4	51.6	30.7
37.7	53.6	31.2

   a) Contrastar H_0: $\mu_1 = \mu_2 = \mu_3$ al nivel α = 0.10.
   b) Si se rechaza H_0, utilizar los contrastes *T* de Bonferroni para determinar con precisión las diferencias entre las medias poblacionales. Sea *OÍ',* a lo sumo, 0.06.

5. Se sabe que se ha arrojado material tóxico a un río que entra en una gran área de pesca comercial en agua salada. Los ingenieros civiles han estudiado la forma en que el agua transporta él material tóxico, midiendo la cantidad de material (en partes por millón) hallado en las ostras recogidas en tres lugares diferentes, desde la salida del estuario hasta la bahía donde se realiza la mayor parte de la pesca comercial. A continuación, se presentan los resultados:

Lugar 1 (estuario)	**Lugar 2** (lejos **de la bahía)**	**Lugar 3 (cerca de la bahía)**
15	19	22
26	15	26
20	10	24
20	26	26
29	11	15
28	20	17
21	13	24
26	15	
	18	

   a) Contrastar las diferencias en las medias de partes por millón de material tóxico hallado en las ostras recogidas en los tres lugares.
   b) Si se rechaza H_0: $\mu_1 = \mu_2 = \mu_3$, utilizar los contrastes T de Bonferroni para señalar las diferencias existentes. Sea α' como máximo, 0.15.
6. Continuar el análisis de los datos del Ejercicio 1 de la Sección 10.1 aplicando el contraste de Duncan de rango múltiple con un $\alpha = 0.01$.
7. Continuar el análisis de los datos del Ejercicio 2 de la Sección 10.1 aplicando el contraste de Duncan de rango múltiple con un $\alpha = 0.01$.
8. Continuar el análisis de los datos del Ejercicio 3 de la Sección 10.1 aplicando el contraste de Duncan de rango múltiple con el ajuste de Kramer con un $\alpha = 0.01$.
9. Continuar el análisis de los datos del Ejercicio 4 de la Sección 10.1 aplicando el contraste de Duncan de rango múltiple con un $\alpha = 0.01$.
10. Continuar el análisis de los datos del Ejercicio 5 de la Sección 10.1 aplicando el contraste de Duncan de rango múltiple con el ajuste de Kramer.
11. Se ha realizado un estudio entre los obreros de empresas en las que éstos son expuestos habitualmente al amianto. Se realizó un programa de protección contra la asbestosis y un estudio posterior dividió la muestra obtenida en tres grupos: los que habían dejado de fumar, los que habían reducido el nivel de fumar y los que habían mantenido el nivel de fumar tras el programa de protección. Se obtuvieron estos datos sobre el nivel de monóxido de carbono alveolar (CO_α) de los tres grupos.

Los que lo dejaron	**Los que lo redujeron**	**Los que continuaron**
36	**34**	**28**
40	**5**	**2**
19	**47**	**51**
25	**37**	**33**
43	**46**	**28**
54		**29**
		30

α) Contrastar la igualdad de las medias.
b) Si se rechaza la hipótesis nula de las medias iguales, señalar las diferencias qué existen realizando los contrastes *T* de Bonferroni adecuados. Mantener un nivel general α de, como máximo, 0.15.

(Basado en las medias halladas en Kaye Kilburn y Raphael Warshaw, «Effects of Individually Motivating Smoking Cessation in Male Blue Collar Workers», *American Journal of Public Health,* noviembre de 1990, págs. 1334-1337.)

10.3. EFECTOS ALEATORIOS (OPCIONAL)

El modelo presentado en la Sección 10.1 se llama *modelo de efectos fijos.* Recuérdese que esto implica que el experimentador selecciona específicamente los niveles del factor, o «tra+ tamientos», por su particular interés. El propósito del experimento es realizar inferencias acerca de las medias de determinadas poblaciones de las que se han extraído las muestras. No preocupa ninguna otra población. Sin embargo, si queremos hacer una generalización más amplia, relativa a un conjunto mayor de poblaciones y no solamente al de las *k* poblaciones de las que tomamos muestras, el modelo se llama entonces *modelo de efectos aleatorios.* En este caso, se considera que las *k* poblaciones muestreadas son una muestra aleatoria de población nes, extraídas del conjunto mayor de poblaciones. La hipótesis que interesa no es $\mu_1 = \mu_2 = \bullet\bullet\bullet = \mu_k$. Más bien, queremos determinar si existe alguna variabilidad entre las medias poblacionales del conjunto mayor. Consideremos el Ejemplo 10.3.1.

Ejemplo 10.3.1. Los medios de cultivo bacteriológico utilizados en los laboratorios de los hospitales proceden de diversos fabricantes. Se sospecha que la calidad de estos medios de cultivo varía de un fabricante a otro. Para comprobar esta teoría, se hace una lista de fabricantes de un medio de cultivo concreto, se seleccionan aleatoriamente los nombres de tres de los que aparecen en la lista y se comparan las muestras de los instrumentos procedentes de éstos. La comparación se realiza colocando sobre una placa dos dosis, en gotas, de una suspensión medida de un microorganismo clásico, *Escherichia coli,* dejando al cultivo crecer durante veinticuatro horas y determinando después el número de colonias (en millares) del microorganismo que aparecen al final del período. El propósito no es comparar precisamente estos tres fabricantes; se les eligió aleatoriamente y únicamente constituyen una muestra de todos los fabricantes de instrumentos. Más bien, lo que buscamos es apoyo estadístico para el supuesto general de que la calidad del instrumental difiere entre fabricantes.

El modelo de efectos aleatorios puede escribirse matemáticamente del siguiente modo:

Modelo

$$X_{ij} \equiv \mu + T_i + E_{ij} \qquad \begin{matrix} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n_i \end{matrix}$$

donde μ = efecto global medio
$T_i = \mu_i - \mu$ (μ_i es la media de la i-ésima población seleccionada para el estudio)
$E_{ij} = X_{ij} - \mu_i$ = error residual o aleatorio

Admitimos lo siguiente:

Supuestos del modelo

1. Las *k* muestras representan muestras aleatorias independientes extraídas de *k* poblaciones seleccionadas aleatoriamente de un conjunto mayor de poblaciones.
2. Todas las poblaciones del conjunto más amplio son normales, de modo que cada una de las *k* poblaciones muestreadas es también normal.

3. Todas las poblaciones del conjunto más amplio tienen la misma varianza, y, por lo tanto, cada una de las *k* poblaciones muestreadas tiene también varianza σ^2.
4. Las variables $T_1, T_2, ..., T_k$ son variables aleatorias normales independientes, cada una con media 0 y varianza común σ^2_{Tr}.

El propio modelo y los tres primeros supuestos del modelo son semejantes a los del modelo de efectos fijos. Sin embargo, el supuesto 4 del modelo expresa matemáticamente una importante diferencia entre los dos. En el modelo de efectos fijos, el experimentador elige minuciosamente los tratamientos, o niveles, utilizados en el experimento por su particular interés. Si se replicase (repitiese) el experimento, se utilizarían los mismos tratamientos. Esto es, se muestrearían las mismas poblaciones cada vez y los *k* efectos del tratamiento $\mu_i - \mu$ no variarían. Esto implica que en el modelo de efectos fijos, los *k* términos $\mu_i - \mu$ se consideran *constantes desconocidas.* En el modelo de efectos aleatorios, no es éste el caso. Puesto que el primer paso en un experimento de efectos aleatorios es seleccionar aleatoriamente *k* poblaciones, las elegidas variarán de replicación en replicación. De este modo, en el modelo de efectos aleatorios, *los k* términos $T_i = \mu_i - \mu$ no son constantes, sino que son, de hecho, *variables aleatorias* cuyos valores para una determinada replicación dependen de la elección de las k poblaciones a estudiar. Estas variables se suponen variables aleatorias normales independientes con media 0 y varianza común σ^2_{Tr}.

Si las medias poblacionales en el conjunto mayor son iguales, no variarán los efectos del tratamiento $T_i = \mu_i - \mu$. Es decir, σ^2_{Tr} será cero. Así, en el modelo de efectos aleatorios, la hipótesis de medias iguales se contrasta considerando:

H_0: $\sigma^2_{Tr} = 0$ (no hay variabilidad en los efectos del tratamiento)
H_1: $\sigma^2_{Tr} \neq 0$

A pesar del hecho de que el modelo de efectos aleatorios difiera teóricamente del de efectos fijos, los datos se manejan exactamente del mismo modo. En la columna de los cuadrados medios esperados, se presenta el único cambio necesario en el formato del análisis de la varianza. Para este modelo, se cambia $E[MS_{Tr}]$ por la expresión:

$$E[MS_{Tr}] = \sigma^2 + n_0\sigma^2_{Tr}$$

donde:

$$n_0 = \frac{N - \sum_{i=1}^{k} n_i^2/N}{k-1}$$

En la Tabla 10.6 se esquematiza el análisis de la varianza. Obsérvese de nuevo que, si *Ho* es cierta ($\sigma^2_{Tr} = 0$), esperaremos que MS_{Tr} y MS_E tengan valores próximos, puesto que ambos estiman el mismo parámetr σ^2. Si H_0 no es cierta, esperaremos que MS_{Tr} sea mayor que MS_E, forzando así al cociente *F* a ser mayor que 1. De este modo, el contraste rechaza H_0, para valores observados del estadístico $F_{k-1,\ N-k}$ que sean demasiado grandes para haberse presentado por azar.

Ejemplo 10.3.2. En el experimento del Ejemplo 10.3.1, se eligen aleatoriamente tres fabricantes de medios de cultivo bacteriológico. De las existencias de cada uno se extraen muestras de tamaño 10 y se comparan. Se obtienen los siguientes datos estadísticos:

$$T_{1\cdot} = 527 \qquad T_{3\cdot} = 480$$
$$T_{2\cdot} = 502 \qquad T_{\cdot\cdot} = 1509$$
$$\sum_{i=1}^{3}\sum_{j=1}^{10} X_{ij}^2 = 76\,511$$

Tabla 10.6. Análisis de la varianza: clasificación simple, diseño completamente aleatorio con efectos aleatorios

Origen de la variación	Grados de libertad *DF*	Suma de cuadrados *SS*	Cuadrado medio *MS*	Cuadrado medio esperado	Cociente *F*
Tratamiento o nivel	$k - 1$	$\sum_{i=1}^{k} \frac{T_{i\cdot}^2}{n_i} - \frac{T_{\cdot\cdot}^2}{N}$	$\frac{SS_{\text{Tr}}}{k-1}$	$\sigma^2 + n_0\sigma_{\text{Tr}}^2$	$\frac{MS_{\text{Tr}}}{MS_E}$
Residuo o error	$N-k$	$SS_{\text{Total}} - SS_{\text{Tr}}$	$\frac{SS_E}{N-k}$	σ^2	
Total	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{N}$			

El ANOVA para estos datos se muestra en la Tabla 10.7. Obsérvese que $P[F_{2,27} > 3.35] = 0.05$ y $P[F_{2,27} > 2.51] = 0.10$. Puesto que el valor observado del estadístico, 3, se sitúa entre 2.51 y 3.35, el valor P se sitúa entre 0.05 y 0.10. Podemos rechazar H_0: $\sigma_{\text{Tr}}^2 = 0$ y concluir diciendo que hay evidencia de la existencia de alguna variabilidad entre la calidad del material de los diferentes fabricantes.

En el modelo de efectos aleatorios no se necesita que se lleven a cabo más contrastes, incluso aunque H_0 haya sido rechazada. El propósito del experimento es hacer un planteamiento general relativo a las poblaciones de las que se extraen las k muestras poblacionales. Esto es lo que se ha hecho.

EJERCICIOS 10.3

1. Se realiza un estudio sobre el comportamiento del carbonero macho. El propósito es determinar si hay diferencia en la altura media (en metros) a la que el pájaro realiza diversas actividades. Se hace una lista con 10 de las principales actividades. Se seleccionan aleatoriamente cuatro de estas actividades para establecer un patrón: canto, alimentación, arreglo y limpieza de las plumas, y reposo. De ellas se obtuvieron los siguientes datos, a partir de 20 observaciones de cada una:

Canto	Alimentación	Limpieza	Reposo
$T_{1\cdot} = 186$	$T_{2\cdot} = 44$	$T_{3\cdot} = 120$	$T_{4\cdot} = 70$

$$\sum_{i=1}^{4}\sum_{j=1}^{20} X_{ij}^2 = 7915.8$$

De acuerdo con estos datos, ¿se puede concluir que las distintas actividades se llevan a cabo a diferentes alturas? Explíquese la respuesta sobre la base del ANOVA y del valor P obtenido.

¿Qué modelo, el de efectos fijos o el de efectos aleatorios, es más probable que se encuentre en la práctica? Expliquese el razonamiento.

Tabla 10.7. ANOVA para los datos del Ejemplo 10.3.2. La hipótesis H_0: $\sigma^2_{Tr} = 0$ es rechazada con $0.05 < P < 0.10$

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	2	110.6	55.3	3
Error	27	497.7	18.43	
Total	29	608.3		

10.4. BLOQUES COMPLETOS ALEATORIZADOS

El proceso que se presenta en esta sección es una extensión del contraste T para datos emparejados, con el fin de comparar medias de dos poblaciones normales (véase Capítulo 9). Recuérdese que la finalidad del emparejamiento es reducir al mínimo el efecto de alguna variable extraña (una variable que no está siendo estudiada en el experimento), emparejando unidades experimentales semejantes con respecto a esta variable. Cada miembro del par recibe un tratamiento distinto, y cualquier tipo de diferencia en la respuesta se atribuye a los efectos del tratamiento, puesto que el efecto de la variable extraña ha sido neutralizado por el emparejamiento. Éste se describe en la Figura 10.3.

Cuando queremos comparar medias de k poblaciones en presencia de una variable extraña, se utiliza un procedimiento conocido como *de bloques.* Un *bloque* es una colección de k (en lugar de dos) unidades experimentales tan parecidas como sea posible con respecto a la variable extraña. Se asigna aleatoriamente cada tratamiento a una unidad dentro de cada bloque. Puesto que, una vez más, se ha neutralizado entre los tratamientos el efecto de la variable extraña emparejando unidades experimentales equivalentes, cualquier diferencia en las respuestas es atribuible a los efectos del tratamiento. En la Figura 10.4 se describe el procedimientp de bloques.

El diseño presentado aquí se llama diseño de *bloque completo aleatorizado* con efectos fijos. La palabra *bloque* se refiere al hecho de que se ha agrupado a las unidades experimentales en función de alguna variable extraña; *aleatorizado* se refiere al hecho de que los tratamientos se asignan aleatoriamente dentro de los bloques, y decir que el diseño es *completo* implica que se utiliza cada tratamiento exactamente una vez dentro de cada bloque. El término efectos fijos se aplica a ambos, bloques y tratamientos. Es decir, se supone que ni los bloques ni los

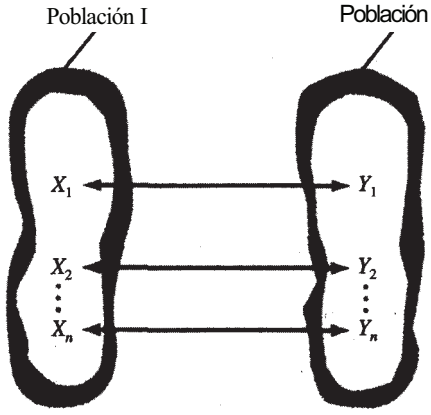


Figura 10.3. Datos emparejados. ¿Es $\mu_1 = \mu_2$?

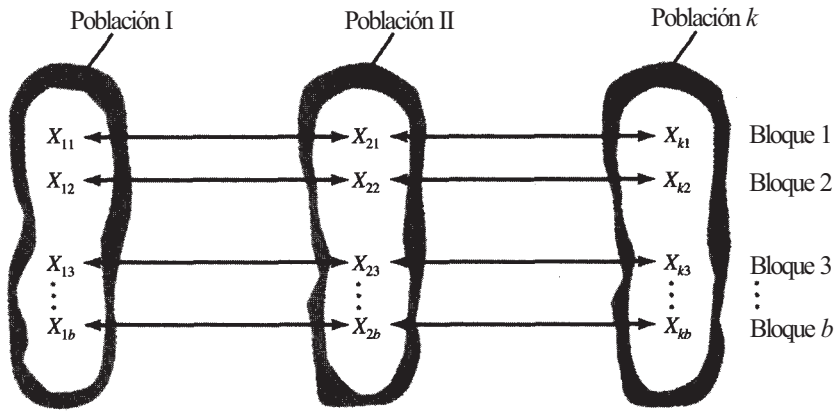


Figura 10.4. Datos asociados. ¿Es $\mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{k\cdot}$?

tratamientos se eligen aleatoriamente. Cualquier inferencia que se haga se aplica solamente) a los k tratamientos y a los b bloques utilizados. La hipótesis nula de interés principal es:

$$H_0: \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{k\cdot}$$

donde $\mu_{i\cdot}$ representa la media del i-ésimo tratamiento.

Ejemplo 10.4.1. Se realiza un experimento para comparar la energía que se requiere para llevar a cabo tres actividades físicas: correr, pasear y montar en bicicleta. La variable de interés es X, número de kilocalorías consumidas por kilómetro recorrido. Se piensa que las diferencias metabólicas entre los individuos pueden afectar al número de kilocalorías requeridas para llevar a cabo una determinada actividad, y se pretende controlar esta variable extraña. Para hacerlo, se seleccionan ocho individuos. Se le pide a cada uno que corra, camine y recorra en bicicleta una distancia medida, y se determina para cada individuo el número de kilocalorías consumidas por kilómetro durante cada actividad. Las actividades se realizan en orden aleatorio, con tiempo de recuperación entre una y otra. Cada individuo es utilizado como un bloque. Cada actividad se monitoriza exactamente una vez para cada individuo y de este modo se completa el diseño. Cualquier diferencia en el número medio de kilocalorías consumidas se atribuirá a diferencias entre las actividades mismas, puesto que se ha neutralizado el efecto de las diferencias individuales por medio de la construcción de bloques. La hipótesis nula de interés es:

$$H_0: \mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$$

donde $\mu_{1\cdot}$, $\mu_{2\cdot}$, $\mu_{3\cdot}$ designan el número medio de kilocalorías consumidas por kilómetro mientras se corre, se pasea, o se monta en bicicleta, respectivamente.

Formato de los datos y notación

En la página siguiente se da el formato en el que se registran convenientemente los datos de un diseño de bloque completo aleatorizado.

Obsérvese que b designa el número de bloques utilizados en el experimento y el número de observaciones por tratamiento; k designa el número de tratamientos que están siendo investigados y el número de observaciones por bloque; $N = kb$ designa el número total de respuestas. La variable X_{ij}, $i = 1, 2, \ldots, k$ y $j = 1, 2, \ldots, b$, es una variable aleatoria que designa la

Disposición de datos en el diseño de bloque completo aleatorizado

	Tratamiento				
Bloque	**1**	**2**	**3**		***k***
1	X_{11}	X_{21}	X_{31}	$\cdots$	X_{k1}
2	X_{12}	X_{22}	X_{32}	$\cdots$	X_{k2}
3	X_{13}	X_{23}	X_{33}	$\cdots$	X_{k3}
	.....				
b	X_{1b}	X_{2b}	X_{3b}	$\cdots$	X_{kb}

respuesta al i-ésimo tratamiento en el j-ésimo bloque. Al utilizar datos muestrales para comparar medias poblacionales, se requieren los siguientes estadísticos muestrales:

$$T_{i\cdot} = \sum_{j=1}^{b} X_{ij} = \text{suma total de las respuestas al i-ésimo tratamiento, } i = 1, 2, \ldots, k$$

$$\bar{X}_{i\cdot} = \frac{T_{i\cdot}}{b} = \text{media muestral del i-ésimo tratamiento, i = 1, 2, ..., } k$$

$$T_{\cdot j} = \sum_{i=1}^{k} X_{ij} = \text{suma total de las respuestas en el j-ésimo bloque, y = 1, 2, ..., } b$$

$$\bar{X}_{\cdot j} = \frac{T_{\cdot j}}{k} = \text{media muestral para el j-ésimo bloque, j = 1, 2,..., } b$$

$$T_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{b} X_{ij} = \sum_{i=1}^{k} T_{i\cdot} = \sum_{j=1}^{b} T_{\cdot j} = \text{suma total de las respuestas}$$

$$\bar{X}_{\cdot\cdot} = \frac{T_{\cdot\cdot}}{N} = \text{media muestral de todas las respuestas}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{b} X_{ij}^2 = \text{suma de cuadrados de cada respuesta}$$

En el Ejemplo 10.4.2, se desarrolla el cálculo de estos estadísticos.

Ejemplo 10.4.2. Al realizarse el experimento del Ejemplo 10.4.1, se obtuvieron como resultado los datos recogidos en la Tabla 10.8 relativos al número de kilocalorías consumidas por kilómetro por cada uno de los individuos, en cada una de las tres actividades. En los márgenes de la tabla se detallan totales por tratamiento y medias, totales por bloque y medias, y el total global y media.

Contraste de H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$.

Para escribir el modelo para el diseño de bloque completo aleatorizado con efectos fijos, se recurre a la siguiente notación:

μ = efecto medio total
μ_i = media del i-ésimo tratamiento, $i = 1, 2, \ldots, k$

Tabla 10.8*

Bloque	1 (corriendo)	2 (caminando)	3 (pedaleando)	Total bloque	Media bloque
1	1.4	1.1	0.7	3.2 ($T_{\cdot 1}$)	1.07 ($\bar{X}_{\cdot 1}$)
2	1.5	1.2	0.8	3.5 ($T_{\cdot 2}$)	1.17 ($\bar{X}_{\cdot 2}$)
3	1.8	1.3	0.7	3.8 ($T_{\cdot 3}$)	1.27 ($\bar{X}_{\cdot 3}$)
4	1.7	1.3	0.8	3.8 ($T_{\cdot 4}$)	1.27 ($\bar{X}_{\cdot 4}$)
5	1.6	0.7	0.1	2.4 ($T_{\cdot 5}$)	0.80 ($\bar{X}_{\cdot 5}$)
6	1.5	1.2	0.7	3.4 ($T_{\cdot 6}$)	1.13 ($\bar{X}_{\cdot 6}$)
7	1.7	1.1	0.4	3.2 ($T_{\cdot 7}$)	1.07 ($\bar{X}_{\cdot 7}$)
8	2.0	1.3	0.6	3.9 ($T_{\cdot 8}$)	1.30 ($\bar{X}_{\cdot 8}$)
Total tratamiento	13.2 ($T_{1\cdot}$)	9.2 ($T_{2\cdot}$)	4.8 (TV)	27.2 ($T_{\cdot\cdot}$)	
Media tratamiento	1.65 ($\bar{X}_{1\cdot}$)	1.15 ($\bar{X}_{2\cdot}$)	0.6 ($\bar{X}_{3\cdot}$)	1.13 ($\bar{X}_{\cdot\cdot}$)	

* Para estos datos $\sum_{i=1}^{3}\sum_{j=1}^{8} X_{ij}^2 = 36.18$.

$\mu_{\cdot j}$ = media del j-ésimo bloque, $j = 1, 2, ..., b$

μ_{ij} = media para el i-ésimo tratamiento y el j-ésimo bloque

$\tau_i = \mu_{i\cdot} - \mu$ = efecto debido al hecho de que la unidad experimental recibe el i-ésimo tratamiento

$\beta_j = \mu_{\cdot j} - \mu$ = efecto debido al hecho de que la unidad experimental está en el j-ésimo bloque

$E_{ij} = X_{ij} - \mu_{ij}$ = error residual o aleatorio

Utilizando esta notación podemos escribir el modelo de la siguiente forma:

Modelo

$$X_{ij} = \mu + \tau_i + \beta_j + E_{ij} \qquad \begin{matrix} i = 1, 2, ..., k \\ j = 1, 2, ..., b \end{matrix}$$

La fórmula expresa simbólicamente la idea de que cada observación puede subdividirse en cuatro componentes reconocibles: un efecto medio global μ, un efecto tratamiento τ_i, un efecto bloque β_j y una desviación aleatoria debida a causas desconocidas E_{ij}. Establecemos los siguientes supuestos:

Supuestos del modelo

1. Las $k \cdot b$ observaciones constituyen muestras aleatorias independientes, cada una de tamaño 1, *de* $k.b$ poblaciones con medias μ_{ij}, $i = 1, 2, ..., k$ y $j = 1, 2, ..., b$.
2. Cada una de las $k \cdot b$ poblaciones es normal.
3. Cada una de las $k \cdot b$ poblaciones tiene la misma varianza, σ^2.
4. Los efectos bloque y tratamiento son aditivos; es decir, no hay interacción entre los bloques y los tratamientos.

Los supuestos 1 a 3 son idénticos a los que se hicieron en el modelo de clasificación simple, salvo que, ahora, se están examinando $k \cdot b$, en vez de k, poblaciones. El cuarto supuesto es nuevo y debemos examinarlo más detenidamente. En resumen, decir que los efectos bloque y tratamiento son aditivos significa que los tratamientos tienen un comportamiento consistente a través de los bloques y que los bloques tienen un comportamiento consistente a través de los tratamientos. Matemáticamente, esto significa que la diferencia de los valores medios para dos tratamientos cualesquiera es la misma en todo un bloque, y que la diferencia en las medias para dos bloques cualesquiera es la misma para cada tratamiento. Si no es éste el caso, decimos que hay interacción entre bloques y tratamientos. Algunos ejemplos numéricos nos ayudarán a aclarar este concepto.

Ejemplo 10.4.3. Se han desarrollado tres programas para ayudar a que los pacientes que han sufrido un ataque cardíaco por primera vez se adapten física y psicológicamente a su situación. La variable de interés es el tiempo, en meses, necesario para que el paciente pueda reiniciar una vida activa. Puesto que se piensa que los varones pueden reaccionar a la enfermedad de manera distinta a como lo harían las mujeres, se controla esta variable mediante bloques. De esta forma, estamos tratando con $k \cdot b = 3 \cdot 2 = 6$ poblaciones normales, supuesta cada una con la misma varianza. Admitamos que las medias para estas seis poblaciones sean:

	Tratamiento		
Bloque	*A*	*B*	*C*
Varones	$\mu_{11} = 4$	$\mu_{21} = 5$	$\mu_{31} = 7$
Mujeres	$\mu_{12} = 3$	$\mu_{22} = 4$	$\mu_{32} = 6$

Obsérvese que la media para el tratamiento A es 1 unidad menos que para el B y 3 menos que para el C en cada bloque; la media para el tratamiento *B* es 2 unidades menos que para el *C* en cada bloque. Es decir, los tratamientos tienen un comportamiento consistente a través de los bloques. Análogamente, la media para el bloque 1 (varones) siempre supera a la media para el bloque 2 (mujeres) en 1, con independencia del tratamiento implicado. Es decir, los bloques tienen un comportamiento consistente a través de los tratamientos. Cuando esto sucede decimos que los efectos bloque y tratamiento son aditivos, o que *no* hay interacción entre tratamientos y bloques. Esta idea se esquematiza gráficamente en la Figura 10. 5.

La Figura 10.5 presenta la gráfica de las medias de los tratamientos recogidas en la tabla precedente. Cuando no existe interacción, los segmentos lineales que unen dos medias cualesquiera serán paralelos a través de los bloques. En términos prácticos, esto significa que es posible hacer consideraciones generales relativas a los tratamientos sin tener que especificar el bloque implicado. Por ejemplo, es correcto decir que el tratamiento *A* es superior al *B* y al C en que da lugar a un menor tiempo medio de recuperación.

Ejemplo 10.4.4. Consideremos el Ejemplo 10.4.3 con estas medias poblacionales:

	Tratamiento		
Bloque	*A*	*B*	*C*
Varones	$\mu_{11} = 4$	$\mu_{21} = 7$	$\mu_{31} = 9$
Mujeres	$\mu_{12} = 3$	$\mu_{22} = 6$	$\mu_{32} = 2$

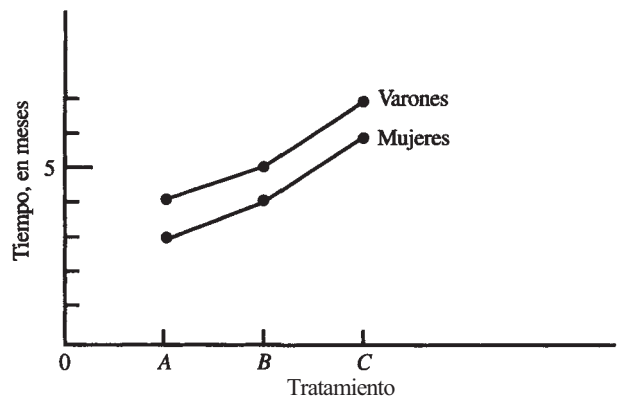


Figura 10.5. No hay interacción: los segmentos son paralelos.

Esta tabla no es aditiva. Para comprobarlo, obsérvese que la media del tratamiento A es 5 unidades menor que la del tratamiento *C* en el bloque 1 (varones), pero 1 más que la del *C* en el bloque 2. Esto es, los tratamientos se comportan de manera diferente en los distintos bloques. En este caso, decimos que los bloques y los tratamientos interactúan.

En la Figura 10.6 se recogen las gráficas para esta tabla. Puesto que no todos los segmentos lineales son paralelos, hay interacción entre bloques y tratamientos. En términos prácticos, esto significa que debemos ser muy cuidadosos cuando hagamos declaraciones relativas a los tratamientos, porque el bloque implicado es también importante. Por ejemplo, no es muy correcto decir que el tratamiento A es superior a los tratamientos *By C* porque da lugar a un menor tiempo medio de recuperación. Esta afirmación es cierta para el bloque 1 (varones), pero no para el bloque 2 (mujeres).

Para obtener la identidad *suma de cuadrados* para este modelo, observemos que puede demostrarse que la aditividad implica que $\mu_{ij} = \mu + (\mu_{i\cdot} - \mu) + (\mu_{\cdot j} - \mu)$. De este modo, el modelo teórico se puede escribir en la forma:

$$X_{ij} - \mu \equiv (\mu_{i\cdot} - \mu) + (\mu_{\cdot j} - \mu) + \{X_{ij} - [\mu + (\mu_{i\cdot} - \mu) + (\mu_{\cdot j} - \mu)]\}$$

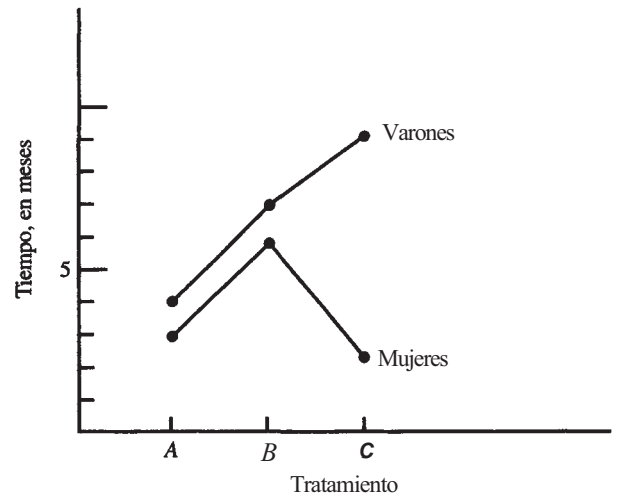


Figura 10.6. Existe interacción: los segmentos no son paralelos.

Sustituyendo cada una de las medias teóricas μ, $\mu_{i\cdot}$, $\mu_{\cdot j}$, por sus estimadores respectivos $\bar{X}_{\cdot\cdot}$, $\bar{X}_{i\cdot}$, $\bar{X}_{\cdot j}$, obtenemos la siguiente identidad:

$$X_{ij} - \bar{X}_{\cdot\cdot} \equiv (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) + (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}) + \{X_{ij} - [\bar{X}_{\cdot\cdot} + (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) + (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})]\}$$

Si cada miembro de esta identidad se eleva al cuadrado y se suma después sobre todos los valores posibles de *i* y j, se obtiene la siguiente identidad suma de cuadrados para el diseño de bloque completo aleatorizado.

Identidad suma de cuadrados

$$\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{b}(X_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{b}(X_{\cdot j} - \bar{X}_{\cdot\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2$$

La interpretación práctica de cada componente es similar a la del modelo de clasificación simple. En particular,

$\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{\cdot\cdot})^2$ = medida de la variabilidad total en los datos
= SS_{Total}

$\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$ = medida de la variabilidad en los datos atribuible a la utilización de diferentes tratamientos
= suma de cuadrados de los tratamientos
= SS_{TR}

$\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2$ = medida de la variabilidad en los datos atribuible a la utilización de bloques diferentes
= suma de cuadrados de los bloques
= $SS_{Bloques}$

$\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2$ = medida de la variabilidad en los datos debida a factores aleatorios
= suma de cuadrados residual, o error
= SS_E

Utilizando esta notación, podemos escribir la identidad suma de cuadrados como

$$SS_{Total} = SS_{Tr} + SS_{Bloques} + SS_E$$

Se puede contrastar la hipótesis nula

H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{k\cdot}$ (no hay diferencia en las medias de los tratamientos)

utilizando los siguientes cuadrados medios.

$$MS_{Tr} = \frac{SS_{Tr}}{k-1} = \text{cuadrado medio de los tratamientos}$$

$$MS_{\text{Bloques}} = \frac{SS_{\text{Bloques}}}{b-1} = \text{cuadrado medio de los bloques}$$

$$MS_E = \frac{SS_E}{(k-1)(b-1)} = \text{cuadrado medio del error}$$

Los valores esperados para estos cuadrados medios son

$$E[MS_{\text{Tr}}] = \sigma^2 + \frac{b}{k-1}\sum_{i=1}^{k}(\mu_{i\cdot} - \mu)^2$$

$$E[MS_E] = \sigma^2$$

Para contrastar H_0, la hipótesis de igualdad de medias en los tratamientos, se utiliza el cociente

$$\boxed{\frac{MS_{\text{Tr}}}{MS_E} = F_{k-1,\ (k-1)(b-1)}} \qquad \text{(Estadístico del contrastó)}$$

Si H_0 es cierta, este cociente tomará un valor próximo a 1, ya que, en este caso, $\sum_{i=1}^{k}(\mu_{i\cdot} - \mu)^2 = 0$ y tanto MS_{TR} como MS_E están estimando σ^2. Si H_0 no es cierta, entonces el valor de este estadístico será mayor que 1. Las fórmulas de cálculo utilizadas para evaluar los estadísticos son similares a las del modelo de clasificación simple y se muestran en la Tabla 10.9.

Ejemplo 10.4.5. Se necesitan los siguientes datos estadísticos del Ejemplo 10.4.2 para continuar el análisis de la energía requerida para correr, andar y montar en bicicleta:

$k = 3$	$T_1 = 3.2$	$T_6 = 3.4$
$b = 8$	$T_2 = 3.5$	$T_7 = 3.2$
$N = 24$	$T_3 = 3.8$	$T_8 = 3.9$
$T_{\cdot\cdot} = 27.2$	$T_4 = 3.8$	
	$T_{\cdot} = 2.4$	

Tabla 10.9. Análisis de la varianza: diseño de bloque completo aleatorizado con efectos fijos

Fuente de la variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	Cuadrado medio esperado	Cociente F
Tratamiento	k- 1	$\sum_{i=1}^{k}\frac{T_{i\cdot}^2}{b} - \frac{T_{\cdot\cdot}^2}{N}$	$\frac{SS_{\text{Tr}}}{k-1}$	$\sigma^2 + \frac{b}{k-1}\sum_{i=1}^{k}(\mu_i - \mu)^2$	$\frac{MS_{\text{Tr}}}{MS_E}$
Bloque	*b*- 1	$\sum_{j=1}^{b}\frac{T_{\cdot j}^2}{k} - \frac{T_{\cdot\cdot}^2}{N}$	$\frac{SS_{\text{Bloques}}}{b-1}$		
Error	$(k-1)(b-1)$	$SS_{\text{Total}} - SS_{\text{Tr}} - SS_{\text{Bloques}}$	$\frac{SS_E}{(k-1)(b-1)}$	σ^2	
Total	*kb*- 1	$\sum_{i=1}^{k}\sum_{j=1}^{b}X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{N}$			

$$\sum_{i=1}^{3}\sum_{j=1}^{8} X_{ij}^2 = 36.18$$

$T_{1.} = 13.2$ (corriendo)
$T_{2.} = 9.2$ (andando)
$T_{3.} = 4.8$ (pedaleando)

El ANOVA se muestra en la Tabla 10. $P = P[F_{2,14} \geq 78.75] < 0.01$. Ya que esta probabilidad es pequeña, rechazamos

$$H_0: \mu_{1.} = \mu_{2.} = \mu_{3.}$$ (no hay diferencias en la energía media requerida)

y concluimos que hay diferencias en las necesidades de energía para las tres actividades.

Si se eligen los bloques aleatoriamente, como probablemente sea el caso del Ejemplo 10.4.1, se dice que el modelo es mixto. El efecto de bloque en el modelo se considera que es una variable aleatoria y se escribe como B_j en lugar de β_j El análisis de los datos no se ve afectado.

Efectividad de la construcción de bloques

Puesto que la construcción de bloques se ha diseñado para controlar el efecto de una variable extraña, la pregunta natural a formular es: ¿ha tenido éxito la construcción de bloques? En caso afirmativo, $SS_{Bloques}$ explicaría una parte sustancial de la suma total de los cuadrados. Esto, a su vez, reduce SS_E, aumentando el valor del cociente F utilizado para contrastar la igualdad de medias de los tratamientos y posibilitando que se rechace H_0,. Se mejorará la potencia del contraste. Obsérvese que el número de grados de libertad para el error en el diseño por clasificación de una vía es $N- k$; en el diseño por bloques completos aleatorizados es más pequeño que éste, es decir $(k - 1)(b - 1) = (N - k) - (b - 1)$. En la Tabla EX del Apéndice B, se puede observar que, a medida que disminuye el número de grados de libertad asociado con el denominador de F, el valor F de la tabla aumenta. Ello implica que si se realiza una construcción innecesaria de bloques, pagamos un precio por este error. Es decir, disminuye el número de grados de libertad para el error, aumenta el punto crítico para contrastar H_0. y es más difícil rechazar H_0. La potencia del contraste será menor. Queda claro que la construcción de bloques puede ayudar cuando sea pertinente, pero debe evitarse la construcción indiscriminada de bloques.

Cuando un experimento se realiza por primera vez, la intuición basada en el conocimiento de la materia es la única guía para decidir si realizar o no la construcción de bloques. Una vez realizado el experimento inicial, puede efectuarse una valoración de la efectividad de la construcción de bloques de forma que puedan diseñarse eficazmente estudios futuros. Parece

Tabla 10.10. ANOVA para los datos del Ejemplo 10.4.5

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	2	4.41	2.205	78.75($F_{2,14}$)
Bloque	7	0.55	0.079	
Error	14	0.39	0.028	
Total	23	5.35		

razonable sugerir que, si las medias de los bloques son iguales, la construcción de bloques es innecesaria; de lo contrario, la construcción de bloques es útil. Sin embargo, no existe una forma válida conocida para contrastar la hipótesis nula de la igualdad de medias de los bloques. Un método utilizado para investigar la efectividad de la construcción de bloques es estimar la eficacia relativa *(RE)* del diseño de bloque completo aleatorizado comparada con la del diseño completo aleatorizado de la Sección 10.1. El desarrollo teórico de la noción de eficacia relativa escapa al alcance de este texto pero puede hallarse en [9] y [11]. La eficacia relativa es un número positivo que puede interpretarse como la relación del número de observaciones por tratamiento necesario para que los dos diseños sean equivalentes. Por ejemplo, si $RE = 3$, entonces el diseño completo aleatorizado requiere tres veces tantas observaciones como el diseño de bloque completo aleatorizado para producir un contraste con las mismas características; en este caso, la construcción de bloques es deseable. Si $RE = 0.5$, entonces la construcción de bloques no es deseable puesto que el diseño completo aleatorizado puede servir igual que el diseño de bloque aleatorizado utilizando la mitad de las observaciones. Si $RE = 1$, entonces los diseños son equivalentes cuando los tamaños muestrales son idénticos.

¿Podemos estimar rápidamente la eficacia relativa a partir de nuestro análisis de varianza original? Afortunadamente, existe una forma fácil de hacerlo. Se ha descubierto [10] que existe una relación lineal entre $\widehat{RE}$ y el cociente

$$\frac{MS_{\text{Bloques}}}{MS_E}$$

donde

$$MS_{\text{Bloques}} = \text{cuadrado medio de los bloques} = \frac{SS_{\text{Bloques}}}{b-1}$$

Esta relación viene dada por

$$\widehat{RE} = c + (1-c)\frac{MS_{\text{Bloques}}}{MS_E}$$

donde $c = b(k-1)/(bk-1)$. Es fácil demostrar que

$\widehat{RE} = 1$ si y sólo si $MS_{\text{Bloques}}/MS_E = 1$

$\widehat{RE} < 1$ si y sólo si $MS_{Bloques}/MS_E < 1$

$\widehat{RE} > 1$ si y sólo si $MS_{Bloques}/MS_E > 1$

Así pues, para emitir un juicio sobre si construir bloques en un experimento en particular ha ayudado o ha perjudicado, podemos utilizar la información disponible de nuestro ANOVA para hallar el valor de MS_{Bloques}/MS_E. A continuación, estimamos la eficacia relativa y decidimos, a partir de consideraciones prácticas como el tiempo, el coste y el esfuerzo requeridos para construir los bloques, si la construcción de bloques vale la pena. A medida que se gane experiencia, será realmente innecesario calcular $\widehat{RE}$. Sólo necesitamos considerar el valor observado de MS_{Bloqae}/MS_E. Los valores considerablemente mayores que 1 indican que la construcción de bloques ha sido provechosa; los valores próximos a 1 indican que la construcción de bloques ni ha ayudado ni ha perjudicado; los valores algo inferiores a 1 indican que la construcción de bloques no ha sido útil. En este último caso, es preferible que en futuros estudios se realice un diseño completamente aleatorizado.

Se observará que algunos textos incluyen un «contraste» para bloques basado en el estadístico MS_{Bloques}/MS_E. Sin embargo, debido a la forma en la que se obtiene la aleatorización en el diseño de bloque completo aleatorizado, el contraste es inadecuado.

Como ejemplo, continuemos el análisis de los datos del Ejemplo 10.4.5.

Ejemplo 10.4.6. En la Tabla 10.10, del análisis de la varianza, observamos que:

$$MS_{Bloques} = 0.079 \text{ y } MS_E = 0.028$$

Para estos datos, $b = 8$, $k = 3$, y

$$c = \frac{b(k-1)}{bk-1}$$
$$= \frac{8(2)}{23}$$
$$= 0.696$$

La eficacia relativa estimada es:

$$\widehat{RE} = c + (1-c)\frac{MS_{\text{Bloques}}}{MS_E}$$
$$= 0.696 + 0.304\,\frac{0.079}{0.028}$$
$$= 1.55$$

Puesto que $\widehat{RE} > 1$, se desprende que la construcción de bloques ha sido beneficiosa en este caso. El diseño completamente aleatorizado requerirá 1.55 veces tantas observaciones como el diseño de bloque completo aleatorizado para producir un contraste de igual potencia. Obsérvese también que, en este caso,

$$\frac{MS_{\text{Bloques}}}{MS_E} = \frac{0.079}{0.028} = 2.82$$

es superior a 1, una indicación de que la construcción de bloques es útil.

Comparaciones por parejas y múltiples

Al igual que en la clasificación de una vía del diseño completamente aleatorizado, las comparaciones por parejas pueden realizarse ejecutando $\binom{k}{2}$ contraste T de tipo Bonferroni con α elegido cuidadosamente de forma que α' se mantenga bajo control. En este caso, los contrastes T realizados son contrastes T dos a dos. Téngase en cuenta el hecho de que este método sólo es factible cuando k es bastante pequeño. Ello se debe al hecho de que los valores grandes de k fuerzan que α sea *extremadamente* pequeño, dando como resultado un contraste con muy poca potencia.

También disponemos de un contraste de Duncan de rango múltiple. El contraste se efectúa como se describió en la Sección 10.2 con:

$$SSR_p = r_p\sqrt{\frac{MS_E}{b}}$$

Obsérvese que en este diseño los tamaños muestrales son iguales.

Nota sobre los cálculos

La mayoría de paquetes informáticos comerciales incluyen un procedimiento para analizar un diseño de bloque completo aleatorizado. En el Ejemplo 10.4.7 tratamos la interpretación del *printout* del SAS obtenido para los datos del Ejemplo 10.4.5. El código utilizado para realizar este *printout* aparece en la sección Herramientas computacionales al final de este capítulo.

Ejemplo 10.4.7. El *printout* del SAS (que se muestra a continuación) no es exactamente como la tabla ANOVA hecha con la calculadora manual. Sin embargo, los valores obtenidas en la calculadora manual se hallan en el *printout* salvo por las diferencias de redondeo. Los números de interés son:

① Grados de libertad para los tratamientos
② Grados de libertad para los bloques
③ Grados de libertad para el error
④ Grados de libertad totales
⑤ Suma total de cuadrados
⑥ Suma de los cuadrados del error
⑦ Media de los cuadrados del error
⑧ Suma de cuadrados de los bloques
⑨ Suma de cuadrados de los tratamientos

ANOVA para los datos del Ejemplo 10.4.7

SAS

General Linear Models Procedure

Dependent Variable: KILOCAL

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Model	9	4.96666667	0.55185185	19.98	0.0001
Error	14 ③	0.38666667 ⑥	0.02761905 ⑦		
Corrected Total	23 ④	5.35333333 ⑤			

R-Square	C.V.	Root MSE	KILOCAL Mean
0.927771	14.66381	0.166190	1.13333333

Source	DF	Type I SS	Mean Square	*F* Value	Pr > *F*
BLOCK	7 ②	0.55333333 ⑧	0.07904762	2.86 ⑫	0.0446
TRTMENT	2 ①	4.41333333 ⑨	2.20666667	79.90 ⑩	0.0001 ⑪

Source	DF	Type III SS	Mean Square	*F* Value	Pr > *F*
BLOCK	7	0.55333333	0.07904762	2.86	0.0446
TRTMENT	2	4.41333333	2.20666667	79.90	0.0001

SAS

General Linear Models Procedure

Duncan's Multiple Range Test for variable: KILOCAL

NOTE: This test controls the type I comparisonwise a error rate, not the experimentwise error rate.

Alpha = 0.05 DF = 14 MSE = 0.027619

Number of Means 2 3

Critical Range 0.178 0.187

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	TRTMENT
A	1.6500	8	1
⑬ B	1.1500	8	2
C	0.6000	8	3

SAS

General Linear Models Procedure

Bonferroni (Dunn) Ttests for variable: KILOCAL

NOTE: This test controls the type I experimentawise error rate, but generally has a higher type II error rate than REGWQ.

Alpha = 0.05 DF = 14 MSE=0.027619

Critical Value of T = 2.72

Minimum Significant Difference = 0.2258

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	TRTMENT
A	1.6500	8	1
⑭ B	1.1500	8	2
C	0.6000	8	3

Obsérvese que el cociente F utilizado para contrastar H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$ se muestra e ⑩ y su valor P viene dado en ⑪. La media de los cuadrados de los bloques utilizada al estimar la eficacia relativa se muestra en ⑫. Los resultados del contraste de Duncan de rango múltiple se dan en ®. Nótese que ninguna de las medias es idéntica. Los resultados del contraste T de Bonferroni concuerdan con los resultados de Duncan y se muestran en ⑭.

EJECICIOS 10.4

Nota: Si es posible, estos problemas deberían hacerse utilizando un ordenador.

1. Para cada tabla de medias poblacionales, decidir si hay interacción entre bloques y tratamientos.

a)

	Tratamiento			
Bloque	***A***	***B***	***C***	***D***
1	**1**	**3**	**4**	**0**
2	**4**	**6**	**7**	**3**
3	**2**	**4**	**5**	**1**

b)

	Tratamiento			
Bloque	***A***	***B***	***C***	***D***
1	**1**	**3**	**0**	**0**
2	**4**	**6**	**5**	**3**
3	**2**	**4**	**5**	**1**

c)

	Tratamiento			
Bloque	***A***	***B***	***C***	***D***
1	**1**	**3**	**4**	**0**
2	**4**	**5**	**7**	**3**
3	**2**	**4**	**5**	**1**

2. El abeto Sitka es, desde el punto de vista económico, el árbol de bosque más importante en el Reino Unido. Sin embargo, posee una regeneración natural escasa, como consecuencia de que los buenos años para la producción de semillas son infrecuentes. Es necesario aumentar la producción de semillas. Se proponen cuatro tratamientos hormonales. Puesto que árboles diferentes tienen distintas características naturales de reproducción, se controla el efecto de las diferencias entre árboles mediante bloques. En el experimento se utilizan diez árboles. Dentro de cada árbol se seleccionan cuatro ramas semejantes. Cada rama recibe exactamente uno de los cuatro tratamientos, siendo éstos aleatoriamente asignados. De este modo, cada árbol constituye un bloque completo. Lo que se mide es el número de semillas producidas por rama. Supongamos que esta variable, si bien discreta, está de manera aproximada normalmente distribuida. Los datos se recogen en la Tabla 10.11.

Tabla 10.11

	Tratamiento					
Bloque (árbol)	*A*	*B*	*C*	*D*	**Total bloque**	**Media bloque**
1	89	59	20	51		
2	87	56	15	47		
3	84	52	14	45		
4	92	67	26	56		
5	95	70	28	60		
6	90	62	22	53		
7	89	60	19	51		
8	88	56	17	50		
9	82	50	14	45		
10	94	63	24	53		
Total tratamiento						
Media tratamiento						

a) Completar la Tabla 10.11 calculando los totales y las medias de los tratamientos muestrales, los totales y las medias de los bloques, y las medias y totales globales
b) Contrastar la hipótesis nula de igualdad de medias en los tratamientos.

3. Explicar el significado de cada una de estas eficacias relativas al evaluar la efectividad de la construcción de bloques.
   a) $RE = 2$
   b) $RE = 10$
   c) $RE = 0.25$
   d) $RE = 0.10$
   e) $RE = 1$
4. Estimar la eficacia relativa de los datos del Ejercicio 2. ¿Ha resultado útil la construcción de bloques? Explicarlo.
5. Se realiza un estudio del efecto de la luz sobre el crecimiento de los helechos. Puesto qué las plantas crecen con distinta velocidad a edades diferentes, se controla esta variable mediante bloques. En el estudio se utilizan cuatro plantas jóvenes (plantas crecidas en la oscuridad durante cuatro días) y cuatro plantas más viejas (plantas crecidas en la oscuridad durante doce días) produciéndose así dos bloques, cada uno de tamaño 4. Se investigan cuatro tratamientos de luz diferentes. Cada tratamiento se asigna aleatoriamente a una planta en cada bloque. Los tratamientos consisten en: exponer a cada planta a una única dosis de luz, ponerla de nuevo en la oscuridad y medir el área de sección transversal del extremo del helecho veinticuatro horas después de que se le administró la luz. Resultaron los siguientes datos (el área de la sección transversal viene dada en micrómetros cuadrados):

	Tratamiento (longitud de onda de la luz)			
Bloque (edad)	**420 nm**	**460 nm**	**600 nm**	**720 nm**
Joven	1017.6	929.0	939.8	1081.5
Adulto	854.7	689.9	841.5	797.4

a) Hallar los totales y medias por tratamiento, bloque y global.
b) Contrastar la hipótesis nula de igualdad de medias en los tratamientos.
c) Estimar la eficacia relativa y comentar la efectividad de la construcción de bloques.

6. Utilizar el contraste de Duncan de rango múltiple con $\alpha = 0.01$ para completar el análisis de los datos del Ejemplo 10.4.5.
7. Utilizar el contraste de Duncan de rango múltiple para completar el análisis de los datos del Ejercicio 2. Úsese $\alpha = 0.01$.
8. Se ha realizado un estudio sobre el efecto de las temporadas de caza del ciervo en los hábitos de éstos. Se seleccionaron cuatro sendas que se sabe que utilizan los ciervos. Antes de comenzar la temporada de caza, durante la temporada y al terminar la temporada, se determinó el promedio de huellas halladas por semana en un área específica de cada senda. Las sendas se trataron como bloques y se obtuvieron los datos siguientes:

Senda	Antes	Durante	Después
1	62.5	57.0	49.0
2	46.5	53.3	50.0
3	45.0	59.3	37.0
4	24.0	35.7	50.0

a) Contrastar H_0: $\mu_1 = \mu_2. = \mu_3..$
b) Si en el apartado *a* se encuentran diferencias, utilizar los contrastes *T* de Bonferroni con un nivel global α no mayor de 0.15, para señalar las diferencias existentes.
c) Estimar la eficacia relativa y comentar la efectividad de la construcción de bloques.
(Basado en un estudio realizado por Daniel Brown, Departamento de Biología, Universidad de Radford y Servicio de Consulting de la Universidad de Radford, octubre de 1990.)

9. Se lleva a cabo un estudio para averiguar el efecto que sobre la flexibilidad de los tobillos de los corredores produce la protección previa de los mismos. La flexibilidad se midió antes y después de la protección, y después de correr con los tobillos protegidos. A un grupo de individuos se les protegió los tobillos utilizando un método estándar; con el otro se utilizó una técnica de protección reforzada. Los sujetos constituyen los bloques. Puntuaciones altas indican mayor flexibilidad. Se obtuvieron los siguientes datos:

	Tratamiento (protección estándar)				**Tratamiento (protección estándar)**		
Bloque (individuo)	**Antes de la protección**	**Después de la protección**	**Después de correr**	**Bloque (individuo)**	**Antes de la protección**	**Después de la protección**	**Después de correr**
1	2.5	1.0	4.0	9	12.0	2.5	0.5
2	9.0	1.0	6.5	10	3.5	7.5	3.5
3	6.0	3.0	2.0	11	0.5	1.0	1.0
4	2.5	0.5	1.5	12	6.0	4.0	5.5
5	6.0	4.5	4.5	13	0.0	4.5	1.0
6	3.5	0.0	3.5	14	5.0	2.0	7.5
7	7.5	5.5	3.5	15	0.5	0.5	1.0
8	4.0	2.0	0.0	16	5.5	-0.5	0.5

	Tratamiento (protección reforzada)				Tratamiento (protección reforzada)		
Bloque (individuo)	Antes de la protección	Después de la protección	Después de correr	Bloque (individuo)	Antes de la protección	Después de la protección	Después de correr
1	6.5	2.5	4.5	9	-3.5	-3.5	0.0
2	4.5	2.5	6.0	10	2.5	2.0	-0.5
3	5.5	4.0	3.0	11	3.0	1.5	1.0
4	5.0	4.0	7.0	12	7.5	8.5	6.5
5	2.0	4.5	4.0	13	9.5	1.0	0.5
6	5.0	2.5	1.5	14	6.0	3.5	2.0
7	10.0	4.0	3.0	15	-0.5	0.5	0.5
8	8.5	0.0	10.5	16	5.5	-0.5	5.0

Analice cada uno de estos conjuntos de datos y comente las semejanzas y las diferencias que encuentre entre la protección estándar y la reforzada. (Basado en un estudio realizado por Jay Canterbury, Departamento de Educación Física, Universidad de Radford, 1997.)

10.5. EXPERIMENTOS FACTORIALES

En muchos experimentos, se investigan con detalle dos o más variables. Ninguna variable se considera extraña; todas tienen el mismo interés. Cuando esto sucede, el experimento se llama un *experimento factorial,* para resaltar el hecho de que el interés se centra en el efecto de dos o más factores sobre una medida de respuesta. Presentamos aquí la *clasificación de dos vías, diseño completamente aleatorio con efectos fijos.* De este modo, nos ocupamos de un modelo en el que se estudian dos factores, *A* y *B,* habiendo seleccionado el experimentador, a propósito, más que aleatoriamente, los niveles de cada factor. No se asocian las unidades experimentales semejantes.

Ejemplo 10.5.1. El *Mirogrex terrae-sanctae* es un pez comercializado semejante a la sardina que se encontró en el Mar de Galilea. Se realizó un estudio para determinar el efecto de la luz y la temperatura sobre el índice gonadosomático (GSI), que es una medida de crecimiento del ovario. Se utilizaron dos fotoperíodos: catorce horas de luz, diez horas de oscuridad y nueve horas de luz, quince horas de oscuridad; y dos niveles de temperatura, 16 y 27 °C. De este modo, el experimentador puede simular situaciones de verano e invierno en la región. Se trata de un experimento factorial con dos factores, luz y temperatura, que son investigados cada uno a dos niveles.

Formato de los datos y notación

En la página siguiente se da el formato en el que se registran adecuadamente los datos recogidos en un diseño de clasificación de dos vías.

Obsérvese que *a* indica el número de niveles del factor *A* utilizados en el experimento, *b* indica el número de niveles del factor *B,* y $a \cdot b$ es el número total de combinaciones de tratamientos, donde una combinación de tratamientos es un nivel del factor *A* aplicado en conjunto con un nivel de *B.* Suponemos que hay *n* observaciones para cada combinación de tratamientos. El número total de respuestas es $N = a \cdot b \cdot n$. La variable X_{iJk}, i = 1,2,.... a, j = 1, 2,…, *b,k=1,* 2,…, *n,* es una variable aleatoria que designa la respuesta de la k-ésima

Disposición de datos en la clasificación de dos vías

	Nivel del factor A				
Factor B	1	2	3	⋯	α
1	X_{111}	X_{211}	X_{311}	⋯	X_{a11}
	X_{112}	X_{212}	X_{312}	⋯	X_{a12}
	.....				
	X_{11n}	X_{21n}	X_{31n}	⋯	X_{a1n}
2	X_{121}	X_{221}	X_{321}	⋯	X_{a21}
	X_{122}	X_{222}	X_{322}	⋯	X_{a22}
	.....				
	X_{12n}	X_{22n}	X_{32n}	⋯	X_{a2n}
⋮					
b	X_{1b1}	X_{2b1}	X_{3b1}	⋯	X_{ab1}
	X_{1b2}	X_{2b2}	X_{3b2}	⋯	X_{ab2}
	.....				
	X_{1bn}	X_{2bn}	X_{3bn}	⋯	X_{abn}

unidad experimental al i-ésimo nivel del factor A y al j-ésimo nivel del factor *B*. Estos estadísticos muestrales son necesarios para analizar los datos. Recordemos que el punto indica el subíndice sobre el que se efectúa la suma.

$$T_{ij\cdot} = \sum_{k=1}^{n} X_{ijk} = \text{suma total de las respuestas al i-ésimo nivel del factor } A \text{ y al j-ésimo nivel del factor } B$$
$$= \text{suma total de las respuestas a la (i - j)-ésima combinación de tratamientos}$$

$$\bar{X}_{ij\cdot} = \frac{T_{ij\cdot}}{n} = \text{media muestral para la (i - j)-ésima combinación de tratamientos}$$

$$T_{i\cdot\cdot} = \sum_{j=1}^{b} T_{ij\cdot} = \text{suma total de las respuestas al i-ésimo nivel, i=1, 2,..., } \alpha, \text{ del factor } A$$

$$\bar{X}_{i\cdot\cdot} = \frac{T_{i\cdot\cdot}}{bn} = \text{media muestral para el i-ésimo nivel del factor } A$$

$$T_{\cdot j\cdot} = \sum_{i=1}^{a} T_{ij\cdot} = \text{suma total de las respuestas al j-ésimo nivel, } j = 1, 2,..., b, \text{ del factor } B$$

$$\bar{X}_{\cdot j\cdot} = \frac{T_{\cdot j\cdot}}{an} = \text{media muestral para el j-ésimo nivel del factor } B$$

$$T_{\cdot\cdot\cdot} = \sum_{i=1}^{a} T_{i\cdot\cdot} = \sum_{j=1}^{b} T_{\cdot j\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij\cdot} = \text{suma total de las respuestas}$$

$$\bar{X}_{...} = \frac{T_{...}}{abn} = \text{media muestral para todas las respuestas}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk}^2 = \text{suma de cuadrados de cada respuesta}$$

En el Ejemplo 10.5.2 se evalúan estos estadísticos.

Ejemplo 10.5.2. El experimento del Ejemplo 10.5.1 se realizó recogiendo 20 hembras en junio. A continuación, se dividió aleatoriamente el grupo en cuatro subgrupos, de tamaño 5 cada uno. Cada subgrupo recibió una de las cuatro posibles combinaciones de tratamientos. Después de tres meses se determinó el GSI para cada pez. Se obtuvieron los siguientes datos:

Factor B (temperatura)	Factor A **(fotoperíodo)**		
	9 horas	**14 horas**	**Total (factor *B*)**
27 °C	(no natural) 0.90 1.06 0.98 $T_{11\cdot} = 5.35$ 1.29 1.12 $\bar{X}_{11\cdot} = 1.07$	(verano simulado) 0.83 0.67 0.57 $T_{21\cdot} = 3.2$ 0.47 0.66 $\bar{X}_{21\cdot} = 0.64$	$T_{\cdot 1\cdot} = T_{11\cdot} + T_{21\cdot} = 8.55$ $\bar{X}_{\cdot 1\cdot} = 0.855$
16 °C	(invierno simulado) 1.30 2.88 2.42 $T_{12\cdot} = 12.20$ 2.66 2.94 $\bar{X}_{12\cdot} = 2.44$	(no natural) 1.01 1.52 1.02 $T_{22\cdot} = 6.5$ 1.32 1.63 $\bar{X}_{22\cdot} = 1.3$	$T_{\cdot 2\cdot} = T_{12\cdot} + T_{22\cdot} = 18.7$ $\bar{X}_{\cdot 2\cdot} = 1.87$
Total (factor *A*)	$T_{1\cdot\cdot} = T_{11\cdot} + T_{12\cdot} = 17.55$ $\bar{X}_{1\cdot\cdot} = 1.755$	$T_{2\cdot\cdot} = T_{21\cdot} + T_{22\cdot} = 9.7$ $\bar{X}_{2\cdot\cdot} = 0.97$	$T_{\cdot\cdot\cdot} = 27.25$ (total global) $\bar{X}_{\cdot\cdot\cdot} = 1.363$

Para estos datos, $\alpha = 2$, $b = 2$, $n = 5 y N = \alpha bn = 20$. También

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{5} X_{ijk}^2 = 48.26$$

Para escribir el modelo para el diseño se utiliza la siguiente notación:

μ = efecto medio global
$\mu_{i\cdot\cdot}$ = media para el i-ésimo nivel del factor *A*, $i = 1, 2, \ldots, \alpha$
$\mu_{\cdot j\cdot}$ = media para el j-ésimo nivel del factor *B*, $j = 1, 2, \ldots, b$
$\mu_{ij\cdot}$ = media para la *(i* - j)-ésima combinación de tratamientos
$\alpha_i = \mu_{i\cdot\cdot} - \mu$ = efecto debido al hecho de que la unidad experimental está en el i-ésimo nivel del factor *A*
$\beta_j = \mu_{\cdot j\cdot} - \mu$ = efecto debido al hecho de que la unidad experimental está en el j-ésimo nivel del factor *B*
$(\alpha\beta)_{ij} = \mu_{ij\cdot} - \mu_{i\cdot\cdot} + \mu_{\cdot j\cdot} + \mu$ = efecto de interacción entre el i-ésimo nivel del factor *A* y el j-ésimo nivel del factor *B*
$E_{ijk} = X_{ijk} - \mu_{ij\cdot}$ = error residual o aleatorio

Utilizando esta notación podemos expresar el modelo en la forma siguiente:

Modelo

$$X_{ijk} \equiv \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk} \quad \begin{matrix} i = 1, 2, ..., a \\ j = 1, 2, ..., b \\ k = 1, 2, ..., n \end{matrix}$$

El modelo expresa simbólicamente la idea de que cada observación puede dividirse en cinco componentes: un efecto medio global (μ), un efecto debido al factor *A* (α_i), un efecto debido al factor *B* (β_j), un efecto debido a la interacción $(\alpha\beta)_{ij}$ y una desviación aleatoria debida a orígenes inexplicados (E_{ijk}). Admitimos los siguientes supuestos:

Supuestos del modelo

1. Las observaciones para cada combinación de tratamientos constituyen muestras aleatorias independientes, cada una de tamaño *n,* de $a \cdot b$ poblaciones con medias μ_{ij}., $i = 1, 2,..., a, j = 1, 2, ..., b$.
2. Cada una de las $a \cdot b$ poblaciones es normal.
3. Cada una de las $a \cdot b$ poblaciones tiene la misma varianza, σ^2.

La identidad suma de cuadrados, obtenida reemplazando cada una de las medias teóricas μ, $\mu_{i\cdot\cdot}$, $\mu_{\cdot j\cdot}$, $\mu_{ij\cdot}$ por sus estimadores $\bar{X}_{\cdots}$, $\bar{X}_{i\cdot\cdot}$, $\bar{X}_{\cdot j\cdot}$, $\bar{X}_{ij\cdot}$, respectivamente, elevando al cuadrado y sumando sobre *i, j y k,* es la siguiente:

Identidad suma de cuadrados

$$SS_{\text{Total}} = SS_A + SS_B + SS_{AB} + SS_E$$

En esta identidad,

$$SS_{\text{Total}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \bar{X}_{\cdots})^2 = \text{medida de la variabilidad total}$$

$$SS_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{X}_{i\cdot\cdot} - \bar{X}_{\cdots})^2 = \text{medida de la variabilidad de los datos atribuible a la utilización de diferentes niveles del factor } A$$

$$SS_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{X}_{\cdot j\cdot} - \bar{X}_{\cdots})^2 = \text{medida de la variabilidad de los datos atribuible a la utilización de diferentes niveles del factor } B$$

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}_{\cdots})^2 = \text{medida de la variabilidad de los datos debida a la interacción entre niveles de los factores } A \text{ y } B$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \bar{X}_{ij\cdot})^2 = \text{medida de la variabilidad en los datos debida a orígenes aleatorios inexplicados}$$

Contraste de los efectos principales e interacción

La primera hipótesis nula a contrastar es la de no interacción. Matemáticamente, esta hipótesis es:

$$H_0: (\alpha\beta)_{ij} = 0 \qquad i = 1, 2, ..., a; \qquad j = 1, 2, ..., b$$

Si no se rechaza esta hipótesis, se continúa el análisis contrastando la hipótesis nula de que no hay diferencias entre los niveles del factor *A,*

$$H_0': \mu_{1\cdot\cdot} = \mu_{2\cdot\cdot} = \cdots = \mu_{a\cdot\cdot}$$

y la hipótesis nula de no diferencia entre niveles del factor *B,*

$$H_0'': \mu_{\cdot 1\cdot} = \mu_{\cdot 2\cdot} = \cdots = \mu_{\cdot b\cdot}$$

Sin embargo, si se rechaza la hipótesis nula de no interacción, entonces no contrastamos H_0 y H_0. En este caso, puesto que los niveles del factor *A* no se comportan consistentemente a través de los niveles del factor *B y* viceversa, contrastamos las diferencias entre los niveles del factor A para cada uno de los *b* niveles del factor *B* individualmente. Esto se realiza ejecutando *b* análisis individuales de clasificación de una vía. Contrastamos estas hipótesis:

$$H_0: \mu_{1j\cdot} = \mu_{2j\cdot} = \cdots = \mu_{aj\cdot} \qquad j = 1, 2, ..., b$$

Las fórmulas utilizadas para calcular SS_A, SS_B y SS_{Total} son similares a aquellas de los modelos previos:

$$SS_A = \sum_{i=1}^{a} \frac{T_{i\cdot\cdot}^2}{bn} - \frac{T_{\cdots}^2}{abn}$$

$$SS_B = \sum_{j=1}^{b} \frac{T_{\cdot j\cdot}^2}{an} - \frac{T_{\cdots}^2}{abn}$$

$$SS_{\text{Total}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk}^2 - \frac{T_{\cdots}^2}{abn}$$

La suma de cuadrados de las interacciones se halla calculando primero la suma de cuadrados de los tratamientos. Es la suma de cuadrados de los tratamientos usual que se obtendría si se analizasen las $a \cdot b$ combinaciones de los tratamientos como un diseño de clasificación de una vía o simple. Es decir,

$$SS_{\text{Tr}} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{T_{ij\cdot}^2}{n} - \frac{T_{\cdots}^2}{abn}$$

Puede demostrarse que $SS_{Tr} = SS_A + SS_B + SS_{AB}$. Ello nos permite calcular la interacción suma de cuadrados por sustracción:

$$SS_{AB} = SS_{Tr} - SS_A - SS_B$$

La suma de cuadrados del error puede obtenerse también por sustracción:

$$SS_E = SS_{Total} - SS_{Tr}$$

El formato del análisis de la varianza para este diseño se muestra en la Tabla 10.12.

Tabla 10.12. Análisis de la varianza: clasificación de dos vías, diseño completamente aleatorio con efectos fijos

Origen de la variación	**Grados de libertad**	**Suma de cuadrados**	**Cuadrado medio**	**Cuadrado medio esperado**	**Cociente *F***
Tratamiento	$ab-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\cdot}^2}{n}-\frac{T_{\cdots}^2}{abn}$	$\frac{SS_{\text{Tr}}}{ab-1}$	$\sigma^2+n\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(\mu_{ij\cdot}-\mu_{\cdots})^2}{ab-1}$	$\frac{MS_{\text{Tr}}}{MS_E}$
A	$a-1$	$\sum_{i=1}^{a}\frac{T_{i\cdot\cdot}^2}{bn}-\frac{T_{\cdots}^2}{abn}$	$\frac{SS_A}{a-1}$	$\sigma^2+nb\sum_{i=1}^{a}\frac{(\mu_{i\cdot\cdot}-\mu_{\cdots})^2}{a-1}$	$\frac{MS_A}{MS_E}$
B	$b-1$	$\sum_{j=1}^{b}\frac{T_{\cdot j\cdot}^2}{an}-\frac{T_{\cdots}^2}{abn}$	$\frac{SS_B}{b-1}$	$\sigma^2+na\sum_{j=1}^{b}\frac{(\mu_{\cdot j\cdot}-\mu_{\cdots})^2}{b-1}$	$\frac{MS_B}{MS_E}$
AB	$(a-1)(b-1)$	$SS_{\text{Tr}}-SS_{AA}-SS_B$	$\frac{SS_{AB}}{(a-1)(b-1)}$	$\sigma^2+n\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	$ab(n-1)$	$SS_{\text{Total}}-SS_{\text{Tr}}$	$\frac{SS_E}{ab(n-1)}$	σ^2	
Total	$abn-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}X_{ijk}^2-\frac{T_{\cdots}^2}{abn}$			

El primer cociente F a considerar en cualquier experimento es

$$F_{(a-1)(b-1),\,ab(n-1)}=\frac{MS_{AB}}{MS_E}$$ (Estadístico **del** contraste)

Este cociente se utiliza para contrastar la hipótesis nula de no interacción. Si no se rechaza esta hipótesis, entonces se emplean los estadísticos F

$$F_{a-1,\,ab(n-1)}=\frac{MS_A}{MS_E}$$ (Estadístico **del** contraste)

y

$$F_{b-1,\,ab(n-1)}=\frac{MS_B}{MS_E}$$ (Estadístico del contraste)

para contrastar la hipótesis nula de no diferencia entre las medias de niveles de los factores A y B, respectivamente.

Estos contrastes se denominan contrastes para «efectos principales». Si se rechaza la hipótesis nula de no interacción, no contrastamos los efectos principales. En cambio, se ejecutan b análisis de una vía para detectar diferencias entre los niveles del factor A en cada nivel del factor B individualmente. En cada caso, el rechazo se da para valores del cociente F que son demasiado grandes para que se deban al azar.

Aplicamos estas ideas completando el análisis de los datos del Ejemplo 10.5.2.

Ejemplo 10.5.3. Se obtuvieron los siguientes totales en el Ejemplo 10.5.2:

$$T_{11\cdot} = 5.35 \quad T_{22\cdot} = 6.5 \quad T_{\cdot 1\cdot} = 8.55$$
$$T_{21\cdot} = 3.2 \quad T_{1\cdot\cdot} = 17.55 \quad T_{\cdot 2\cdot} = 18.7$$
$$T_{12\cdot} = 12.20 \quad T_{2\cdot\cdot} = 9.7 \quad T_{\cdots} = 27.25$$
$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{5} X_{ijk}^2 = 48.26$$

Estos totales se utilizan para obtener las necesarias sumas de cuadrados:

$$SS_{\text{Tr}} = \frac{(5.35)^2}{5} + \frac{(3.2)^2}{5} + \frac{(12.20)^2}{5} + \frac{(6.5)^2}{5} - \frac{(27.25)^2}{20} = 8.86$$

$$SS_A = \frac{(17.55)^2}{10} + \frac{(9.7)^2}{10} - \frac{(27.25)^2}{20} = 3.08$$

$$SS_B = \frac{(8.55)^2}{10} + \frac{(18.7)^2}{10} - \frac{(27.25)^2}{20} = 5.15$$

$$SS_{AB} = SS_{\text{Tr}} - SS_A - SS_B = 8.86 - 3.08 - 5.15 = 0.63$$

$$SS_{\text{Total}} = 48.26 - \frac{(27.25)^2}{20} = 11.13$$

$$SS_E = SS_{\text{Total}} - SS_{\text{Tr}} = 11.13 - 8.86 = 2.27$$

El ANOVA se da en la Tabla 10.13.

Observemos primero el cociente F empleado para comprobar la interacción (* en la tabla). El valor P para este estadístico es:

$$P = P[F_{1,16} \geq 4.5] \cong 0.05$$

Puesto que la probabilidad es pequeña, rechazamos la hipótesis nula de no interacción y concluimos que existe interacción entre el ciclo de luz utilizado y la temperatura.

El origen de esta interacción puede observarse en la Figura 10.7. Compruébese que los segmentos que unen las medias muestrales de los dos niveles del factor B no son paralelos. Mientras que existe un decremento en el GSI en cada caso, a medida que variamos de 9 a 14 horas de luz, el decremento es más rápido a 16 °C que a 27 °C. En el comportamiento de las dos temperaturas hay una inconsistencia. Para completar el análisis realizamos dos análisis de una vía. En particular, comparamos la media del GSI con 9 horas de luz, con la de 14 horas de

Tabla 10.13. ANOVA para los datos del Ejemplo 10.5.3

Origen	*DF*	*SS*	*MS*	*F*
Tratamiento	3	8.86	2.95	21.07
A	1	3.08	3.08	22.0
B	1	5.15	5.15	36.79
AB	1	0.63	0.63	4.5*
Error	16	2.27	0.14	
Total	19	11.13		

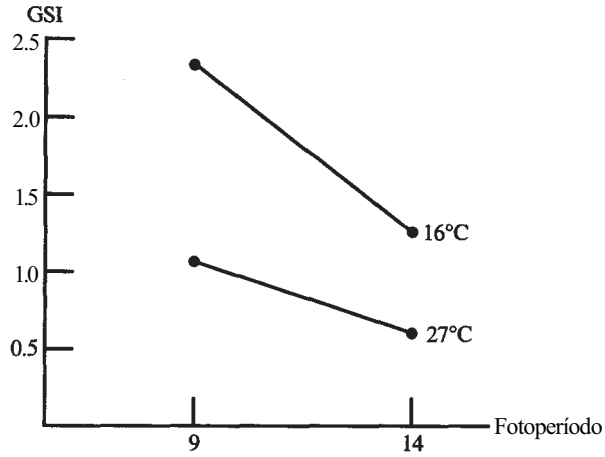


Figura 10.7. Los segmentos no son paralelos, indicando una interacción entre la temperatura y el fotoperíodo.

luz, primero a 16 °C y después a 27 °C. Los resultados de estos contrastes se muestran en las Tablas 10.14 y 10.15, respectivamente. Obsérvese que en cada caso se puede rechazar la hipótesis nula de igualdad de las medias con $P < 0.01$. Así, se puede concluir que a cada temperatura existe una diferencia entre las medias del GSI de los dos fotoperíodos.

Obsérvese que en este diseño los tamaños de las celdas son iguales. Hay n observaciones tomadas en cada combinación del tratamiento. Esto es *esencial* para que los análisis que se acaban de presentar sean válidos. Si los tamaños de las celdas no son iguales, se dice que el diseño está desequilibrado (o «no balanceado»). Estos diseños son difíciles de manejar. Si aparece esta situación durante un proyecto de investigación, se deberá buscar la ayuda de un profesional de la estadística para analizar sus datos.

Tabla 10.14. Análisis de la varianza de una vía utilizado para comparar el GSI medio con 9 horas de luz y 14 horas de luz, cuando la temperatura es de 16 °C

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	1	3.249	3.249	12.306
Error	8	2.1122	0.264025	
Total	9	5.3612		

Tabla 10.15. Análisis de la varianza de una vía utilizado para comparar el GSI medio con 9 horas de luz y 14 horas de luz, cuando la temperatura es de 27 °C

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	1	0.46225	0.46225	23.228
Error	8	0.1592	0.0199	
Total	9	0.62145		

Comparaciones múltiples y por parejas

Si no se rechaza la hipótesis nula de no interacción en un análisis de la varianza de dos vías, el interés se centrará entonces en los efectos principales. Si se encuentran diferencias entre los niveles del factor *A o B,* pueden utilizarse los contrastes *T* de Bonferroni o el contraste de Duncan de rango múltiple para señalar las diferencias. Los contrastes de Bonferroni son contrastes *T* de varianza conjunta que utilizan MS_E como estimador de σ^2.

Para señalar las diferencias entre los niveles del factor *A* utilizando el contraste de Duncan

$$SSR_p = r_p\sqrt{\frac{MS_E}{bn}}$$

Para determinar las diferencias existentes entre los niveles del factor *B,*

$$SSR_p = r_p\sqrt{\frac{MS_E}{an}}$$

Nota sobre los cálculos

Como se puede apreciar, la realización completamente manual de un análisis de la varianza de dos vías requiere mucho tiempo. Esto es especialmente así para grandes conjuntos de datos. En la práctica, los datos de este tipo se analizan con un ordenador. El Ejemplo 10.5.4 ilustra el análisis mediante ordenador de los datos del Ejemplo 10.5.3. En la sección de Herramientas Computacionales, al final del capítulo, se puede consultar el código del programa del SAS empleado para producir este *output.*

Ejemplo 10.5.4. En la Tabla 10.16 se presenta la tabla del análisis de la varianza de dos vías para los datos del Ejemplo 10.5.3. La tabla no tiene el mismo formato que la Tabla 10.13, pero pueden encontrarse los valores dados. En particular, estos son los valores de interés:

① Grados de libertad para tratamientos (celdas)
② Suma de cuadrados de los tratamientos
③ Media de cuadrados de los tratamientos
④ Grados de libertad totales
⑤ Suma total de cuadrados
⑥ Grados de libertad del error
⑦ Suma de cuadrados del error
⑧ Media de cuadrados del error
⑨ Grados de libertad de la interacción
⑩ Suma de cuadrados de la interacción
⑪ Media de cuadrados de la interacción
⑫ Cociente *F* utilizado para contrastar la hipótesis nula de no interacción
⑬ Valor *P* para el contraste de interacción
⑭ Grados de libertad del factor *B*
⑮ Suma de cuadrados del factor *B*
⑯ Media de cuadrados del factor *B*
⑰ Cociente *F* utilizado para contrastar las diferencias globales entre los niveles del factor *B*
⑱ Valor *P* para el contraste de efectos principales de *B*

Tabla 10.16. Análisis de la varianza de dos vías para los datos del Ejemplo 10.5.3. Obsérvese que la interacción está presente. El análisis se completa remitiéndose a las Tablas 10.17 y 10.18

SAS

General Linear Models Procedure

Dependent Variable: GSI

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Model	3 ①	8.86237500 ②	2.95412500 ③	20.81	0.0001
Error	16 ⑥	2.27140000 ⑦	0.14196250 ⑧		
Corrected Total	19 ④	11.13377500 ⑤			

R-Square	C.V.	Root MSE	GSI Mean
0.795990	27.65351	0.376779	1.36250000

Source	DF	Type I *SS*	Mean Square	*F* Value	Pr > *F*
PHOTOPRD	1	3.08112500	3.08112500	21.70	0.0003
TEMP	1	5.15112500	5.15112500	36.29	0.0001
PHOTOPRD*TEMP	1	0.63012500	0.63012500	4.44	0.0513

Source	DF Type III *SS*	Mean Square	*F* Value	Pr > *F*
PHOTOPRD	1 ⑲ 3.08112500 ⑳	3.08112500 ㉑	21.70 ㉒	0.0023 ㉓
TEMP	1 ⑭ 5.15112500 ⑮	5.15112500 ⑯	36.29 ⑰	0.0001 ⑱
PHOTOPRD*TEMP	1 ⑨ 0.63012500 ⑩	0.63012500 ⑪	4.44 ⑫	0.0513 ⑬

⑲ Grados de libertad del factor *A*
⑳ Suma de cuadrados del factor *A*
㉑ Media de cuadrados del factor *A*
㉒ Valor *F* utilizado para contrastar las diferencias globales entre los niveles del factor *A*
㉓ Valor *P* para el contraste de efectos principales de *A*

Obsérvese que, en este caso, primero contrastamos la interacción. Puesto que la interacción está presente, no nos molestamos con los contrastes para efectos principales. En su lugar, continuamos interpretando el ANOVA de una vía dado en las Tablas 10.17 y 10.18. En la Tabla 10.17 observamos que existe una diferencia en los niveles de la media del GSI entre los dos fotoperíodos *(P* = 0.008) a 16 °C. La tabla 10.18 muestra que también existe una diferencia en los niveles de la media del GSI entre los dos fotoperíodos a 27 °C *(P* = 0.0013).

Tabla 10.17. Análisis de la varianza de una vía utilizado para comparar fotoperíodos cuando la temperatura es de 16 °C

SAS

Analysis-of-Variance Procedure

Dependent Variable: GSI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > *F*
Model	1	3.24900000	3.24900000	12.31	0.0080
Error	8	2.11220000	0.26402500		
Corrected Total	9	5.36120000			

R-Square	C.V.	Root MSE	GSI Mean
0.606021	27.47773	0.513834	1.87000000

Source	DF	Anova SS	Mean Square	F Value	Pr > *F*
PHOTOPRD	1	3.24900000	3.24900000	12.31	0.0080

Tabla 10.18. Análisis de la varianza de una vía utilizado para comparar fotoperíodos cuando la temperatura es de 27 °C

SAS

Analysis-of-Variance Procedure

Dependent Variable: GSI

Source	DF	Sum of Squares	Mean Square	FValue	Pr > *F*
Model	1	0.46225000	0.46225000	23.23	0.0013
Error	8	0.15920000	0.01990000		
Corrected Total	9	0.62145000			

R-Square	C.V.	Root MSE	GSI Mean
0.743825	16.49911	0.141067	0.85500000

Source	DF	Anova *SS*	Mean Square	*F* Value	Pr > *F*
PHOTOPRD	1	0.46225000	0.46225000	23.23	0.0013

EJERCICIOS 10.5

Nota: Si es posible, realícense estos ejercicios con la ayuda de un ordenador.

1. Se lleva a cabo un estudio para determinar el efecto del nivel del agua y del tipo de planta sobre la longitud global del tallo de las plantas de guisantes. Para ello, se utilizan tres niveles de agua y dos tipos de plantas. Se dispone para el estudio de dieciocho plantas sin hojas. Se dividen aleatoriamente en tres subgrupos, y después se asignan los niveles de agua aleatoriamente a los grupos. Un procedimiento similar se sigue con 18 plantas convencionales. Se obtuvieron los siguientes datos (la longitud del tallo se da en centímetros):

Factor *B*	**Factor *A* (nivel de agua)**			
(tipo de planta)	**Bajo**	**Medio**	**Alto**	**Total (factor *B*)**
Sin hojas	69.0	96.1	121.0	1788
	71.3	102.3	122.9	
	73.2	107.5	123.1	
	75.1	103.6	125.7	
	74.4	100.7	125.2	
	75.0 (438)	101.8 (612)	120.1 (738)	
Convencional	71.1	81.0	101.1	1578
	69.2	85.8	103.2	
	70.4	86.0	106.1	
	73.2	87.5	109.7	
	71.2	88.1	109.0	
	70.9 (426)	87.6 (516)	106.9 (636)	
Total (factor *A*)	864	1128	1374	3366

a) Comprobar los totales dados.
b) Para estos datos,

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{6} X_{ijk}^2 = 327\,431.42$$

Utilizarlo para hallar SS_{Total}.
c) Hallar SS_{Tr}. Emplearlo junto con SS_{Total} para calcular SS_E.
d) Hallar SS_A *y* SS_B. Emplearlos junto con SS_{Tr} para calcular SS_{AB}.
e) Obtener la tabla ANOVA y utilizarla para contrastar las hipótesis apropiadas.
f) Si es adecuado, continuar el análisis utilizando el contraste de Duncan de rango múltiple.

2. Se realiza un estudio sobre el efecto del fotoperíodo y del genotipo sobre el período latente de infección del moho de cebada aislado AB3. Se obtienen cincuenta hojas de cuatro genotipos distintos y se dividen aleatoriamente en cinco subgrupos, cada uno de tamaño 10. Cada grupo es infectado y posteriormente expuesto a diferente fotoperíodo. La respuesta anotada es el número de días hasta la aparición de síntomas visibles. Se hallaron los siguientes *totales* para los niveles y tratamientos:

Factor *B* (genotipo)	**Factor *A* (fotoperíodo: horas de oscuridad por ciclo de 24 horas)**					
	0	**2**	**4**	**8**	**16**	**Total (factor *B*)**
Armelle	630	610	560	570	590	2960
Golden Promise	640	630	600	620	620	3110
Emir	640	630	650	620	580	3120
Vacia	660	660	620	610	630	3180
Total (factor A)	2570	2530	2430	2420	2420	12370

a) Para estos datos,

$$\sum_{i=1}^{5}\sum_{j=1}^{4}\sum_{k=1}^{10} X_{ijk}^2 = 773\,377.2$$

Utilizarlos para hallar SS_{Total}.
b) Hallar SS_{Tr}. Utilizarlo junto a SS_{Total} para calcular SS_E.
c) Hallar SS_A *y* SS_B. Emplearlos junto a SS_{TR} para hallar SS_{AB}.
d) Construir la tabla ANOVA y contrastar las hipótesis oportunas.
e) Donde sea apropiado, continuar el análisis utilizando el contraste de Duncan de rango múltiple, con un $\alpha = 0.05$.

3. Se realiza un estudio sobre la solubilidad capsular en líquidos biológicos de dos de las preparaciones de enzimas que más comúnmente se elaboran en cápsulas. El propósito es determinar el efecto del tipo de cápsula y del líquido biológico sobre el tiempo que transcurre hasta la disolución de la cápsula. Nos referimos a dos líquidos biológicos: los jugos gástricos y duodenales, y a dos tipos de cápsulas, *C y V*. De este modo, están implicados dos factores, cada uno estudiado a dos niveles. Para llevar a cabo el estudio se obtuvieron 10 cápsulas vacías de cada tipo y se dividieron aleatoriamente en dos subgrupos, cada uno de tamaño 5. Se disolvió uno de los grupos en jugos gástricos y el otro en

jugos duodenales. La respuesta anotada es el instante en que se liberan las primeras burbujas de aire a través de las perforaciones en las cápsulas. Se obtuvieron los siguientes datos (el tiempo está en minutos):

	Factor *A* (tipo de líquido)		
Factor *B* (tipo de cápsula)	**Gástrico**	**Duodenal**	**Total (factor *B*)**
C	39.5 **45.7** 49.8 50.2 63.8 (249)	31.2 33.5 36.7 42.0 38.1 (181.5)	430.5
V	47.4 43.5 39.8 36.1 **41.2** (208)	44.0 41.2 47.3 45.3 42.7 (220.5)	428.5
Total (factor *A*)	457	402	859

a) Comprobar los totales dados.

b) Para estos datos,

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{5} X_{ijk}^2 = 37\,847.26$$

Utilizarlo para hallar SS_{Total}.

c) Hallar SS_{TT}. Utilizarlo junto a SS_{Total} para hallar SS_E.

d) Calcular SS_A y SS_B. Utilícense junto a SS_{TR} para hallar SS_{AB}.

e) Esbozar una gráfica de las medias muestrales para los niveles del factor *A* para cada nivel del factor *B*. ¿Puede sospecharse la presencia de interacción? Explicarlo.

f) Completar el análisis de los datos hallando la tabla ANOVA y contrastando las oportunas hipótesis.

4. La cotinina es uno de los principales metabolitos de la nicotina. Actualmente se la considera el mejor indicador de la exposición al humo del tabaco. Se ha realizado un estudió para detectar las posibles diferencias raciales en el nivel de cotinina en adultos jóvenes. Se han obtenido los datos de la tabla siguiente sobre el nivel de cotinina en miligramos por mililitro.

	Blancos	**Negros**
Varones	210 300 150 325 100(1085)	245 347 125 250 260(1227)
Mujeres	177 300 106 150 160(893)	252 152 315 267 275(1261)

a) Construir una tabla de análisis de la varianza de dos vías para estos datos y utilizarla para contrastar la hipótesis nula de no interacción.
b) Si no se halla interacción, contrastar los efectos principales.
c) Si se detecta interacción, construir un diagrama similar al de la Figura 10.6 para investigar el origen de la interacción.
d) Si se detecta interacción, comparar el nivel medio de cotinina entre mujeres de raza blanca y de raza negra mediante el análisis de la varianza de una vía. Hacer lo mismo para los varones.

(Basado en las medias halladas en Lynne Wagenknecht *et al.,* «Racial Differences in Serum Cotinine Levels Among Smokers in the Coronary Artery Risk Development in Young Adults Study», *American Journal of Public Health,* septiembre de 1990, págs. 1053-1056.)

5. Se ha realizado un estudio sobre el efecto que produce la descarga de aguas residuales de una planta sobre la ecología del agua natural. En el estudio se utilizaron dos lugares de muestreo. Un lugar está aguas arriba del punto en el que la planta introduce aguas residuales en la corriente; el otro está aguas abajo. Se tomaron muestras durante un período de tres semanas y se obtuvieron estos datos sobre el número de diatomeas halladas:

	Semana		
Lugar	1	2	3
Arriba	689	831	558
	756	916	423
Abajo	204	56	34
	229	73	78

a) Construir una tabla de análisis de la varianza de dos vías para estos datos.
b) Contrastar la interacción y continuar el análisis de forma adecuada según los resultados de este contraste.

Tabla 10.19. Análisis de la varianza para el Ejercicio 6

SAS

SPECIES = spirochetes

Analysis-of-Variance Procedure

Dependent Variable: COUNT

Source	DF	Sum of Squares	Mean Square	*F* Valué	Pr > *F*
Model	5	518.7500000	103.7500000	0,.40	0.8321
Error	6	1549.5000000	258.2500000		
Corrected Total	11	2068.2500000			

R-Square	C.V.	Root MSE	COUNT Mean
0.250816	64.92993	16.07016	24.7500000

Source	DF	Anova *SS*	Mean Square	*F* Value	Pr > F
SITE	1	90.7500000	90.7500000	0**.35**	0.5750
WEEK	2	84.5000000	42.2500000	0**.16**	0.8527
SITE*WEEK	2	343.5000000	171.7500000	0**.67**	0.5484

Tabla 10.20. Análisis de la varianza para el Ejercicio 7

SAS

SPECIES = protists

Analysis-of-Variance Procedure

Dependent Variable: COUNT

Source	DF	Sum of Squares	Mean Square	*F* Valué	Pr > *F*
Model	5	84.670.00000	16.934.00000	65.93	0.0001
Error	6	1.541.00000	256.83333		
Corrected Total	11	86.211.00000			

.R-Square	C.V.	Root MSE	COUNT Mean
0.982125	16.10655	16.02602	99.5000000,

Source	DF	Anova *SS*	Mean Square	*F* Valué	Pr > *F*
SITE	1	83.333.33333	83.333.33333	324.46	0.0001
WEEK	2	738.50000	369.25000	1.44	0.3090
SITE*WEEK	2	598.16667	299.08333	1.16	0.3738

c) Utilizar los contrastes *T* de Bonferroni con cada contraste *T* realizado al nivel $\alpha = 0.01$ para señalar cualquier diferencia existente. ¿Cuál es la probabilidad global máxima de efectuar al menos un rechazo incorrecto?

(Basado en un estudio de Joseph Hutton, Departamento de Biología y Servicio de Consulting Estadístico, Universidad de Radford, octubre de 1990.)

6. Considérese el experimento descrito en el Ejercicio 5. El *printout* de la Tabla 10.19 es el análisis para el número de espiroquetas halladas. Interpretar el *printout.*
7. Considérese el experimento descrito en el Ejercicio 5. El *printout* de la Tabla 10.20 es el análisis para el número de protistas hallados. Interpretar el *printout.*

HERRAMIENTAS COMPUTACIONALES

TI83

XXIV. ANOVA de una vía

La calculadora TI83 lleva a cabo el contraste *F* necesario para contrastar H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$. Los pasos necesarios se ilustran utilizando los datos del Ejemplo 10.2.1. Los datos se introducirán en las cinco primeras columnas del editor de datos.

Tecla/Comando de la TI83	Propósito
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 40 ENTER 45 ENTER 51 ENTER	2. Introduce los datos para la temperatura I en la columna L1.

3. ▷ 36 ENTER 42 ENTER 40 ENTER	3. Introduce los datos para la temperatura II en la columna L2.
4. ▷ 49 ENTER 51 ENTER 50 ENTER	4. Introduce los datos para la temperatura III en la columna L3.
5. ▷ 47 ENTER 49 ENTER 51 ENTER	5. Introduce los datos para la temperatura IV en la columna *LA*.
6. ▷ 55 ENTER 60 ENTER 61 ENTER	6. Introduce los datos para la temperatura V en la columna L5.
7. STAT ◁ ALPHA F	7. Accede a la pantalla necesaria para realizar el ANOVA de una vía ANOVA(.
8. 2ND L1 , 2ND L2 , 2ND L3 , 2ND L4 , 2ND L5)	8. Indica que los datos para el análisis están en las columnas L1 a L5.

9. ENTER	9. Muestra el ANOVA; reproduce la información que se encuentra en la Tabla 10.3.

Paquete estadístico SAS

X. ANOVA de una vía con comparaciones múltiples

El código SAS empleado para producir el *output* mostrado en el Ejemplo 10.2.5 se muestra a continuación.

Código SAS	**Proposito**
OPTIONS LS = 80 PS = 60 NODATE;	Establece las opciones de impresión.
DATA TOXIC;	Nombra el conjunto de datos.
INPUT TEMP PERCENT;	Nombra las variables.
LINES;	Indica que los datos van a continuación.
1 40 1 45 1 51 2 36 2 42 2 40 5 55 5 60 5 61	Líneas de datos.
	Indica el final de los datos.
PROC ANOVA;	Solicita el procedimiento para el análisis de
CLASSES TEMP;	la varianza; indica que los datos están
MODEL PERCENT = TEMP;	agrupados por temperatura.
MEANS TEMP/DUNCAN BON;	Solicita que se ejecuten los contrastes de Duncan y Bonferroni para comparaciones múltiples

XI. Bloques completos aleatorizados

El código SAS utilizado para realizar el *output* que se muestra en el Ejemplo 10.4.7 se da a continuación.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Establece las opciones de impresión.
DATA EXERCISE;	Nombra el conjunto de variables.
INPUT BLOCK TRTMENT KILOCAL;	Nombra las variables.

LINES;	Indica que los datos vienen a continuación.
1 1 1.4	
2 1 1.5	
3 1 1.8	
8 1 2.0	
1 2 1.1	
2 2 1.2	Líneas de datos.
3 2 1.3	
8 2 1.3	
1 3 0.7	
2 3 0.8	
3 3 0.7	
8 3 0.6	Indica que finalizan los datos.
PROC GLM;	Solicita que el análisis se realice por el procedimiento de modelo lineal general.
CLASSES TRTMENT BLOCK;	Indica que los datos están agrupados por bloque y tratamiento.
MODEL KILOCAL = TRTMENT BLOCK;	Solicita que se realicen comparaciones
MEANS TRTMENT/DUNCAN BON;	múltiples sobre la media de kilocalorías por las técnicas de Duncan y Bonferroni.

XII. ANOVA de dos vías

El código SAS utilizado para producir el *ouput* que se da en el Ejemplo 10.5.4 se muestra a continuación.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Establece la configuración de impresión.
DATA MIROGREX;	Nombra el conjunto de los datos.
INPUT PHOTOPRD TEMP GSI;	Nombra las variables.
LINES;	Indica que los datos van a continuación.
1 1 0.9	
1 1 1.06	
1 1 1.12	
1 2 0.83	
1 2 0.67	Líneas de datos.
1 2 0.66	
2 1 1.3	
2 1 2.88	

```
2    1    2.94
2    2    1.01
2    2    1.52

2    2    1.63
```

Indica el final de los datos.

```
PROC GLM;
CLASSES PHOTOPRD TEMP;
```

Solicita que se analicen los datos por el procedimiento del modelo lineal general; indica que los datos están agrupados por fotoperíodo y temperatura.

```
MODEL GSI = PHOTOPRD
   TEMP PHOTOPRD*TEMP;
```

Indica que la variable de respuesta (a la izquierda) es GSI; indica que los factores son fotoperíodo y temperatura; incluir el término de la interacción PHOTOPRD*TEMP.

El siguiente código puede añadirse al programa anterior para producir una gráfica que muestra visualmente la posible interacción entre temperatura y fotoperíodo. La gráfica, por supuesto, utiliza las medias muestrales como estimador de las medias poblacionales en cuestión.

```
PROC SORT; BY PHOTOPRD
   TEMP;
```

Ordena los datos por fotoperíodo y temperatura.

```
PROC MEANS; BY PHOTOPRD
   TEMP;
```

Calcula las medias muestrales para cada de las cuatro combinaciones de tratamientos

```
OUTPUT OUT=NEW MGSI=MEAN;
```

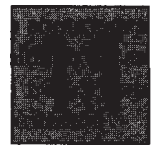
Saca las cuatro medias muestrales a un conjunto llamado NEW; llama a las medias MGSI.

```
PROC PLOT;
PLOT MGSI*PHOTOPRD=TEMP;
```

Dibuja las medias de cada temperatura y fotoperíodo.

```
TITLE'BÚSQUEDA DE
   INTERACCIÓN;
```

Titula el *output.*



Regresión y correlación

En este capítulo comentamos dos tipos de problemas. El primero, llamado *regresión,* implica necesariamente la obtención de una ecuación mediante la cual pueda estimarse el valor medio de una variable aleatoria, desde el conocimiento de los valores tomados por una o más variables diferentes. El segundo problema, llamado *correlación,* consiste en medir la fuerza de la relación lineal entre dos variables aleatorias. Comenzamos con una breve introducción al tema de la regresión.

11.1. INTRODUCCIÓN A LA REGRESIÓN LINEAL SIMPLE

En un problema de regresión, estamos interesados principalmente en una variable aleatoria simple *Y.* Se supone que el valor tomado por esta variable aleatoria depende o está influenciado por los valores tomados por una o más variables diferentes. La variable aleatoria Y se denomina *variable dependiente* o *respuesta;* las variables que influyen en *Y* se denominan *variables independientes, variables predictoras* o *regresores.* Al realizar estimaciones o predicciones, los regresores no se tratan como variables aleatorias. Por el contrario, son entidades que pueden asumir valores diferentes, pero cuyos valores, en el momento en que debe hacerse la predicción, no se determinan al azar. Para ilustrarlo, supóngase que deseamos obtener una ecuación para describir la temperatura del agua fuera de la plataforma continental. Puesto que la temperatura depende en parte de su profundidad, hay dos variables implicadas: *X,* profundidad del agua, e *Y,* temperatura. No estamos interesados en hacer inferencias sobre la profundidad del agua. En cambio, queremos describir el comportamiento de la temperatura del agua bajo la suposición de que su profundidad se conoce de antemano con precisión. La temperatura del agua es la respuesta; la profundidad es el único regresor considerado.

Para ilustrar la notación que se utilizará, obsérvese que, incluso si la profundidad del agua está fijada en algún valor *x,* la temperatura del agua variará debido a otras influencias aleatorias. Por ejemplo, si se toman varias mediciones de temperatura en diferentes lugares, cada una a una profundidad de $x = 1000$ pies, los valores de las mediciones variarán. Por esta razón, debemos admitir que para una x dada, estamos realmente tratando con una variable aleatoria «condicional», que indicamos mediante $Y|x$ (y dado que $X = x$). Esta variable aleatoria condicional tiene una media a la que se notará mediante $\mu_{Y|x}$. Resulta obvio que la temperatura media del agua del océano depende en parte de la profundidad del agua; no

esperamos que la temperatura media $x = 1000$ pies sea la misma que $x = 5000$ pies. Es *decir,* es razonable suponer que $\mu_{Y|x}$ es una función de x. A la gráfica de esta función la denominamos *curva de regresión de Y sobre X.* Esta idea se ilustra en la Figura 11.1.

Nuestro problema inmediato es estimar la forma de $\mu_{Y|x}$ a partir de los datos obtenidos en algunos valores seleccionados $x_1, x_2, x_3, \ldots, x_n$ de la variable predictora X. Los valores reales utilizados para desarrollar el modelo no son demasiado importantes. Si existe una relación funcional, ésta se manifestará independientemente de qué valores de X se utilicen para descubrirlo. Sin embargo, por razones prácticas, estos valores deberían representar una gama bastante amplia de los valores posibles de la variable independiente X. A veces se pueden preseleccionar los valores utilizados. Por ejemplo, al estudiar la relación entre la temperatura del agua y su profundidad, podemos saber que nuestro modelo debe utilizarse para predecir la temperatura del agua a profundidades de 1000 a 5000 pies. Podemos medir las temperaturas del agua a cualquier profundidad deseada dentro de este rango. Por ejemplo, podemos tomar mediciones en incrementos de 1000 pies. De esta forma, prefijamos nuestros valores a $x_1 = 1000, x_2 = 2000, x_3 = 3000, x_4 = 4000$ y $x_5 = 5000$. Cuando se preseleccionan los valores X utilizados para desarrollar la ecuación de regresión, se dice que el estudio está *controlado:* A menudo, los valores X utilizados para desarrollar la ecuación se eligen mediante algún mecanismo aleatorio. Por ejemplo, al estudiar el efecto de la calidad del aire sobre el pH del agua de lluvia, nos veremos forzados a seleccionar una muestra de días, anotar la lectura de la calidad del aire ese día y medir el pH del agua de lluvia. En este caso, los valores de X utilizados para desarrollar la ecuación de regresión no están preseleccionados por el investigador. Representan un conjunto de valores típicos de X. Los estudios de este tipo se denominan *estudios observacionales.* Los dos ejemplos siguientes ilustran este tipo de estudios.

Ejemplo 11.1.1. Un médico quiere predecir la concentración de un determinado fórmico en la corriente sanguínea, cinco minutos después de su administración *(Y),* basándose en el conocimiento del tamaño de la dosis inicial (X). En este caso, la variable aleatoria Y es la variable dependiente; X es la variable independiente. En un experimento controlado en el laboratorio, el experimentador selecciona los valores tomados por X. Por ejemplo, podríamos elegir experimentar con dosis de 0.05, 0.10, 0.20 y 0.30 mL. Puesto que la elección de las dosis experimentales está en manos del investigador, éste es un estudio controlado.

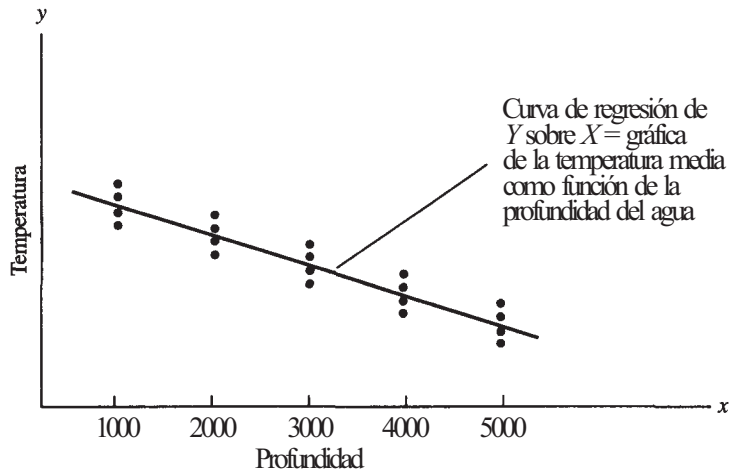


Figura 11.1. Para una profundidad de agua dada, la temperatura varía alrededor de un valor medio desconocido $\mu_{Y|x}$. La curva que une estos valores medios se denomina curva de regresión de Y sobre X.

Ejemplo 11.1.2. Un ecólogo desea predecir el cambio de temperatura del agua que tiene lugar en un punto situado una milla por debajo de una planta industrial, tras el vertido de aguas residuales calientes en la corriente *(Y).* La predicción se basará en la cantidad de agua liberada *(X).* La respuesta es el cambio de la temperatura del agua; el único regresor es la cantidad de agua liberada. Puesto que la cantidad de agua caliente liberada varía dependiendo del nivel de actividad en la planta, el experimentador no controla los valores del regresor utilizado para desarrollar la ecuación de predicción. Más bien, los valores se miden simplemente en el momento de liberación del agua. Este es un ejemplo de estudio observacional.

Independientemente de si el estudio es controlado u observacional, el objeto del estudio de regresión es el mismo: encontrar una ecuación de predicción o regresión razonable. Las curvas típicas de regresión se recogen en la Figura 11.2. Obsérvese que estas curvas son

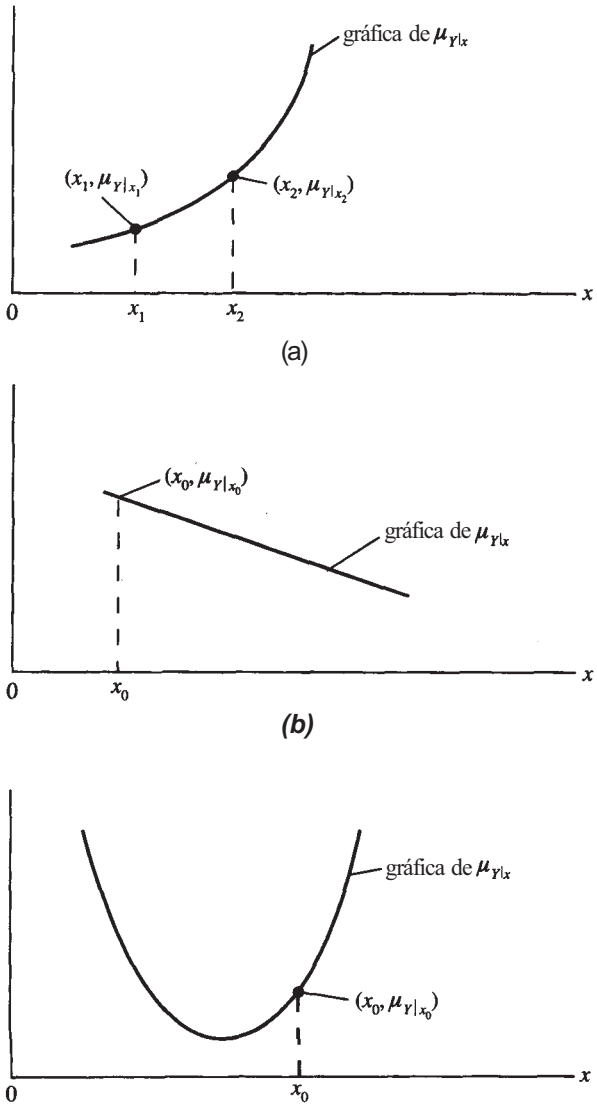


Figura 11.2. a) Curva de regresión no lineal; *b)* curva de regresión lineal (recta); *c)* curva de regresión no lineal.

teóricas. Son los gráficos de las medias teóricas de la variable dependiente *Y* para valores dados de la variable independiente X. Sirven como curvas *ideales* de predicción. Generalmente, no conocemos las ecuaciones exactas para estas curvas. Nuestro problema es estimarlas a partir de datos observados sobre X e *Y.* Cuando el gráfico de $\mu_{Y|x}$ es una línea recta (Figura 11*2b),* decimos que la regresión de *Y* sobre X es *lineal.* De lo contrario, se dice que es la regresión *no lineal* (Figuras 11.*2a* y 11.2c). El término *regresión lineal* se define, de manera más precisa, como sigue:

Definición 11.1.1. Regresión lineal. Una curva de regresión de *Y* sobre *X* se dice que es una *regresión lineal* si y sólo si:

$$\mu_{Y|x} = \alpha + \beta x$$

para α y β números reales, $\beta \neq 0$.

El parámetro α de la ecuación anterior es el punto de intersección y. Es el punto en que la recta corta al eje vertical o eje y. El parámetro β es la pendiente de la recta. Si $\beta > 0$, la recta se inclina hacia arriba a medida que nos movemos de izquierda a derecha; si $\beta < 0$, la recta se inclina hacia abajo; y si $\beta = 0$, la recta es horizontal. Por ejemplo, puesto que la recta de la Figura 11.1 se inclina hacia abajo de izquierda a derecha, su pendiente es negativa. La consecuencia práctica de esto es que, a medida que aumenta la profundidad del agua, disminuye la temperatura. En muchos casos, el signo algebraico de la pendiente puede anticiparse basándonos sólo en el conocimiento de la materia.

El título de esta sección es *introducción a la regresión lineal simple.* La frase implica ttes cosas: *Regresión,* supone que el propósito del experimento es la predicción. *Simple,* significa que intentaremos obtener una ecuación mediante la cual pueda predecirse el valor de una variable dependiente *Y,* basándonos en el conocimiento del valor tomado por *una* variable independiente *X.* Si se utilizara más de una variable independiente para predecir el valor de *Y* entonces emplearíamos el término *regresión múltiple.* El término *lineal* alude a que la ecuación de predicción tomará la forma de una línea recta.

En el análisis de la regresión, lo primero que debemos hacer es determinar si es razonable suponer que la curva de regresión de *Y* sobre X es una línea recta. Una forma de ayudar a hacerlo es dibujar un gráfico de los pares observados *(x,* y). A tal gráfico se le denomina *nube de puntos* o *diagrama de dispersión.* Los puntos, normalmente, no están formando exactamente una línea recta. Sin embargo, si la regresión lineal es aplicable, deberán mostrar una notable tendencia a la linealidad. Los dos ejemplos siguientes ilustran el uso de una nube de puntos para detectar la tendencia a la linealidad.

Ejemplo 11.1.3. Se realiza un experimento para estudiar la relación entre la altura de la concha (X) y su longitud *(Y),* cada una medida en milímetros, de *Patelloida pygmaea,* tina lapa pegada a las rocas y conchas a lo largo de las costas protegidas en el área Indo-Pacífica. Se tienen los siguientes datos:

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
0.9	**3.1**	**1.9**	**5.0**	**2.1**	**5.6**	**2.3**	**5.8**
1.5	**3.6**	**1.9**	**5.3**	**2.1**	**5.7**	**2.3**	**6.2**
1.6	**4.3**	**1.9**	**5.7**	**2.1**	**5.8**	**2.3**	**6.3**
1.7	**4.7**	**2.0**	**4.4**	**2.2**	**5.2**	**2.3**	**6.4**
1.7	**5.5**	**2.0**	**5.2**	**2.2**	**5.3**	**2.4**	**6.4**
1.8	**5.7**	**2.0**	**5.3**	**2.2**	**5.6**	**2.4**	**6.3**
1.8	**5.2**	**2.1**	**5.4**	**2.2**	**5.8**	**2.7**	**6.3**

La nube de puntos recogida en la Figura 11.3 se obtiene señalando los valores de la variable independiente X a lo largo del eje horizontal y los de la variable dependiente Y a lo largo del eje vertical. Incluso si estos puntos no están sobre una línea recta, hay una tendencia lineal. La tendencia es lo que vamos buscando. Ésta identifica el problema como uno en que es aplicable una regresión lineal simple. Es razonable suponer que el gráfico de $\mu_{Y|x}$ es una línea recta. Podemos visualizar la línea teórica de regresión de Y y X como se muestra en la Figura 11.4. Nuestro problema es utilizar los datos para estimar los valores de α y β con los que estimar la línea teórica de regresión de Y sobre X.

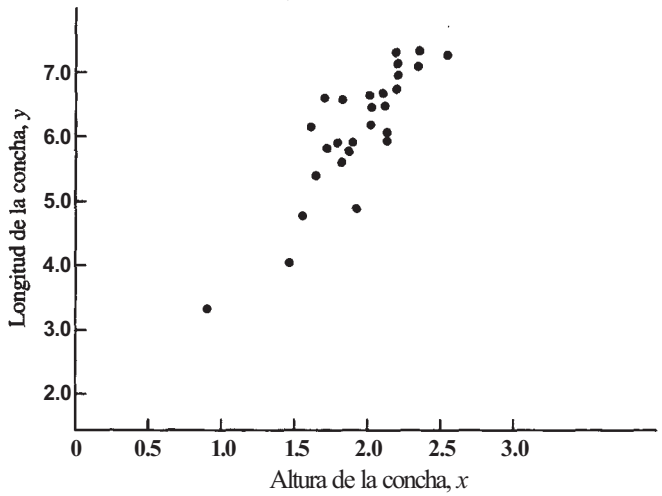


Figura 11.3. La nube de puntos de la altura de la concha frente a su longitud muestra una tendencia lineal.

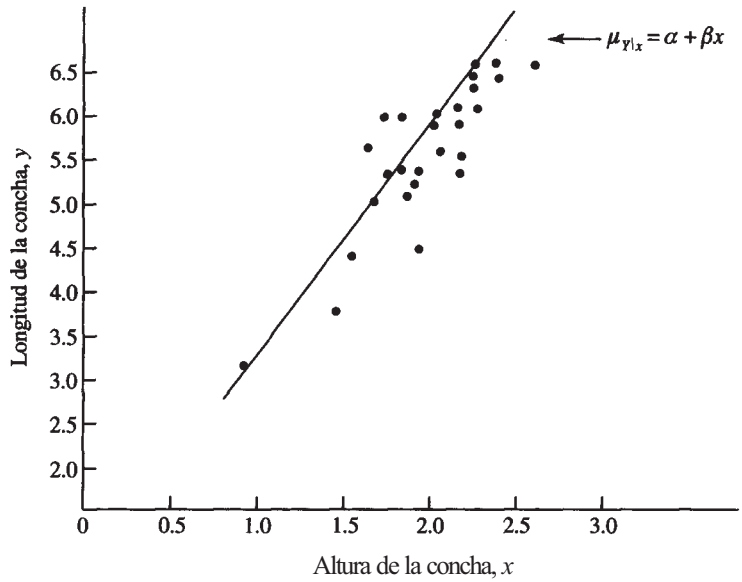


Figura 11.4. Curva teórica lineal (recta) de regresión y curva ideal para predecir la longitud de la concha basándose en su altura.

Ejemplo 11.1.4. En un estudio sobre depredación, un depredador, *Didinium nasutum,* se introduce en un medio que contiene a su presa natural, *Paramecium caudatum.*

El propósito es predecir la tasa de alimentación de *Didinium (Y)* basándose en el conocimiento de la densidad inicial de *Paramecium* en el medio (X). En la Figura 11.5 se presenta la nube de puntos de los datos obtenidos. Claramente, los puntos *no* muestran una tendencia lineal. En este caso, es inapropiado suponer que el gráfico $\mu_{Y|x}$ sea lineal. La curva hiperbólica que aparece en la Figura 11.6 es una elección más razonable. Para estimar $\mu_{Y|x}$ *no* utilizaríamos las técnicas de la regresión lineal simple.

Gráficos del tipo de los mostrados en las Figuras 11.3 y 11.5 se obtienen fácilmente utilizando la TI83 u otras calculadoras gráficas. También pueden realizarse con SAS u otros paquetes informáticos. Dibujar una nube de puntos debería ser siempre el primer paso en cualquier estudio de regresión lineal simple.

Nuestro problema en el Ejemplo 11.1.3 es estimar matemáticamente la ecuación para $\mu_{Y|x}$ basándonos en los datos observados. Fijémonos en que, a través de la nube de puntos, pueden trazarse muchas líneas rectas que pasan por algunos de los datos puntuales, y pueden trazarle otras que estén «próximas» a la mayor parte de los datos puntuales. ¿Cuál de estas líneas debemos escoger como nuestro estimado para $\mu_{Y|x}$? ¿Cuál es la línea que mejor se «ajusta» a los datos? Responderemos a estas preguntas en la Sección 11.2.

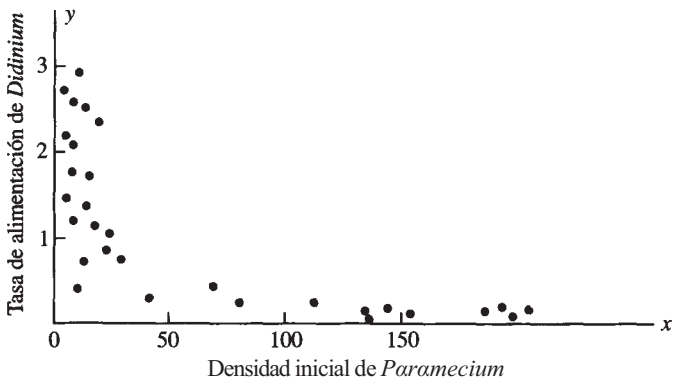


Figura 11.5. Nube de puntos de la densidad inicial de *Pararmecium* frente a la tasa de alimentación de *Didinium.* La tendencia no lineal es evidente.

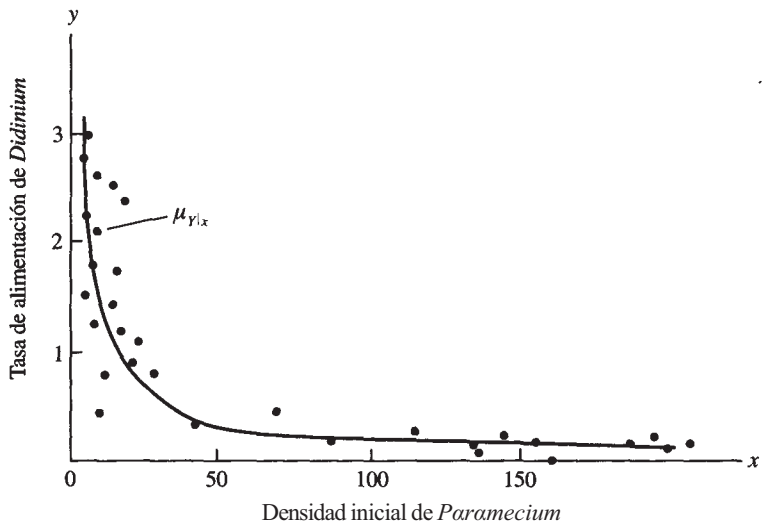


Figura 11.6. Curva teórica no lineal de regresión y curva ideal para predecir la tasa de alimentación de *Didinium* basándose en la densidad inicial de *Paramecium.*

EJERCICIOS 11.1

1. Considérense las siguientes observaciones sobre una variable dependiente y y una variable independiente *X.*

x	*y*	*x*	*y*
1.0	3.0	3.0	7.0
1.1	3.2	3.0	7.1
1.5	4.1	3.1	7.4
1.7	4.2	3.2	6.0
2.0	5.0	3.5	8.1
2.5	6.2	3.6	8.0
3.0	7.3	4.0	9.0

Dibujar la nube de puntos. ¿Parece aplicable la regresión lineal?

2. Considérense las siguientes observaciones sobre la variable aleatoria dependiente *Y,* lapso de atención de un niño en minutos, y *X,* el CI del niño.

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
75	2.0	95	5.2	110	7.2	130	3.8
80	3.0	100	5.5	115	6.8	135	2.9
85	4.5	105	6.0	115	6.4	140	2.0
85	4.7	110	6.5	120	5.5		
90	5.0	110	6.7	125	4.2		

Dibujar la nube de puntos. ¿Parece aplicable la regresión lineal?

3. Se realiza un experimento para estudiar la relación entre el período de incubación (número de días desde que se pusieron los huevos) y la media del tiempo de incubación (número medio de minutos dedicados ininterrumpidamente a la incubación en el nido) en un ave marina: el charrán de pico de gaviota. Se pretende obtener una ecuación mediante la cual se pueda predecir el tiempo medio de incubación *Y* a partir del conocimiento del período de incubación *X.* Utilizando fotografías a intervalos se obtuvieron los siguientes datos:

x	*y*	*x*	*y*	*x*	*y*
0.25	30	4	18	12	38
0.50	18	5	26	18	55
0.50	25	6	21	19	35
1.0	21	7	52	20	30
1.0	22	8	62	20	50
1.0	40	9	45	20	155
1.5	19	10	39	21	35
2	10	10	120	21	38
2.5	55	11	18		
3	23	11	50		

Dibujar la nube de puntos. ¿Parece aplicable la regresión lineal?

4. Se realiza un estudio para investigar la relación entre el nivel de humedad del suelo y la tasa de mortalidad en lombrices de tierra. La tasa de mortalidad, *Y,* es la proporción de lombrices de tierra que mueren tras un período de dos semanas; el nivel de humedad, x, viene medido en milímetros de agua por centímetro cuadrado de suelo. Se obtuvieron los siguientes datos:

x	*y*	*x*	*y*
0	0.5	0.632	0
0	0.4	0.947	0.1
0	0.5	0.947	0.2
0.316	0.2	0.947	0.1
0.316	0.3	1.26	0.6
0.316	0.3	1.26	0.5
0.632	0	1.26	0.4
0.632	0.1		

Dibujar una nube de puntos para estos datos. ¿Muestran los datos una tendencia lineal? ¿Qué tipo de curva debe usarse para describir la tendencia mostrada por los datos? (Basado en un estudio realizado por Jeffrey A. Hollar, Departamento de Biología, Universidad de Radford, 1996.)

11.2. MÉTODO DE LOS MÍNIMOS CUADRADOS

Recordemos que en la regresión lineal simple se da por supuesto que el gráfico de la media de la variable dependiente *Y,* para valores dados de la variable independiente X, es una línea recta. Es decir,

donde α y β son parámetros desconocidos cuyos valores es necesario estimar. Al método empleado para estimar α y β se le llama el *método de los mínimos cuadrados.* El procedimiento se explica en el Ejemplo 11.2.1.

Ejemplo 11.2.1. Se realiza un estudio de fotoperiodismo en aves acuáticas. Se pretende establecer una ecuación mediante la cual pueda predecirse la duración de la estación de cría, y, a partir del conocimiento del fotoperíodo (número de horas de luz por día) bajo el que se inició la reproducción, *x.* Se obtuvieron los siguientes datos, observando el comportamiento de once *Aythya* (patos buceadores).

x (horas de luz por día)	*y* (días de la estación de cría)
12.8	110
13.9	54
14.1	98
14.7	50
15.0	67
15.1	58
16.0	52

x (horas de luz por día)	*y* (días de la estación de cría)
16.5	50
16.6	43
17.2	15
17.9	28

En la Figura 11.7 se recogen los datos formando la nube de puntos, junto con una línea teórica imaginaria de regresión. Para estimar la línea teórica de regresión, debemos estimar α, valor «y» de la intersección de la línea con eje vertical, y β, su pendiente. Las estimaciones para estos parámetros se designan como α y b, respectivamente. De este modo, la línea de regresión estimada viene dada por

$$\hat{\mu}_{Y|x} = a + bx$$

El razonamiento que está detrás del método de los mínimos cuadrados es muy sencillo. De las muchas líneas rectas que pueden trazarse a través de la nube de puntos, conviene elegir aquella que «mejor se ajusta» a los datos. El ajuste es el mejor en el sentido de que los valores adoptados por αy b serán aquellos que minimicen la suma de los cuadrados de las distancias entre los datos puntuales y la línea de regresión ajustada. De esta forma, estamos determinando la línea recta que está tan próxima como sea posible a todos los datos puntuales simultáneamente.

La distancia entre un dato puntual y el valor estimado por la recta de regresión se denomina *residuo*. Los residuos se indican mediante e_1, e_2, e_3,..., e_n. Se ilustran en la Figura 11.8. Utilizando esta notación,

$$\begin{aligned} e_1 &= \text{diferencia entre el primer dato puntual y el estimado por la línea de regresión} \\ &= y_1 - (a + bx_1) \\ e_2 &= y_2 - (a + bx_2) \\ e_3 &= y_3 - (a + bx_3) \\ &\vdots \\ e_{11} &= y_{11} - (a + bx_{11}) \end{aligned}$$

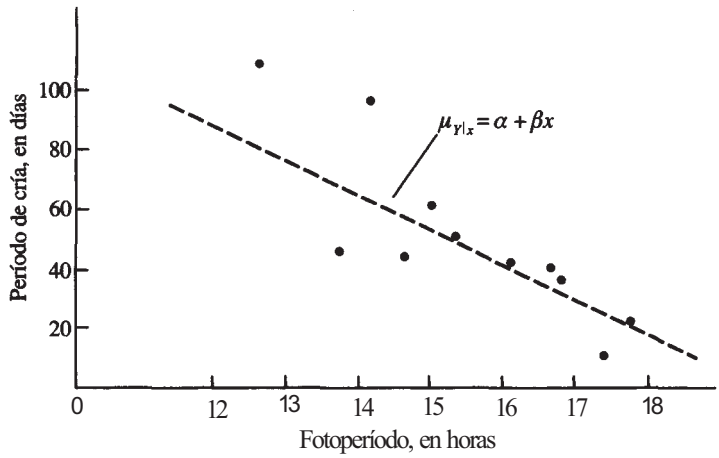


Figura 11.7. Línea de regresión teórica y curva ideal para predecir la duración del período de cría, desconocida por el experimentador, basándose en el fotoperíodo.

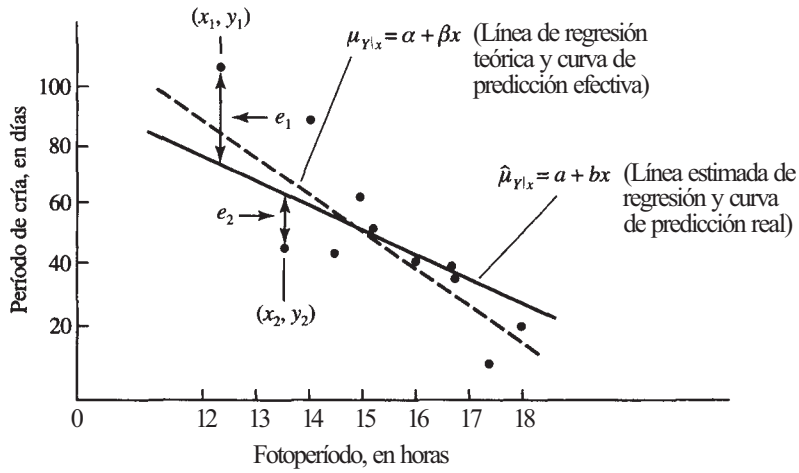


Figura 11.8. La curva de predicción efectiva minimiza la suma de los cuadrados de las distancias e_i, donde e_i es la distancia vertical del dato puntual (x_i, y_i) a la recta de regresión estimada.

La suma de los cuadrados de estas diferencias es:

$$\sum_{i=1}^{11} e_i^2 = \sum_{i=1}^{11} [y_i - (a + bx_i)]^2$$

Esta suma, indicada mediante SS_E, es una función tanto de α como de b. Es decir, su valor depende de los valores numéricos elegidos para la pendiente y el término independiente de la recta de regresión estimada. Deseamos elegir a yb de forma que SS_E sea lo más pequeña posible. Las técnicas de cálculo pueden utilizarse para demostrar que esto ocurre siempre que se satisfagan estas ecuaciones:

$$\sum_{i=1}^{11} y_i = 11a + b \sum_{i=1}^{11} x_i$$

$$\sum_{i=1}^{11} x_i y_i = a \sum_{i=1}^{11} x_i + b \sum_{i=1}^{11} x_i^2$$

A estas ecuaciones se les llama *ecuaciones normales.* Se resuelven para a y b, obteniéndose;

$$\hat{\beta} = b = \frac{11 \sum_{i=1}^{11} x_i y_i - \sum_{i=1}^{11} x_i \sum_{i=1}^{11} y_i}{11 \sum_{i=1}^{11} x_i^2 - \left[\sum_{i=1}^{11} x_i\right]^2}$$

$$\hat{\alpha} = a = \bar{y} - b\bar{x}$$

Así, para calcular a y b en este conjunto de datos, necesitaremos solamente calcular cuatro valores:

$$\sum_{i=1}^{11} x_i \quad \sum_{i=1}^{11} x_i^2 \quad \sum_{i=1}^{11} y_i \quad \sum_{i=1}^{11} x_i y_i$$

Para estos datos, estos valores son:

$$\sum_{i=1}^{11} x_i = 169.8 \qquad \sum_{i=1}^{11} y_i = 625$$

$$\sum_{i=1}^{11} x_i^2 = 2645.02 \qquad \sum_{i=1}^{11} x_i y_i = 9286.2$$

De este modo:

$$\hat{\beta} = b = \frac{11(9286.2) - 169.8(625)}{11(2645.02) - (169.8)^2} = -15.11$$

$$\hat{\alpha} = a = \frac{625}{11} - (-15.11)\frac{(169.8)}{(11)} = 290.06$$

La línea de regresión estimada y la ecuación de predicción real basada en estos datos es:

$$\hat{\mu}_{Y|x} = 290.06 - 15.11x$$

Para predecir la duración media de la época de cría, cuando el fotoperíodo bajo el cual se inició la reproducción es 14.5 horas, sustituimos $x = 14.5$ en la ecuación anterior. Así, nuestro valor predicho es:

$$\hat{\mu}_{Y|x} = 290.06 - 15.11(14.5) = 70.97 \text{ días}$$

Las ecuaciones normales y las estimadas para α y β dadas en el Ejemplo 11.2.1 se generalizan reemplazando el número 11, tamaño de la muestra para el conjunto de datos dado, por *n,* tamaño de la muestra para un conjunto de datos en general. De este modo, las ecuaciones normales para el modelo de regresión lineal simple son:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \qquad \text{(Ecuaciones normales)}$$

Las fórmulas generales para $\hat{\alpha}$ y $\hat{\beta}$ son:

$$\hat{\alpha} = a = \bar{y} - b\bar{x}$$

$$\hat{\beta} = b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \qquad \text{(Estimaciones para } \alpha \text{ y } \beta\text{)}$$

Debemos llamar la atención sobre un punto: un conjunto de datos proporciona una prueba de linealidad solamente sobre aquellos valores de X cubiertos por el conjunto de datos. Para valores de *X* fuera de los cubiertos no hay evidencia de lineahdad. Es peligroso, por tanto,

utilizar una línea de regresión estimada para predecir valores de Y correspondientes a valores de X que estén fuera del campo de los valores de X cubiertos por el conjunto de datos. Aclararemos este punto en el Ejemplo 11.2.2.

Ejemplo 11.2.2. Se obtuvieron los siguientes datos sobre los tiempos más rápidos «Y» (en segundos) invertidos por corredores de la carrera de la milla, en los campeonatos mundiales que tuvieron lugar entre 1954 y 1972:

X **(año)**	Y **(tiempo invertido)**
54	239.4 (Bannister rompe la barrera de los 4 minutos)
54	238.0
56	238.1
56	238.5
58	234.5
58	236.2
60	235.3
60	234.8
62	235.1
62	234.4
64	234.1
64	234.9
66	231.3 (Ryun)
66	232.7
68	231.4
68	231.8
70	232.0
70	231.9
72	231.4
72	231.5

En la Figura 11.9 se representa la nube de puntos de estos datos. Parecen tener una tendencia lineal. Para ajustar una línea de regresión se precisan los siguientes valores:

$$\sum_{i=1}^{20} x_i = 1260 \qquad \sum_{i=1}^{20} y_i = 4687.3$$

$$\sum_{i=1}^{20} x_i^2 = 80\,040 \qquad \sum_{i=1}^{20} x_i y_i = 295\,024.2$$

$$\bar{x} = 63 \qquad \bar{y} = 234.37$$

$$\hat{\beta} = b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{20(295\,024.2) - (1260)(4687.3)}{20(80\,040) - 1260^2} = -0.42$$

$$\hat{\alpha} = a = \bar{y} - b\bar{x} = 234.37 - (-0.42)(63) = 260.83$$

La línea de regresión estimada es:

$$\hat{\mu}_{Y|x} = 260.83 - 0.42x$$

Esta línea puede utilizarse sin problemas para predecir el tiempo invertido por un corredor dq la milla en un campeonato mundial celebrado entre 1954 y 1972, los años cubiertos por el conjunto de datos. El peligro está en utilizar esta ecuación para predecir el tiempo para la

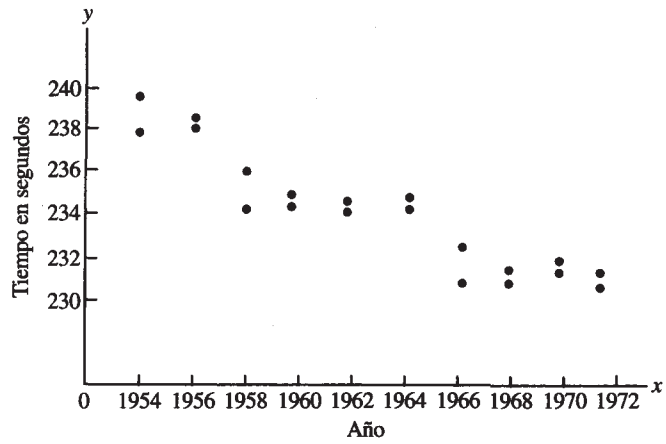


Figura 11.9. Nube de puntos de *años* frente a *tiempo empleado en la carrera de la milla* por campeones del mundo.

carrera de la milla mucho después de 1972, porque, a partir del conjunto de datos, no hay prueba de que la tendencia lineal continúe indefinidamente. En particular, supongamos que suponemos erróneamente que la tendencia lineal continúa y tratamos así de predecir el tiempo invertido en la carrera para un corredor en el año 2521. Para dicho año, $x = 621$. Así, el valor predicho de Y es:

$$\hat{\mu}_{Y|x} = 260.83 - 0.42(621) = 0.01 \text{ segundos}$$

¡Este valor es inconcebible como tiempo posible para correr una milla! El problema surge porque en algún punto después de 1972, cesa la tendencia lineal y los tiempos se mantienen en un determinado valor.

Estimando una respuesta individual

Al escribir la recta de regresión estimada como:

$$\hat{\mu}_{Y|x} = a + bx$$

estamos subrayando el hecho de que los puntos sobre la recta representan la respuesta *media* estimada cuando la variable predictora asume el valor x. Esta recta también puede utilizarse para predecir la respuesta misma. El sentido común nos dice que una elección lógica del valor predicho de Y para un valor x dado es su valor medio estimado cuando $X = x$. Por ejemplo, si nos pidieran que predijéramos la temperatura del agua oceánica en un punto concreto donde la profundidad es de 1000 pies, una elección lógica para esta predicción es la temperatura media estimada a esta profundidad. Si nos piden que predigamos la duración del tiempo de reproducción de un pato buceador en particular, que comenzó la puesta cuando el fotoperíodo era de 14.5 horas, nuestra mejor estimación es la duración media estimada bajo estas condiciones. En el Ejemplo 11.2.1 se ha hallado que este promedio estimado era de 70.97 días. Por lo tanto, es correcto escribir:

$$\hat{y} = \hat{\mu}_{Y|x} = 290.066 - 15.11(14.5) = 70.97$$

Al leer los resultados de un estudio de regresión hay que ser cuidadoso. La ecuación obtenida es la ecuación del valor medio de Y como función de x. Se utiliza para estimar tanto la respuesta media como la individual cuando el regresor asume el valor x.

Hemos proporcionado una forma lógica para estimar α y *β,* y así ajustar una línea recta a cualquier conjunto de datos. También hemos visto que una nube de puntos da una cierta idea sobre si es aplicable o no una regresión lineal. En todo caso, todavía hay que comprobar una cuestión importante: ¿Tiene realmente sentido la regresión *lineal* o se ajustaría más a los datos algún otro tipo de curva, proporcionando así una mejor ecuación de predicción? Esta pregunta no puede ignorarse. Nos dedicamos a ella en la Sección 11.4.

Nota sobre los cálculos

La regresión lineal simple es una técnica estadística muy útil. Por esta razón, muchas calculadoras manuales están programadas para efectuar análisis de regresión introduciendo simplemente los pares *(x,* y). Consulte el manual de su calculadora para ver si tiene esta capacidad. La sección de Herramientas Computacionales, al final del capítulo, mostrará cómo utilizar la TI83 para estimar una recta de regresión.

El análisis de regresión para grandes conjuntos de datos se realiza, habitualmente, por ordenador. En el Ejemplo 11.2.3 se trata el análisis por ordenador de los datos del Ejemplo 11.2.1. El código SAS utilizado para generar el *printout* se muestra en la sección de Herramientas Computacionales de este capítulo.

Ejemplo 11.2.3. Para comenzar un análisis por ordenador de un conjunto de datos, en primer lugar le pedimos al ordenador que dibuje la nube de puntos de los datos. Si la regresión lineal simple es apropiada, la nube de puntos presentará una visible tendencia lineal. La Figura 11.10 muestra la nube de puntos generada mediante el S AS de los datos del Ejemplo 11.2.1. La respuesta, duración de la temporada de reproducción, está representada sobre el eje vertical; el regresor, el fotoperíodo, está representado sobre el eje horizontal. Existe una visible tendencia lineal descendente, indicando que, a medida que aumenta el número de horas de luz por día, disminuye la duración del tiempo de reproducción. La pendiente de la recta de regresión estimada será negativa. La Figura 11.11 da la salida del programa utilizado para estimar el término independiente y la pendiente de la recta de regresión. Estas vienen dadas por ① y (2), respectivamente. Obsérvese que la pendiente estimada es negativa, tal como se esperaba. La recta de regresión estimada es:

$$\hat{\mu}_{Y|x} = 290.07044608 - 15.11057071x$$

Esto concuerda, al margen de las diferencias de redondeo, con la ecuación hallada anteriormente. La duración media estimada del tiempo de reproducción, cuando la reproducción comienza durante las 14.5 horas de luz, se muestra en ③.

EJERCICIOS 11.2

Nota: Aunque en esta sección se han dado las fórmulas para calcular *a y b,* los problemas deberían ser resueltos utilizando un ordenador o una calculadora estadística cuando sea posible.

1. Para los datos del Ejercicio 1 de la Sección 11.1, hallar $\sum x, \sum x^2, \sum y, \sum xy, \bar{x}, \bar{y}$. Utilizar estas cantidades para estimar α y *β.* Escribir la ecuación de la línea de regresión estimada y emplearla para predecir el valor medio de Y cuando $x = 3.7$; así como el mismo valor de *Y* cuando $x = 3.7$.
2. Utilizar la ecuación de regresión obtenida en el Ejemplo 11.2.1 para predecir la duración de la época de reproducción si ésta se inició cuando había catorce horas de luz por día. ¿Puede emplearse esta ecuación con cierta seguridad para estimar la duración de la época de reproducción si ésta se inició cuando había diez horas de luz por día? Explicarlo.

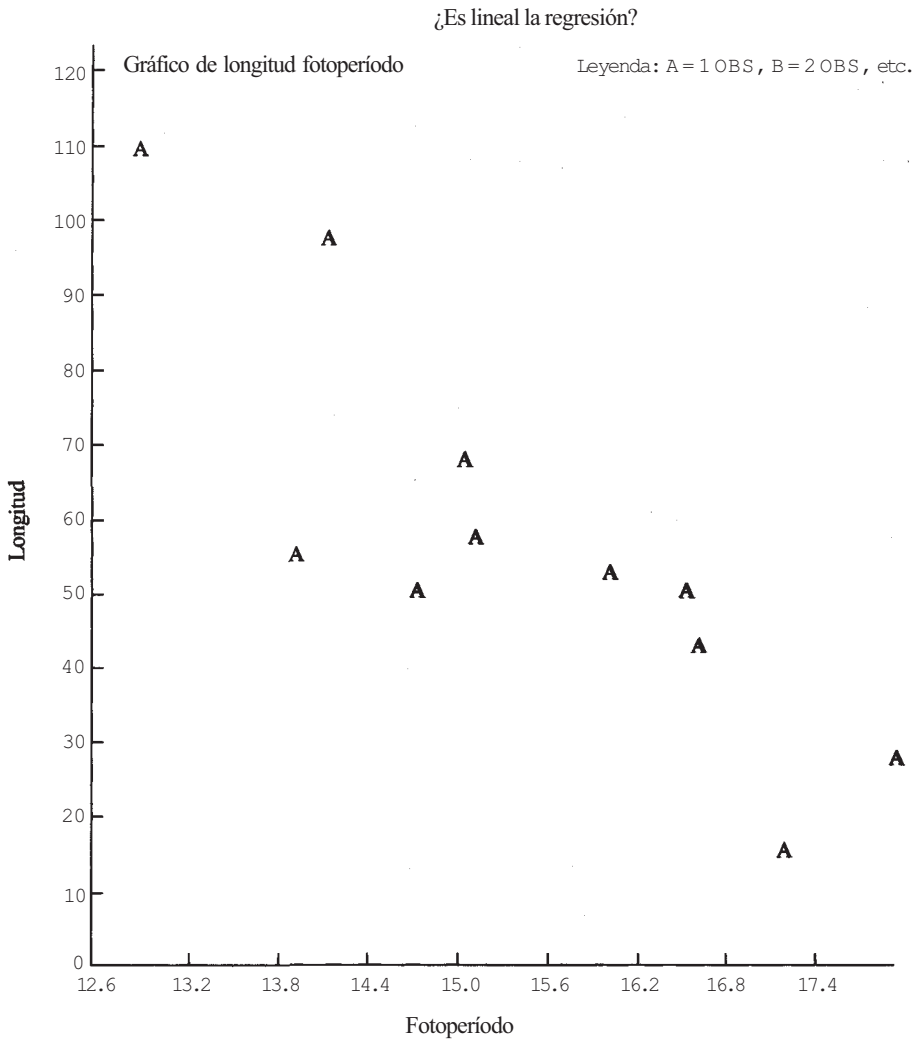


Figura 11.10. Nube de puntos para los datos del Ejemplo 11.2.1. Los datos presentan una tendencia lineal descendente.

3. Se ha establecido un índice numérico del grado de enfermedad de pacientes que sufren la enfermedad de Crohn. El índice requiere que el paciente lleve un diario e incluya información sobre ocho variables clínicas. El índice, si bien útil, es molesto de obtener en la práctica y-se ha ideado un nuevo índice que es más fácil de calcular. Se cree que los valores obtenidos con el nuevo índice pueden ser utilizados para predecir el valor que se habría obtenido utilizando el antiguo índice ya comprobado. Se evaluó a ciento seis pacientes utilizando ambos índices. Los valores de X recorren de 0.5 a 14.0. La nube de puntos para los datos muestra una tendencia lineal. Se tiene:

$$\sum x = 366.1 \qquad \sum y = 12\,623$$
$$\sum x^2 = 2435.63 \qquad \bar{y} = 119.08$$
$$\bar{x} = 3.45 \qquad \sum xy = 75\,989.6$$

```
                          IS THE REGRESSION LINEAR?
                        GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE:  LENGTH
SOURCE                    DF    SUM OF      SQUARES      MEAN SQUARES       F VALUE
MODEL                      1    5462.88341888         5462.88341888         23.86
ERROR                      9    2060.75294475          228.97254942         PR > F
CORRECTED TOTAL           10    7523.63636364                               0.0009

R-SQUARE              C.V.           STD DEV          LENGTH MEAN
0.726096           26.6320        15.13183893          56.81818182

SOÜRCE                    DF       TYPE I SS     F  VALUE   PR > F
PHOTOPRD                   1    5462.88341888       23.86    0.0009

SOÜRCE                    DF       TYPE IV SS    F  VALUE   PR > F
PHOTOPRD                   1    5462.88341888       23.86    0.0009

                                              T FOR H0:   PR > |T|      STD ERROR OF
PARAMETER              ESTIMATE           PARAMETER = 0                   ESTIMATE
INTERCEPT        290.07044608 (1)                6.05       0.0002       47.97110790
PHOTOPRD         -15.11057071 (2)               -4.88       0.0009        3.09358185

                          IS THE REGRESSION LINEAR?

                OBS     PHOTOPRD      LENGTH      PREDICT

                  1         12.8         110      96.6551
                  2         13.9          54      80.0335
                  3         14.1          98      77.0114
                  4         14.7          50      67.9451
                  5         15.0          67      63.4119
                  6         15.1          58      61.9008
                  7         16.0          52      48.3013
                  8         16.5          50      40.7460
                  9         16.6          43      39.2350
                 10         17.2          15      30.1686
                 11         17.9          28      19.5912
                 12         14.5                  70.9672
```

Figura 11.11. Salida de un programa SAS, utilizado para estimar la recta de regresión y aproximar la duración media del tiempo de reproducción cuando ésta comienza en el momento en que hay 14.5 horas de luz al día.

Manejar esta información para estimar α, β y $\mu_{Y|x}$. ¿Cuál es la puntuación estimada mediante el índice antiguo para un paciente cuya puntuación por el nuevo índice es 5.5? ¿Se puede predecir con cierta seguridad la puntuación mediante el índice antiguo de un paciente con una puntuación de $x = 16$ por el nuevo índice? Razonar la respuesta.

4. Se realiza un estudio para establecer una ecuación mediante la cual se pueda utilizar la concentración de estrona en la saliva, para predecir la concentración del esteroide en plasma libre. Se extrajeron los siguientes datos de 14 varones sanos:

x (concentración de estrona en saliva, pg/mL)	***y* (concentración de estrona en plasma libre, pg/mL)**
7.4	30.0
7.5	25.0

x (concentración de estrona en saliva, pg/mL)	y (concentración de estrona en plasma libre, pg/mL)
8.5	31.5
9.0	27.5
9.0	39.5
11.0	38.0
13.0	43.0
14.0	49.0
14.5	55.0
16.0	48.5
17.0	51.0
18.0	64.5
20.0	63.0
23.0	68.0

a) Dibujar la nube de puntos.
b) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum xy$.
c) Estimar α, β, $\mu_{Y|x}$.
d) Utilizar la línea de regresión estimada para predecir el nivel de estrona en plasma libre de un varón cuyo nivel de estrona en saliva es de 17.5 pg/mL.

5. Utilizar los datos del Ejemplo 11.1.3 para:
   a) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum xy$.
   b) Estimar α, β, $\mu_{Y|x}$.
   c) Utilizar la línea estimada de regresión para predecir la longitud de la concha de una lapa cuya altura de concha es 2.25 milímetros.
6. Medir la capacidad de una persona para soplar una vela es una forma ordinaria de evaluar las velocidades respiratorias máximas. Se realiza la prueba de la vela sosteniendo una vela encendida, perpendicular a un tablero de madera colocado sobre una bandeja ajustable, de forma que la llama esté a la misma altura que la boca del sujeto. La vela se coloca a 5 cm del sujeto y se le permiten tres intentos para apagarla. Si el sujeto tiene éxito, se desplaza la vela 5 cm hacia atrás y se repite el experimento. Entre los intentos, se le permite un período de descanso. La variable predictora es *X,* la distancia más lejana a la que puede apagarse la vela. Las respuestas son:

$$\begin{aligned} \text{FVC} &= \text{capacidad vital forzada} \\ \text{FEV}_1 &= \text{volumen de espiración forzada en 1 segundo} \\ \text{PEFR} &= \text{flujo máximo espiratorio} \\ \text{FET} &= \text{tiempo de espiración forzada} \end{aligned}$$

Se obtuvieron las siguientes ecuaciones de regresión estimadas:

$$\begin{aligned} \widehat{\text{FVC}} &= 0.04039x + 0.9606 \\ \widehat{\text{FEV}_1} &= 0.037659x + 0.4983 \\ \widehat{\text{PEFR}} &= 4.3379x + 195.5 \\ \widehat{\text{FET}} &= -0.071331x + 9.5591 \end{aligned}$$

a) ¿Cuál es la respuesta media estimada para cada una de estas variables en pacientes para los que la distancia máxima a la que pueden apagar la vela es de 10 cm?

b) Se observa a un paciente y se determina que la distancia máxima a la que puede apagar la vela es 10 cm. ¿Cuáles son los valores estimados para cada respuesta de este paciente?

c) Si se representaran estas líneas, ¿qué línea se inclinaría hacia arriba formando el ángulo más agudo?

d) Si se representaran estas líneas, ¿qué línea se inclinaría hacia abajo?

(Ecuaciones de regresión halladas en Bayur Teklu et al., «The Match Test Revisited-Blowing Out a Candle as a Screening Test for Airflow Obstruction», *Journal of Family Practice,* noviembre de 1990, págs. 557-562.)

7. Se lleva a cabo un estudio para investigar el efecto de la descomposición de las agujas del pino en el pH del suelo. Se añadieron diversas cantidades de agujas de pino a una muestra de suelo y se dejó que se descompusiesen. Se obtuvo el pH del suelo resultante, con los siguientes resultados:

Agujas de pino (g por 170 g de suelo)	pH	Agujas de pino (g por 170 g de suelo)	pH
5	7.50	15	6.75
5	7.30	15	6.73
5	7.00	15	6.70
5	**6.95**	15	6.68
5	6.88	15	6.70
10	6.85	20	6.55
10	6.83	20	6.65
10	6.80	20	6.63
10	6.78	20	6.62
10	**6.80**	20	6.60

a) Dibujar la nube de puntos para estos datos.

b) Si parece apropiada la regresión lineal, estimar la recta de regresión

c) Basándose en la ecuación de regresión estimada, ¿cuál es el pH medio de las muestras de suelo sometidas a 18 g de agujas de pino? ¿Cuál es el pH estimado para una muestra individual de suelo sometida a 18 g de agujas de pino?

(Basado en un estudio realizado por Amy Payne, Departamento de Biología, Universidad de Radford, 1996.)

8. Se realiza un estudio sobre el efecto de la intensidad lumínica en la tasa de fotosíntesis de hojas de espinaca. El estudio se llevó a cabo utilizando discos de hojas a las que se habían realizado agujeros. Las hojas son sumergidas en agua. La intensidad lumínica es medida en candelas-pie, la tasa de fotosíntesis se mide anotando el tiempo en segundos que tardan en flotar la mitad de las hojas. Una rápida subida de las hojas indica una alta tasa de fotosíntesis. Se obtuvieron estos datos:

Intensidad (candelas-pie)	**Tiempo en flotar** (segundos)	**Intensidad (candelas-pie)**	**Tiempo en flotar (segundos)**
400	4500	800	2200
400	4450	800	2000
400	4200	800	1800
400	3900	800	1500

Intensidad (candelas-pie)	Tiempo en flotar (segundos)	Intensidad (candelas-pie)	Tiempo en flotar (segundos)
400	3700	800	1400
600	3100	1000	1200
600	2800	1000	1000
600	2750	1000	980
600	2600	1000	990
600	2500	1000	995

a) Dibujar una nube de puntos para los datos.
b) Si es apropiado, estimar la recta de regresión.
c) Basándose en la recta de regresión obtenida, estimar la tasa media de fotosíntesis para las muestras sometidas a una intensidad de 650 candelas-pie. Estimar la tasa fotosintética para una muestra individual sometida a una intensidad de 650 candelas-pie.

(Basado en un estudio de Shane Bryant, Departamento de Biología, Universidad de Radford, 1996.)

11.3. INTRODUCCIÓN A LA CORRELACIÓN

Recuérdese que el análisis de regresión estadística aborda la relación entre una variable independiente X y la media de una variable dependiente Y. No estamos interesados en extraer conclusiones sobre X, puesto que esta variable sólo nos interesa debido al hecho de que ayuda a estimar la respuesta media $\mu_{Y|x}$, o a predecir una respuesta individual $Y|x$. En la determinación de la regresión, la variable predictora X *no* es aleatoria, mientras que la respuesta Y es una variable aleatoria. Hemos estado particularmente interesados en situaciones en las que la relación entre x y $\mu_{Y|x}$ es lineal. Nuestra atención ha estado centrada en estimar la recta:

$$\mu_{Y|x} = \alpha + \beta x$$

de forma que la ecuación obtenida se puede utilizar para estimar respuestas futuras y respuestas medias para valores de X dados.

En el análisis de correlación, tanto X como Y son variables aleatorias y tienen el mismo interés. Deseamos determinar si existe o no una relación lineal entre estas dos variables aleatorias. Puesto que deseamos responder a la pregunta: ¿es

$$Y = \alpha + \beta X$$

para algunos parámetros α y β donde $\beta \neq 0$? Vamos a desarrollar un parámetro que mida la fuerza de la asociación lineal que existe entre X e Y.

La medida de asociación lineal más frecuentemente utilizada entre dos variables aleatorias es ρ, el coeficiente de correlación momento-producto de Pearson. Este parámetro se define en términos de covarianza entre X e Y, que es una medida de la forma en que X e Y varían conjuntamente. La definimos del siguiente modo:

Definición 11.3.1. Covarianza. Sean X e Y variables aleatorias con medias μ_X y μ_Y respectivamente. La *covarianza* entre X e Y, designada por Cov (X, Y), viene dada por:

$$\text{Cov}\,(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y]$$

Obsérvese que si los valores pequeños de X tienden a estar asociados con valores pequeños de y Y los valores grandes de X con valores grandes de Y, entonces $X - \mu_X$ e $Y - \mu_Y$ tenderán a tener el mismo signo algebraico. Esto implica que $(X - \mu_X)(Y - \mu_Y)$ tenderá a ser positivo produciendo una covarianza positiva. Si es cierto lo contrario y los valores pequeños de X tienden a ser asociados con valores grandes de Y y viceversa, entonces $X - \mu_X$ e $Y - \mu_Y$ tenderán a tener signos algebraicos opuestos. Ello produce la tendencia de $(X - \mu_X)(Y - \mu_Y)$ a ser negativo, proporcionando una covarianza negativa.

Es evidente que podemos decir algo sobre la relación entre X e Y a partir del signo algebraico de la covarianza. Sin embargo, este parámetro no está acotado. Puede asumir cualquier valor real, y su magnitud no tiene significado. Para corregir este problema, dividimos la covarianza por $\sqrt{(\text{Var } X)(\text{Var } Y)}$ para formar el coeficiente de correlación de Pearson presentado en la Definición 11.3.2.

Definición 11.3.2. Coeficiente de correlación de Pearson. Sean X eY variables aleatorias con medias μ_X y μ_Y y varianzas σ_X^2 y σ_Y^2, respectivamente. La *correlación p* entre X e Y es

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{(\text{Var } X)(\text{Var } Y)}}$$

A pesar de que la prueba está fuera del objetivo de este texto, se puede demostrar que ρ se sitúa entre -1 y +1, ambos inclusive. Si $p = 1$, decimos que existe una *correlación positiva perfecta* entre XeY. Los valores grandes de X están asociados con valores grandes de Y, y los valores pequeños de X están asociados con valores pequeños de Y. En este caso, los puntos (x, y) se situarán en una línea recta con pendiente ascendente tal como se muestra en la Figura 11.1*2a.* Si $\rho = -1$, decimos que existe una *correlación negativa perfecta* entre X e Y. Aquí, los valores grandes de X están asociados con valores pequeños de Y, y los valores pequeños de X están asociados con valores grandes de Y. Los puntos se situarán en una línea recta con pendiente descendente, tal como se muestra en la Figura 11.1*2b.* Si $p = 0$, decimos que XeY *no están correlacionados.* Esto sólo significa que no existe asociación *lineal* entre X e Y. No significa que X e Y no estén relacionados. Sin embargo, ello implica que si existe una relación entre estas dos variables aleatorias, ésta no será lineal. Las Figuras 11.12c y *d* ilustran dos casos en los que $\rho = 0$. En la parte c no existe una relación obvia entre X e F.En la parte *d,* hay una relación aparente, pero no es lineal.

Ejemplo 11.3.1. Considérense la variable aleatoria X, altura de un varón adulto, e Y, su peso. Puesto que las personas más altas tienden a pesar más que las de baja estatura, podemos esperar que X e Y estén positivamente correlacionadas.

Ejemplo 11.3.2. Sea X la altitud e Y el índice de diversidad del arbolado. Puesto que el número de especies de árboles que pueden sobrevivir a grandes altitudes es bastante pequeño, a medida que aumenta la altitud debería disminuir el número de especies halladas. Por lo tanto, los valores grandes de X deberán asociarse con valores pequeños de Y y viceversa. La correlación entre estas dos variables aleatorias será negativa.

Estimación de ρ

Téngase en cuenta que Cov (X, Y) *y p* son parámetros *teóricos.* Ni uno ni otro pueden calcularse sin el conocimiento de la distribución de probabilidad del par de variables (X, Y). El problema estadístico es estimar sus valores a partir de un conjunto de datos. Puesto que Cov (X, Y)

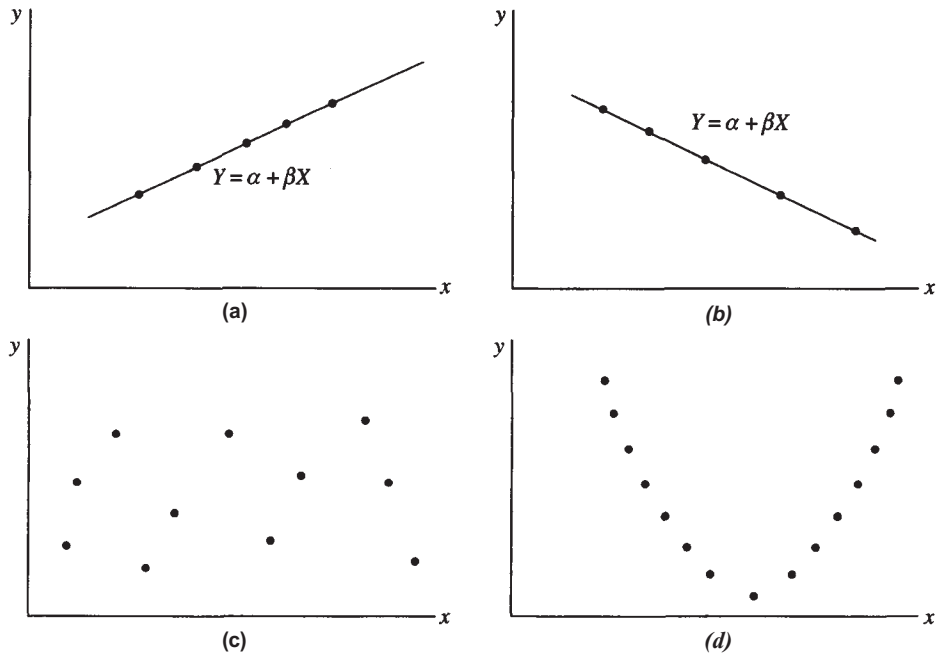


Figura 11.12. (a) Correlación positiva perfecta: $\rho = 1$, $\beta > 0$, todos los puntos están sobre una línea recta con pendiente positiva; *ib)* correlación negativa perfecta: $\rho = -1$, $\beta < 0$, todos los puntos están sobre una línea recta con pendiente negativa; (c) no correlacionados: $\rho = 0$, los puntos están aleatoriamente dispuestos; (d) no correlacionados: $\rho = 0$, los puntos indican una relación entre X e Y, pero no lineal.

puede expresarse como diferencia entre las medias teóricas $E[XY]$ y $E[X]E[Y]$, es posible estimarla fácilmente reemplazando cada media teórica por su correspondiente media muestral. De este modo estimamos:

$$E[XY] \quad \text{mediante} \quad \frac{\sum_{i=1}^{n} x_i y_i}{n}$$

$$E[X] \quad \text{mediante} \quad \frac{\sum_{i=1}^{n} x_i}{n}$$

$$E[Y] \quad \text{mediante} \quad \frac{\sum_{i=1}^{n} y_i}{n}$$

Sustituyendo, la covarianza estimada se convierte en:

$$\widehat{\text{Cov}\,(X, Y)} = \sum_{i=1}^{n} \frac{x_i y_i}{n} - \sum_{i=1}^{n} \frac{x_i}{n} \sum_{i=1}^{n} \frac{y_i}{n}$$

$$= \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2}$$

Igualmente, puesto que $\text{Var}\, X = E[X^2] - (E[X])^2$,

$$\widehat{\text{Var}\, X} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left[\sum_{i=1}^{n} \frac{x_i}{n}\right]^2$$

$$= \frac{n\sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2}{n^2}$$

Combinando estos resultados, obtenemos una estimación lógica para el coeficiente de correlación ρ:

$$\widehat{\rho} = \frac{\dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2}}{\sqrt{\left[\dfrac{n\sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2}{n^2}\right]\left[\dfrac{n\sum_{i=1}^{n} y_i^2 - \left[\sum_{i=1}^{n} y_i\right]^2}{n^2}\right]}}$$

Se cancelan los términos n^2 para obtener la expresión de $\widehat{\rho}$ dada en la Definición 11.3.3.

Definición 11.3.3. Estimación para *p.* La *estimación para* ρ, coeficiente de correlación de Pearson, a la que se designa r, es:

$$\widehat{\rho} = r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

El Ejemplo 11.3.3 nos ilustra el cálculo de *r.*

Ejemplo 11.3.3. Los investigadores están estudiando la correlación entre la obesidad y la respuesta individual al dolor. La obesidad se mide como porcentaje sobre el peso ideal *(X).* La respuesta al dolor se mide utilizando el umbral de reflejo de flexión nociceptiva (*Y),* que es una medida de sensación de punzada. Obsérvese que ambas, *X* e *Y,* son variables aleatorias. Queremos estimar ρ, el coeficiente de correlación para estas variables. Se obtienen los siguientes datos:

x (porcentaje de sobrepeso)	*y* (umbral de reflejo de flexión nociceptiva)
89	2
90	3
75	4
30	4.5
51	5.5
75	7
62	9
45	13
90	15
20	14

En la Figura 11.13 se dibuja la nube de puntos. Parece haber alguna tendencia a que los valores pequeños de X estén asociados con valores grandes de Y, y viceversa. Sin embargo, no es una tendencia fuerte, como lo demuestran los puntos (30,4.5), en los que se empareja un valor pequeño de X con un valor pequeño de Y, y (20, 14), para el que ocurre lo contrario. Basándonos en estos comentarios, esperaremos que r, la estimación par ρ, sea negativa; pero no que esté muy próxima a - 1 . Es decir, esperaremos que X e Y presenten una *ligera* correlación negativa. Para verificar estas observaciones, estimamos ρ utilizando la Definición 11.3.3. Se necesitan los siguientes estadísticos muestrales:

$$\sum x = 627 \qquad \sum y = 77 \qquad \sum xy = 4461.5$$
$$\sum x^2 = 45\,141 \qquad \sum y^2 = 799.5$$

$$\hat{\rho} = r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$
$$= \frac{10(4461.5) - 627(77)}{\sqrt{[10(45\,141) - 627^2][10(799.5) - 77^2]}}$$
$$= -0.33$$

Conviene hacer una llamada de atención. Hemos proporcionado un estimador puntual lógico para ρ y un método un poco burdo de interpretar el valor estimado. Todavía hay que resolver un problema. Puesto que estimamos p a partir de un conjunto de datos, es improbable que $\hat{\rho}$ tome siempre los valores fácilmente interpretables de 1, -1 ó 0. Lo habitual es que tengamos que hacer frente al problema de interpretar un valor tal como el obtenido en el Ejemplo 11.3.3 (-0.33), que no es claramente próximo a ninguno de los valores extremos 1, - 1 , ó 0.

La Figura 11.14 proporciona una escala sugerida para interpretar r. De acuerdo con la escala, la correlación -0.33 se describe como la correlación negativa «débil». La escala se justificará en la Sección 11.4. Al leer la escala, se asumirá que los valores que son iguales a uno de los puntos de corte de la escala se sitúan en la clasificación superior. Por ejemplo, una correlación de 0.5 se considerará una correlación positiva moderada, mientras que una correlación de -0.9 se considera una correlación negativa fuerte. La escala presentada aquí no es «la ley» («la Biblia»), sólo es una interpretación sugerida. Los coeficientes de correlación son, en cierta medida, dependientes del objeto en cuestión. En experimentos de laboratorio biológico

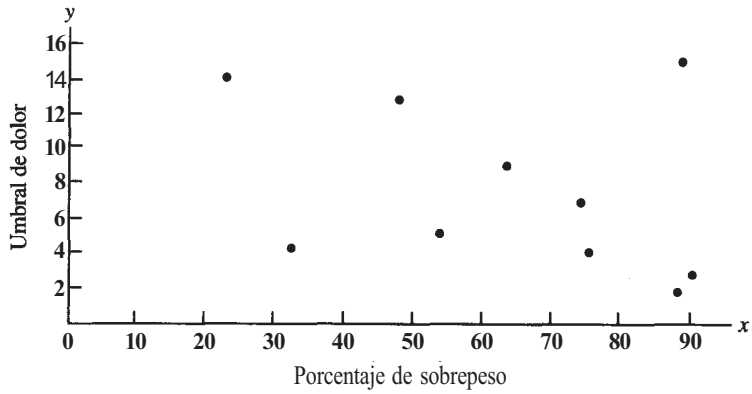


Figura 11.13. Nube de puntos del porcentaje de sobrepeso frente al umbral de reflejo de flexión nociceptiva, indicando una ligera correlación negativa.

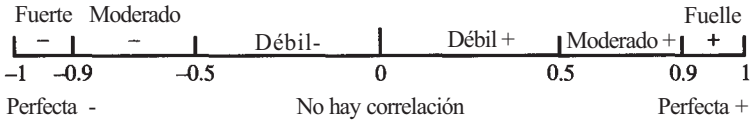


Figura 11.14. Interpretación del coeficiente de correlación de Pearson, estimando r.

o químico controlados cuidadosamente, se puede esperar que los coeficientes de correlación sean bastante altos. Sin embargo, en experimentos con seres humanos o en estudios de campo observacionales, habitualmente se observan coeficientes de correlación más bajos. Estos valores más bajos aún pueden ser considerados altamente informativos para el experto en la materia.

Nota sobre los cálculos

El cálculo manual de *r* es tedioso, por lo que la mayoría de calculadoras manuales están programadas para hallar *r.* Se hace necesario utilizar una calculadora para realizar los cálculos aritméticos de la Definición 11.3.3. En la sección de Herramientas Computacionales, al final del capítulo, se desarrollará el procedimiento correspondiente utilizando la calculadora TI83 y SAS. El código utilizado para hacer la Figura 11.15 se muestra allí.

Los paquetes informáticos que calculan *r* generalmente incluyen un contraste de:

H_0: $\rho = 0$ *(X* e *Y* no están correlacionados)

H_1: $\rho \neq 0$ (existe correlación entre *X* e *Y)*

Este contraste debe verse con precaución. Sólo comprueba si *X* e *Y* están o no correlacionados. De ninguna forma contrasta si la correlación que existe tiene alguna importancia práctica. Para un conjunto de datos grande, se puede comprobar que una correlación de 0.05 es diferente de cero. Sin embargo, como hemos visto, esta correlación se considera débil. En la sección siguiente explicaremos lo débil que es. Verá en artículos de investigación que ciertas correlaciones son «estadísticamente significativas». Recuerde el hecho de que esto generalmente significa que se ha realizado el contraste anterior y se ha rechazado H_0: la correlación no es cero. Sin embargo, usted puede juzgar por sí mismo desde el punto de vista del conteni-

```
                    ARE PAIN AND OBESITY CORRELATED

VARIABLE   N          MEAN        STD DEV           SUM        MINIMUM        MAXIMUM

PERCENT   10    62.70000000   25.44733123   627.0000000   20.00000000   90.00000000
PAIN      10     7.70000000    4.79119563    77.0000000    2.00000000   15.00000000

          CORRELATION COEFFICIENTS/PROB > |R| UNDER H0:RH0 = 0/N = 10

                                  PERCENT          PAIN

                   PERCENT        1.00000       -0.33391 (1)
                                  0.0000         0.3457 (2)

                   PAIN          -0.33391        1.00000
                                  0.3457         0.0000
```

Figura 11.15. Salida de SAS para hallar el coeficiente de correlación de los datos del Ejemplo 11.3.3. No se rechaza la hipótesis nula de no correlación.

do si la correlación tiene o no algún uso práctico. El *printout* que se muestra en la Figura 11.15 da la salida de SAS del programa utilizado para hallar el coeficiente de correlación de los datos del Ejemplo 11.3.3. El coeficiente de correlación, -0.33391, se muestra en ①. El valor *P* para el contraste:

$$H_0: \rho = 0 \quad (X \text{ e } Y \text{ no están correlacionados})$$

$$H_1: \rho \neq 0 \quad (\text{existe correlación entre } X \text{ e } Y)$$

se muestra en ②. Dado que este valor P(0.3457) es alto, no rechazaremos H_0. La correlación negativa débil no es lo suficientemente grande como para poder concluir que, de hecho, existe una asociación lineal entre el umbral de reflexión nociceptiva y el porcentaje de sobrepeso.

EJERCICIOS 11.3

1. Considérense las siguientes observaciones sobre las variables aleatorias *X* e *Y:*

x	*y*
2.0	**5.0**
2.5	**5.5**
3.0	**6.2**
3.5	**6.4**
4.0	**7.0**

   a) Dibujar la nube de puntos.
   b) Basándose en la nube de puntos, ¿se puede esperar que r, el coeficiente de correlación estimado, esté próximo a 1,-1 ó 0?
   c) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$.
   d) Hallar *r* y clasificarlo según la escala de la Figura 11.14.

2. Considérense las siguientes observaciones sobre las variables aleatorias X e Y:

x	*y*
2.0	7.2
2.5	7.0
3.0	6.5
3.5	6.0
4.0	5.3

   a) Dibujar la nube de puntos.
   b) Basándose en la nube de puntos, ¿se puede esperar que *r* esté próximo a 1, -1 o 0?
   c) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$.
   d) Determinar *r* y clasificarlo según la escala de la Figura 11.14.

3. Considérense las siguientes observaciones sobre las variables aleatorias X e Y:

x	y
2.0	4.0
2.1	4.4
2.5	6.3
3.0	9.0
3.5	6.2
3.9	4.3
4.0	4.0

a) Dibujar la nube de puntos.
b) Basándose en la nube de puntos, ¿se puede esperar que r esté próximo a 1, -1 ó 0?
c) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$.
d) Calcular r.

4. Se realiza un estudio para estimar la correlación entre las variables aleatorias X, el valor de un cierto índice de obesidad para cada individuo, e Y, la tasa metabólica en reposo dé cada individuo. Respecto al índice de obesidad, un valor elevado indica un alto grado de obesidad; la tasa metabólica se mide en mililitros de oxígeno consumidos por minuto. Se mide cada variable sobre 43 sujetos. Se obtienen estos estadísticos muestrales:

$$\sum x = 1482.5 \qquad \sum y = 10\,719 \qquad \sum xy = 379\,207.5$$
$$\sum x^2 = 53\,515.25 \qquad \sum y^2 = 2736.063$$

Hallar r.

5. Para estudiar el efecto de las aguas residuales de las alcantarillas que afluyen a un lago, se toman medidas de la concentración de nitrato en el agua. Para monitorizar la variable, se ha utilizado un antiguo método manual. Se idea un nuevo método automático. Si se pone de manifiesto una alta correlación positiva entre las medidas tomadas empleando los dos métodos, se hará uso habitual del método automático. Los datos obtenidos son los siguientes (las unidades son microgramos de nitrato por litro de agua):

x (manual)	y (automático)
25	30
40	80
120	150
75	80
150	200
300	350
270	240
400	320
450	470
575	583

a) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$.
b) Estimar ρ.
c) ¿Aconsejaría poner en uso el método automático? Explicarlo.

6. Se ha realizado un estudio para evaluar la precisión con que las madres pueden juzgar el consumo de alimentos de sus hijos. Se obtuvieron datos de las madres y de un observador externo que pasó mucho tiempo observando la preparación de los alimentos y los hábitos nutritivos del niño. Entre el informe de la madre y el del observador se hallaron estas correlaciones:

Alimento	**r**	**Alimento**	***r***
Grasas, g	0.52	Calcio, mg	0.28
Grasas saturadas, g	0.38	Fósforo, mg	**-0.10**
Grasas monoinsaturadas, g	0.41	Hierro, mg	**0.82**
Grasas poliinsaturadas, g	0.58	Vitamina A, UI	**0.66**
Colesterol, mg	0.65	Tiamina, mg	**0.68**
Proteínas, g	0.50	Riboflavina, mg	0.52
Hidratos de carbono, g	0.48	Niacina, mg	0.72
Sodio, mg	0.27	Vitamina C, mg	**0.50**
Potasio, mg	0.54	Calorías	**0.71**

a) Explicar en un sentido práctico el significado de la correlación negativa del fósforo.
b) Categorizar cada correlación utilizando la escala de la Figura 11.14.
(Correlación registrada en Charles Basch et al., «Validation of Mothers' Reports of Dietary Intake by Four to Seven Year-Old Children», *American Journal of Public Health,* noviembre de 1990, págs. 1314-1317.)

7. Se realiza un estudio para investigar la depresión en los adolescentes. Entre los factores considerados, están la preocupación y la satisfacción con el entorno inmediato. Las puntuaciones altas indican altos niveles de depresión, preocupación o satisfacción. Se hacen las siguientes afirmaciones:

> «La depresión está positivamente correlacionada con la preocupación, $r = 0.3$, $P < 0.001$.»
> «La depresión está negativamente correlacionada con la satisfacción, $r = 0.36$, $P < 0.001$.»
> «Las puntuaciones de la satisfacción y la preocupacion están correlacionadas negativamente, $r = -0.16$, $P < 0.02$.»

a) Construya nubes de puntos para ilustrar la forma en que usted cree que aparecerán los datos en cada caso.
b) Un amigo, que no sabe nada sobre estadística, le pide que interprete estas afirmaciones en un sentido práctico. ¿Qué diría usted?
c) Observe que cada afirmación viene acompañada por un valor *P.* Indique las hipótesis nula y alternativa implicadas.
(Afirmaciones halladas en Lirio Covey y Debbie Tam, «Depressive Mood, The Single Parent Home and Adolescent Cigarette Smoking», *American Jaurnal of Public Health,* noviembre de 1990, págs. 1330-1333.)

11.4. EVALUACIÓN DE LA CONSISTENCIA DE LA RELACIÓN LINEAL (OPCIONAL)

Como hemos puesto de manifiesto en la Sección 11.2, el método de los mínimos cuadrados puede utilizarse para ajustar una línea recta a *cualquier* conjunto de datos. Sin embargo, la

utilidad de esta línea como pronóstico de valores futuros de la variable dependiente Y está condicionada por completo a que el *supuesto de linealidad sea apropiado.* Por esta razón, necesitamos un método analítico para determinar la bondad del ajuste de la línea recta a los lados puntuales. Presentamos dos métodos. El primero utiliza un estadístico llamado coeficiente de determinación; el segundo emplea una técnica de análisis de la varianza para descubrir si la variabilidad de Y está bien explicada por medio de una relación lineal con X.

Coeficiente de determinación

El coeficiente de determinación estimado es un estadístico que se utiliza para evaluar la fuerza de la relación lineal existente entre X e Y, tanto en una determinación de regresión como de correlación. Tiene una interpretación fácilmente comprensible y nos permitirá justificar la escala utilizada para interpretar el valor r dado en la última sección.

Puesto que este coeficiente está asociado tanto con el análisis de regresión como con la correlación, proporciona una conexión entre los dos procedimientos. Para definir este estadístico debemos determinar la relación entre r, coeficiente de correlación estimado, y b pendiente estimada de la línea de regresión mínimo cuadrada. Utilizamos la siguiente notación:

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y})$$
$$= \frac{n\sum xy - \sum x \sum y}{n} = \text{medida de la covarianza entre } X \text{ e } Y$$
$$S_{xx} = \sum (x - \bar{x})^2$$
$$= \frac{n\sum x^2 - (\sum x)^2}{n} = \text{medida de variabilidad en } X$$
$$S_{yy} = \sum (y - \bar{y})^2$$
$$= \frac{n\sum y^2 - (\sum y)^2}{n} = \text{medida de variabilidad en } Y$$

Con esta notación, podemos expresar b y r del siguiente modo:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{S_{xy}}{S_{xx}}$$

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Obsérvese que:

$$b\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} = \frac{S_{xy}\sqrt{S_{xx}}}{S_{xx}\sqrt{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = r$$

Así, la fórmula que relaciona b y r es:

$$r = b\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$$

La consecuencia práctica de esta relación es que *b y r* tienen siempre el mismo signo algebraico. Así, una correlación positiva implica una línea de regresión con pendiente positiva (una línea que se eleva de izquierda a derecha); una correlación negativa implica una línea de regresión con pendiente negativa (una línea que desciende de izquierda a derecha).

Reconsideremos ahora la suma de cuadrados SS_E que está siendo minimizada en el proceso de mínimos cuadrados:

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

La suma mide la variabilidad de los datos puntuales y_i en torno a la línea de regresión ajustada $a + bx_i$. Si la línea se ajusta estrechamente a los datos puntuales, entonces SS_E será pequeña; de lo contrario, será grande.

Se puede utilizar un argumento algebraico bastante astuto que comprende el uso de las reglas de la suma dadas en el Apéndice A, para demostrar que:

$$SS_E = S_{yy} - bS_{xy}$$

El argumento viene dado enteramente en [11]. Dividiendo cada miembro de la ecuación por S_{yy}, obtenemos:

$$\frac{SS_E}{S_{yy}} = 1 - b\frac{S_{xy}}{S_{yy}}$$

Sustituyendo S_{xy}/S_{xx} en esta ecuación por la *b,* tenemos:

$$\frac{SS_E}{S_{yy}} = 1 - \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Puesto que $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$, concluimos que:

$$\frac{SS_E}{S_{yy}} = 1 - r^2$$

o que:

$$r^2 = 1 - \frac{SS_E}{S_{yy}} = \frac{S_{yy} - SS_E}{S_{yy}}$$

Dado que S_{yy} mide la variación total en Y y SS_E mide la variación aleatoria de *Y* en torno a la línea de regresión, S_{yy} - SS_E es una medida de variación de *Y* que no es aleatoria. Es decir, S_{yy} - SS_E es una medida de la variabilidad de Y que puede ser atribuida a su asociación lineal con *X.*

El estadístico r^2 se llama *coeficiente de determinación.* En la práctica,

$$r^2 = \frac{\text{variación de } Y \text{ debida a la linealidad}}{\text{variación total en } Y}$$

Si multiplicamos r^2 por 100, obtenemos el porcentaje de la variación de Y que puede atribuirse a la relación lineal entre X e Y. Si r^2 es grande, debemos concluir que hay una asociación lineal fuerte entre X e Y.

El coeficiente de determinación puede utilizarse para justificar la escala de correlación de la Figura 11.14. Cualquier correlación que se sitúe estrictamente entre -0.5 y 0.5 se considera débil, porque el r^2 de estos valores es inferior a 0.25. Para estos valores, menos del 25 % de la variación en Y se atribuye a una asociación lineal con X; más del 75 % de la variación en Y no está explicada. Los coeficientes de correlación moderados tienen valores r^2 que, como mínimo, son 0.25, pero inferiores a 0.81. Para decir que una correlación es fuerte, deseamos que su valor absoluto sea al menos 0.9. En este caso, el porcentaje de variación en Y explicado por su asociación lineal con X es sustancial en el sentido de que es del 81% o más. El Ejemplo 11.4.1 ilustra el uso del coeficiente de determinación.

Ejemplo 11.4.1. Volvamos a examinar la relación entre X, porcentaje por encima del peso ideal, e Y, umbral individual del dolor. En el Ejemplo 11.3.3 hallamos que $r = -0.33$ e interpretamos que hay una ligera correlación negativa entre las dos variables. Para hacerse una idea mejor de la consistencia de la relación lineal, calculamos r^2, coeficiente de determinación. El valor de este estadístico es $r^2 = (-0.33)^2 = 0.1089$. Multiplicando por 100, concluimos que solamente el 10.89% de la variación de Y es atribuible a su asociación lineal con X. Puesto que este porcentaje es pequeño, una correlación de -0.33 no es realmente muy fuerte. Ello indica que no hay una fuerte tendencia de los individuos obesos a mostrar un bajo umbral de dolor, y viceversa.

Análisis de la varianza

Existe una técnica de análisis de la varianza que se utiliza para comprobar si una línea recta muestra una cantidad significativa de la variabilidad observada de Y. Como en cualquier proceso de análisis de la varianza, la idea es dividir la variabilidad total de Y, S , en componentes que puedan ser atribuidas a orígenes reconocibles. Esto se puede hacer fácilmente puesto que ya hemos establecido que SS_E, la variabilidad aleatoria en torno a la línea de regresión ajustada, puede escribirse en la forma:

$$SS_E = SS_{yy} - bS_{xy}$$

Resolviendo esta ecuación en S_{yy}, vemos que:

$$S_{yy} = bS_{xy} + SS_E$$

La segunda componente de la derecha, SS_E, se llama *suma de cuadrados de los errores o residuos;* es una medida de la variabilidad en Y, aleatoria o no explicada. La primera componente de la derecha, bS_{xy}, la *suma de cuadrados de la regresión,* mide la variabilidad en Y atribuible a la asociación lineal entre X e Y. La suma de cuadrados de la regresión se designa por SS_R. De este modo, hemos dividido S_{yy} en las dos componentes:

S_{yy}	—	SS_R	+	SS_E
(variabilidad total en Y)		(variabilidad en Y debida a la regresión sobre X)		(variación no explicada o aleatoria)

Lógicamente, si el supuesto de regresión lineal es válido, entonces SS_R explicará la mayor parte de la variabilidad de Y, siendo aleatoria o no explicada solamente una cantidad pequeña. Así estaremos en condiciones de utilizar los tamaños relativos de SS_R y SS_E para decidir de alguna manera si el supuesto de regresión lineal es razonable.

Esto puede hacerse estableciendo nuevas hipótesis respecto a la variable dependiente Y. Supongamos que estamos tratando con k valores específicos de la variable independiente $x_1, x_2, x_3, \ldots, x_k$. Esto implica que estamos trabajando con k variables aleatorias $Y|x_1, Y|x_2, Y|x_3, \ldots, Y|x_k$. Suponemos que son variables aleatorias normales independientes con la misma varianza, σ^2. Si la regresión lineal es válida, entonces las medias de estas variables estarán situadas sobre la línea recta $\mu_{Y|x} = \alpha + \beta x$. La idea se recoge en la Figura 11.16.

Supongamos ahora que se selecciona de cada distribución una muestra aleatoria de tamaño n_i, $i = 1, 2, \ldots, k$. Sea Y_{ij} elj-ésimo elemento de la muestra aleatoria de la distribución de $Y|x_i$. La variable Y_{ij} es una variable aleatoria con media $\alpha + \beta x_i$ y varianza σ^2. No se espera que su valor observado caiga exactamente en el valor medio, sino que se desvíe de él en alguna cantidad aleatoria, E_{ij}. Así podemos escribir la siguiente expresión, que sirve como modelo para la regresión lineal simple:

Modelo

$$Y_{ij} = \alpha + \beta x_i + E_{ij} \qquad \begin{matrix} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n_i \end{matrix}$$

$$\begin{bmatrix} \text{Valor observado de } Y \text{ para un valor particular de } X \end{bmatrix} = \begin{bmatrix} \text{Valor medio de } Y \text{ para este valor de } X \end{bmatrix} + \begin{bmatrix} \text{Desviación aleatoria de la media} \end{bmatrix}$$

Las hipótesis hechas sobre Y implican los siguientes supuestos respecto al modelo:

Supuestos del modelo. Las desviaciones aleatorias E_{ij} son variables aleatorias normales independientes con media 0 y varianza σ^2.

Con ellos es posible formular matemáticamente las hipótesis nula y alternativa apropiadas, y contrastar la hipótesis nula utilizando SS_R y SS_E. Queremos contrastar:

H_0: la variación en Y no está explicada por el modelo lineal

H_1: una parte significativa de la variación en Y está explicada por el modelo lineal

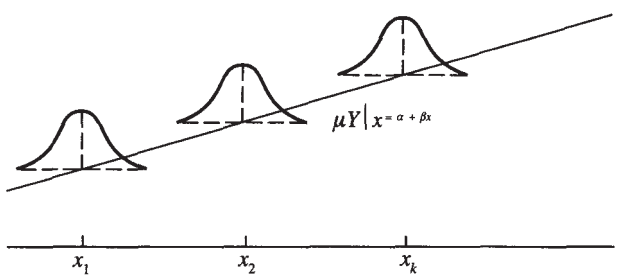


Figura 11.16. Cada variable aleatoria $Y|x_i$ está normalmente distribuida con la misma varianza.

Obsérvese que si $\beta = 0$ el modelo se transforma en $Y_{ij} = \alpha + E_{ij}$. Es decir, si $\beta = 0$, entonces se supone que *toda* la variabilidad de Y es aleatoria; si $\beta \neq 0$, entonces se supone que al menos una parte de la variabilidad se debe a la regresión lineal de Y sobre X. La hipótesis nula de no regresión lineal puede expresarse en la forma H_0: $\beta = 0$. Estamos, pues, interesados en contrastar:

$$H_0: \beta = 0 \text{ (no hay regresión lineal)}$$
$$H_1: \beta \neq 0$$

El contraste utiliza dos estadísticos que son funciones de SS_R y SS_E. El primero, llamado *cuadrado medio de la regresión* MS_R, se halla dividiendo SS_R por 1. El segundo, *cuadrado medio del error* se halla dividiendo SS_E por $n - 2$, donde $n = \sum_{i=1}^{k} n_i$ indica el tamaño global de la muestra. Es decir, definimos:

$$MS_R = \frac{SS_R}{1} \qquad MS_E = \frac{SS_E}{n-2}$$

Así, si H_0 es cierta, entonces el estadístico:

$$\boxed{\frac{MS_R}{MS_E}}$$

sigue una distribución F con 1 *y* n-2 grados de libertad. Además, si H_0, es cierta, entonces el valor observado del estadístico estará próximo a 1; en otro caso, sería mucho mayor. Por lo tanto, se rechaza la hipótesis nula de que no hay regresión lineal, si el valor observado del cociente F es demasiado *grande* para que se deba al azar.

Todo lo afirmado puede sintetizarse convenientemente en una tabla de análisis de la varianza, como la Tabla 11.1. La tabla incluye las fórmulas necesarias para calcular S_{yy}, suma total de cuadrados, y SS_R, la suma de cuadrados de la regresión. La suma de cuadrados de Jos errores SS_E se obtiene sustrayendo SS_R de S . Recordemos que:

$$S_{xy} = \frac{n\sum xy - \sum x \sum y}{n} \qquad S_{xx} = \frac{n\sum x^2 - (\sum x)^2}{n}$$
$$S_{yy} = \frac{n\sum y^2 - (\sum y)^2}{n} \qquad b = \frac{S_{xy}}{S_{xx}}$$

Tabla 11.1. ANOVA utilizado para contrastar H_0 : $\beta = 0$ (no hay regresión lineal)

Origen de la variación	Grados de libertad *DF*	Suma de cuadrados *SS*	Cuadrado medio *MS*	Cociente *F*
Regresión (modelo)	1	bS_{xy}	$\frac{SS_R}{1}$	$F_{1,n-2} = \frac{MS_R}{MS_E}$
Error	n - 2	$S_{yy} - bS_{xy}$	$\frac{SS_E}{n-2}$	
Total	n - 1	S_{yy}		

En el Ejemplo 11.4.2 continuamos el análisis de los datos del Ejemplo 11.1.3.

Ejemplo 11.4.2. En el Ejemplo 11.1.3, las variables de interés son *X* e *Y,* altura y longitud, respectivamente, de la concha de la lapa *Patelloida pygmaea.* La nube de puntos indica que el supuesto de regresión lineal es válido. Comprobamos estadísticamente ese supuesto en lugar de confiar únicamente en el modelo gráfico de los datos. Necesitamos los siguientes valores:

$$\sum x = 56.6 \qquad \sum y = 151.1 \qquad \sum xy = 311.96$$
$$\sum x^2 = 117.68 \qquad \sum y^2 = 832.85 \qquad S_{xy} = 6.52$$
$$S_{xx} = 3.27 \qquad S_{yy} = 17.45 \qquad N = 28$$

A partir de ellos se obtiene:

$$\hat{\alpha} = a = 1.36 \qquad r = 0.8638$$
$$\hat{\beta} = b = 1.99 \qquad r^2 = 0.7461$$

Puesto que $r^2 = 0.7461$, el 74.61 % de la variación de *Y* puede atribuirse a una asociación lineal con *X.* ¿Es este porcentaje lo suficientemente grande para concluir que mediante el modelo lineal se explica una cantidad significativa de la variabilidad de *Y?*

Para contestar a esta pregunta, contrastamos:

$$H_0: \beta = 0 \text{ (no hay regresión lineal)}$$
$$H_1: \beta \neq 0$$

utilizando la técnica de análisis de la varianza. La ANOVA para estos datos se muestra en la Tabla 11.2. Basándonos en la distribución $F_{1,26}$, se puede rechazar H_0 con $P < 0.01$. Podemos concluir que, como se esperaba, el supuesto de regresión lineal es válido. En términos prácticos, ello significa que serán aceptables las predicciones basadas en la línea de regresión estimada:

$$\hat{\mu}_{Y|x} = 1.36 + 1.99x$$

Observemos que, cuando se rechaza la hipótesis nula de que no hay regresión lineal, se ha llegado a la conclusión de que mediante el modelo lineal puede explicarse una parte significa-

Tabla 11.2. ANOVA para los datos del Ejemplo 11.4.2. La hipótesis nula de no regresión lineal se rechaza con P< 0.01

Origen	***DF***	***SS***	***MS***	***F***
Regresión	1	$bS_{xy} = 12.97$	12.97	$\frac{12.97}{0.17} = 76.29$
Error	26	17.45 - 12.97 = 4.48	$\frac{4.48}{26} = 0.17$	
Total	27	$S_{yy} = 17.45$		

tiva de la variabilidad de *Y*. Ello no quiere decir que el modelo lineal sea necesariamente el mejor modelo a utilizar, sino que es razonable hacerlo.

Nota sobre los cálculos

Ahora es posible explicar algo más del *printout* que se muestra en la Figura 11.11. En el Ejemplo 11.2.3, hemos observado que en la nube de puntos parece haber una tendencia lineal en los datos. La recta de regresión estimada aparece como:

$$\hat{\mu}_{Y|x} = 290.07044608 - 15.11057071x$$

El valor r^2 de estos datos, 0.726096, se muestra en el *printout* bajo R-SQUARE. Así, el 72.6096 % de la variación en la duración de la época de reproducción está asociada con X, el fotoperíodo. El análisis de la varianza utilizado para contrastar H_0: $\beta = 0$ se da en la parte superior de la Figura 11.11. En SAS, el origen de variación que hemos llamado «regresión» se denomina MODEL. En el *printout* vemos que el valor observado del estadístico *F* utilizado para contrastar H_0 es 23.86 y su valor *P* es 0.0009. Dado que *P* es pequeño, podemos rechazar H_0 y concluir que resulta adecuada la regresión lineal.

EJERCICIOS 11.4

1. Se lleva a cabo un estudio sobre las características corporales y el modo de actuar de los levantadores de peso olímpicos, superiores y de primera clase. Se estudian dos variables, la *X,* peso corporal del sujeto, e *Y,* su mejor levantamiento dictaminado en cuanto a limpieza y empuje. Se obtuvieron los siguientes datos en libras:

x	*y*	*x*	*y*
134	185	190	336
138	238	190	339
154	260	205	341
178	290	205	358
176	312	206	359

   a) Dibujar la nube de puntos. Basándose en ella, ¿se puede esperar que *b* sea positivo o negativo?
   b) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$, y *r*.
   c) Hallar e interpretar el coeficiente de determinación.
   d) Comprobar la idoneidad del modelo lineal de regresión. Si es adecuado, hallar la línea de regresión estimada de *Y* sobre *X,* y utilizarla para estimar el mejor levantamiento en cuanto a limpieza y empuje para un levantador de peso que pese 200 libras.
2. Se lleva a cabo un estudio, por medio de técnicas de seguimiento de marcadores radiactivos, de la capacidad corporal para absorber hierro y plomo. Participan en el estudio diez sujetos. A cada uno se le da una dosis oral idéntica de hierro (sulfato ferroso) y de plomo (cloruro de plomo-203). Después de doce días, se mide la cantidad de cada componente

retenida en el sistema corporal y, a partir de ésta, se determina el porcentaje absorbido por el cuerpo. Se obtuvieron los siguientes datos:

x (porcentaje de hierro absorbido)	y (porcentaje de plomo absorbido)
17	8
22	17
35	18
43	25
80	58
85	59
91	41
92	30
96	43
100	58

a) Dibujar la nube de puntos. Basándose en ella, ¿se puede determinar si b será positivo o negativo?
b) Hallar $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$ y r.
c) Hallar e interpretar el coeficiente de determinación.
d) Comprobar la idoneidad del modelo lineal de regresión. Si es apropiado, estimar la verdadera línea de regresión y utilizarla para predecir el porcentaje de hierro absorbido por un individuo cuyo sistema corporal absorbe el 15% del plomo ingerido.

3. Verificar la tabla del análisis de la varianza de la Figura 11.11.
4. *a)* Hallar e interpretar el coeficiente de determinación para los datos del Ejercicio 4 de la Sección 11.2.
   b) Comprobar la idoneidad del modelo lineal de regresión utilizando los datos del Ejercicio 4 de la Sección 11.2. ¿Cree el lector que la predicción dada en el apartado *d* del Ejercicio 4 de la Sección 11.2 es un buen estimador para y?
5. Utilizar los datos del Ejercicio 5 de la Sección 11.3 para hallar el coeficiente de determinación para las variables X, lectura de nitrato hecha mediante una antigua técnica manual, e Y, lectura hecha mediante un nuevo método automático. Sobre esta base, ¿cree el lector que el nuevo método reflejará con precisión las lecturas que habrían sido obtenidas con la técnica antigua?
6. Considérense las correlaciones del Ejercicio 6 de la Sección 11.3.
   a) Hallar el coeficiente de determinación para cada tipo de alimento.
   b) ¿Cuál es el porcentaje de variación máxima en Y (el informe de la madre) explicado por su asociación lineal con X (informe del observador)?
   c) Basándose en el valor r^2, ¿piensa usted que la correlación negativa asociada al fósforo tiene alguna importancia práctica? Explicarlo.
7. Considérese el Ejercicio 7 de la Sección 11.3. Hallar el coeficiente de determinación de cada correlación dada.
8. Considérense las ecuaciones de regresión del Ejercicio 6 de la Sección 11.2. Los valores de r para las ecuaciones son 0.805 (FVC), 0.836 (FEV_1), 0.825 (PEFR) y -0.689 (FET). En cada caso, ¿qué porcentaje de variación en la respuesta está ligado a la asociación lineal con X, la distancia más lejana a la que puede apagarse la vela?

11.5. ESTIMACIONES POR INTERVALOS DE CONFIANZA (OPCIONAL)

Es posible ahora hallar estimaciones puntuales para α, β, $\mu_{Y|x}$, y $Y|x$, valor de Y para un valor específico de X. Como anteriormente, es natural querer extender estas estimaciones puntuales a estimaciones por intervalos, de forma que se pueda informar sobre los niveles de confianza. Ello puede hacerse con los supuestos del modelo que hemos establecido. Consideremos el Ejemplo 11.5.1.

Ejemplo 11.5.1. Aunque no todos los productores primarios necesitan silicona disuelta en el agua de mar, hay investigaciones que relacionan la falta de esta sustancia con una productividad decreciente. Se lleva a cabo un estudio para poder establecer un criterio con respecto al comportamiento de la silicona disuelta. Las dos variables estudiadas son *X,* la distancia en kilómetros de la costa, e Y la concentración de silicona en microgramos por litro (μg/litro). Se realizan estas medidas basándose en tomas efectuadas en la plataforma continental del noroeste africano. (Obsérvese que *X,* la variable independiente, *no* es aleatoria. Se eligen seis distancias de la costa y se hacen cuatro medidas a cada distancia.)

x	y	x	y	x	y
5	6.1	25	3.7	42	3.4
5	6.2	25	3.7	42	3.6
5	6.1	25	3.8	42	3.5
5	6.0	25	3.9	42	3.2
15	5.2	32	3.9	55	3.7
15	5.0	32	3.8	55	3.9
15	4.9	32	3.9	55	3.6
15	5.1	32	3.7	55	3.8

En la Figura 11.17 se muestra la nube de puntos. Parece existir una tendencia lineal descendente. Por lo tanto, anticipamos una pendiente negativa para la línea de regresión estillada. Primero comprobamos la idoneidad del modelo lineal de regresión. Para estos datos,

$$\sum x = 696 \qquad \sum y = 103.7 \qquad \sum xy = 2692.5$$

$$\sum x^2 = 26\,752 \qquad \sum y^2 = 469.81 \qquad S_{xy} = -314.8$$

$$\bar{x} = 29 \qquad \bar{y} = 4.32 \qquad b = \frac{S_{xy}}{S_{xx}} = -0.048$$

$$S_{xx} = 6568 \qquad S_{yy} = 21.74 \qquad a = \bar{y} - b\bar{x} = 5.71$$

En la Tabla 11.3 se muestra el análisis de la varianza. En la tabla F observamos que $P[F_{1,22} \geq 7.95] = 0.01$. Dado que 50.37, el valor observado de F, excede este punto, $P < 0.01$. Este valor P es pequeño, de forma que podemos rechazar H_0: $\beta = 0$ en favor de H_1: $\beta \neq 0$ y concluir que está garantizado el supuesto de linealidad. Disponemos ahora de estimaciones *puntuales* para α y β, a saber, $\hat{\alpha} = \alpha = 5.71$ y $\hat{\beta} = b = -0.048$.

Supongamos que estamos particularmente interesados en la concentración de silicona a una distancia de 10 km de la costa. Se pueden plantear dos preguntas: ¿Cuál es la concentración media a esta distancia?, y, si se extrae una *única* muestra de agua a esta distancia, ¿cuál será la concentración de silicona para la muestra? Estas preguntas pueden responderse ya utilizando estimadores puntuales. En concreto, se nos está pidiendo estimar $\mu_{Y|x} = 10$ e $Y|x = 10$. Se trata de la misma estimación, es decir:

$$\hat{\mu}_{Y|x} = \hat{y} = a + bx$$
$$= 5.71 - 0.048x = 5.23 \text{ μg/litro}$$

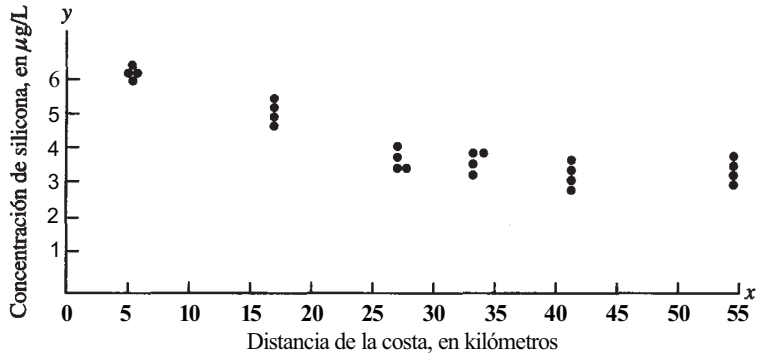


Figura 11.17. Nube de puntos de la distancia de la costa frente a la concentración de silicona en aguas de la plataforma continental del noroeste africano.

Nuestro problema es extender cada una de estas estimaciones puntuales a intervalos de estimación, de modo que se pueda establecer un nivel de confianza.

La obtención de intervalos de confianza de α, β, $\mu_{Y|x}$, e $Y|x$ está fuera de los objetivos del texto. Sin embargo, el modelo implicado en cada caso es el mismo que el que hemos visto anteriormente. Es decir, construimos el intervalo de confianza añadiendo una cantidad específica a la estimación puntual para obtener un límite superior del intervalo de confianza; el límite inferior se halla restando este valor de la estimación puntual. La cantidad a sumar y restar depende de la confianza deseada y de la desviación típica estimada del estimador puntual del parámetro en cuestión. En esta determinación, σ^2 se estima mediante $MS_E = SS_E/(n-2)$. Para simplificar la notación en el trabajo futuro, indicaremos el estimador de σ por S. En este caso, $S = \sqrt{MS_E}$.

Los intervalos de confianza para el término independiente y la pendiente de la recta de regresión son:

$$a \pm \frac{tS\sqrt{\sum x^2}}{\sqrt{nS_{xx}}}$$

(Intervalo de confianza de α, término independiente de la verdadera recta de regresión)

$$b \pm \frac{tS}{\sqrt{S_{xx}}}$$

(Intervalo de confianza de β, pendiente de la verdadera recta de regresión)

Tabla 11.3. ANOVA para los datos del Ejemplo 11.5.1. La hipótesis nula de regresión no lineal puede rechazarse con $P < 0.01$

Origen	*DF*	*SS*	*MS*	*F*
Regresión	1	15.11	15.11	50.37
Error	22	6.63	0.30	
Total	23	21.74		

En cada caso, el punto t es el punto asociado a la distribución T_{n_2} cuyo valor depende de la confianza deseada. El Ejemplo 11.5.2 continúa el análisis de los datos del Ejemplo 11.5.1 construyendo intervalos de confianza del 95 % para la pendiente y el término independiente de la verdadera recta de regresión.

Ejemplo 11.5.2. Para construir un intervalo de confianza de α del 95 % basándose en los datos del Ejemplo 11.5.1, solamente necesitamos sustituir los correspondientes valores en la fórmula:

$$a \pm t \frac{S\sqrt{\sum x^2}}{\sqrt{nS_{xx}}}$$

Para los datos de que disponemos,

$$\sum x^2 = 26\,752 \qquad n = 24 \qquad a = 5.71$$

$$S_{xx} = 6568 \qquad S = \sqrt{\frac{SS_E}{n-2}} = \sqrt{0.30} = 0.5477$$

El punto t necesario se muestra en la Figura 11.18. Su valor 2.074 se obtiene de la tabla de la distribución T con n - 2 = 22 grados de libertad. El intervalo de confianza es, por tanto,

$$5.71 \pm 2.074 \frac{\sqrt{0.3}\sqrt{26\,752}}{\sqrt{24(6568)}} = 5.71 \pm 0.47$$

Es decir, podemos tener un 95 % de confianza en que el punto de intersección de la línea de regresión con el eje Y está entre 5.24 y 6.18.

El intervalo de confianza de β al 95% se halla sustituyendo los valores t = 2.074, $S = \sqrt{0.3}$, S_{xx} = 6568 y b = -0.048 en:

$$b \pm t \frac{S}{\sqrt{S_{xx}}}$$

El intervalo de confianza que resulta es:

$$-0.048 \pm 2.074 \frac{\sqrt{0.3}}{\sqrt{6568}} = -0.048 \pm 0.014$$

Podemos tener un 95 % de confianza en que la pendiente de la línea de regresión esté entre -0.062 y -0.034. Obsérvese que el procedimiento de análisis de la varianza ya ha indicado que $\beta \neq 0$. De este modo, no sería sorprendente que el intervalo de confianza sobre β *no* contenga al número 0.

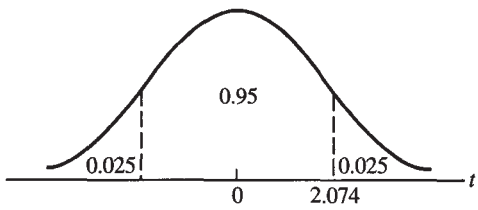


Figura 11.18. El punto T necesario para construir un intervalo de confianza del 95% para α, término independiente de la recta de regresión de los datos del Ejemplo 11.5.1.

Recuérdese que la recta de regresión estimada puede utilizarse para obtener una estimación puntual de la respuesta media para un valor de X dado, $\mu_{Y|x}$, o para una respuesta individual, $Y|x$. Las dos estimaciones puntuales son idénticas. Sin embargo, sus estimaciones de intervalo no son idénticas. Puesto que es mucho más difícil predecir el comportamiento individual que el comportamiento en grupo, es mucho más difícil precisar $Y|x$ que $\mu_{Y|x}$. Un intervalo previsto para incluir $Y|x$ debe ser más amplio que aquél cuyo objetivo es incluir $\mu_{Y|x}$. Para distinguir estos intervalos tan similares, nos referiremos al intervalo utilizado para contener a $\mu_{Y|x}$, como intervalo de confianza; un intervalo construido para contener a $Y|x$ se denominará *intervalo de predicción.* Se utilizan las fórmulas siguientes:

$$\hat{\mu}_{Y|x} \pm tS\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

(Intervalo de confianza de la respuesta media cuando $X = x$)

$$\hat{y} \pm tS\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

(Intervalo de predicción de la respuesta individual cuando $X = x$)

En cada caso, el punto t asociado es relativo a la distribución T_{n-2}. Estas fórmulas se ilustran en el Ejemplo 11.5.3.

Ejemplo 11.5.3. El intervalo de confianza al 95 % de la concentración media de silicona hallada a 10 km de la costa se calcula sustituyendo los valores $t = 2.074$, $S = \sqrt{0.3}$, $n = 24$, $x = 10$, $\bar{x} = 29$, $S_{xx} = 6568$ y $\hat{\mu}_{Y|x} = 5.23$ en:

$$\hat{\mu}_{Y|x} \pm tS\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

El intervalo de confianza resultante es:

$$5.23 \pm 2.074\sqrt{0.3}\sqrt{\frac{1}{24} + \frac{(10-29)^2}{6568}} = 5.23 \pm 0.35$$

Podemos tener un 95 % de confianza en que la concentración *media* de silicona a esta distancia está entre 4.88 y 5.58 μg/litro.

El intervalo de confianza al 95 % de la concentración de silicona para una única muestra de agua extraída a 10 km de la costa se halla sustituyendo los valores $t = 2.074$, $S = \sqrt{0.3}$, $n = 24$, $x = 10$, $\bar{x} = 29$, $S_{xx} = 6568$, e $\hat{y} = 5.23$ en:

$$\hat{y} \pm tS\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

El intervalo resultante es:

$$5.23 \pm 2.074\sqrt{0.3}\sqrt{1 + \frac{1}{24} + \frac{(10-29)^2}{6568}} = 5.23 \pm 1.19$$

Podemos tener un 95 % de confianza en que la concentración de silicona para la siguiente muestra que se tome a 10 km de la costa estará entre 4.04 y 6.42 μg/litro. Como se esperaba, el intervalo de predicción de $Y|x = 10$ es más amplio que el intervalo de confianza de $\mu_{Y|x=10}$.

EJERCICIOS 11.5

1. Se obtuvieron los siguientes datos para *X,* latitud del espacio natural de cría, e *Y,* duración del período de cría, en días, de 11 especies de patos buceadores:

x	*y*	*x*	*y*
29	112	53	42
42	98	54	50
45	58	55	18
45	68	60	51
50	28	65	49
50	46		

   α) Dibujar la nube de puntos. Basándose en ella, ¿cree el lector que es aplicable la regresión lineal?
   b) Utilizar el método de análisis de la varianza para comprobar la idoneidad del modelo lineal.
   c) Hallar estimaciones puntuales para α y *β.*
   d) Hallar intervalos de confianza al 95 *%* para α y *β.*
   e) Hallar una estimación puntual para la duración media del período de cría para ayes cuyo espacio natural de cría está a una latitud de 35 grados. Hallar un intervalo de confianza al 95 % para este parámetro.
   f) Hallar una estimación puntual para la longitud del período de cría para una única ave cuyo espacio natural de cría está a una latitud de 35 grados. Hallar un intervalo de confianza al 95 % para el valor predicho.

2. Se han obtenido importantes ventajas del hecho de enseñar a los diabéticos a medir su propia glucosa en sangre. Se investiga una nueva técnica menos costosa que el procedimiento habitual. La técnica utiliza una varilla con la enzima glucosa oxidasa. La varilla desarrolla dos colores simultáneamente y estos colores son comparados a simple vista con una tarjeta que da el nivel de glucosa. Si se puede probar que este procedimiento es preciso, se generalizará su uso. Se obtuvieron los siguientes datos de *X,* nivel de glucosa en sangre medido por un paciente diabético utilizando la varilla, e *Y,* nivel de glucosa en sangre del paciente medido en laboratorio. (Los datos están dados en milimoles por litro.) Para nuestros datos,

$$\sum x = 295 \qquad \sum y = 303.5 \qquad \sum xy = 3073.55 \qquad S_{xx} = 915.335$$
$$\sum x^2 = 3090.96 \qquad \sum y^2 = 3120.59 \qquad S_{xy} = 835.2375 \qquad S = 1.21$$
$$\bar{x} = 7.375 \qquad \bar{y} = 7.5875 \qquad S_{yy} = 817.78375$$

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1.3	2.4	3.2	4.4	7.0	7.7	15.0	14.9
2.0	3.0	3.6	4.3	8.0	8.0	15.0	13.8
2.4	2.3	3.7	4.3	8.0	10.0	17.5	17.6
2.6	3.0	3.7	5.0	10.0	10.0	18.7	17.5
2.5	2.2	3.8	4.4	10.2	9.5	6.0	6.0
2.6	2.4	4.4	4.5	10.2	11.2	8.7	8.8

x	*y*	*x*	y	x	*y*	*x*	*y*
2.7	2.5	4.3	5.0	12.5	11.0	5.6	5.7
3.0	3.8	5.0	4.5	11.3	13.0	9.1	9.0
3.7	2.5	5.0	6.2	13.0	13.1	16.2	12.5
3.7	3.5	6.3	6.2	14.5	13.8	9.0	14.0

a) Dibujar la nube de puntos. Basándose en ella, ¿parece apropiada la regresión lineal?
b) Hallar e interpretar el coeficiente de determinación.
c) Utilizar el procedimiento del análisis de la varianza para comprobar la idoneidad del modelo lineal.
d) Hallar estimaciones puntuales para α y β.
e) Hallar intervalos de confianza de α y β al 90 %.
f) Hallar una estimación puntual del nivel de glucosa, establecido en laboratorio, de un paciente que lo sitúa en 4.0 mmol/litro. Hallar un intervalo de confianza de este valor al 90%.

11.6. REGRESIÓN MÚLTIPLE (OPCIONAL)

Como se ha mencionado anteriormente, cuando se utiliza más de un regresor para estimar la respuesta media, el problema es de regresión múltiple. El modelo de la regresión lineal múltiple toma la forma:

$$\mu_{Y|x_1, x_2, x_3, \ldots, x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Aquí Y es la respuesta y $\mu_{Y|x_1, x_2, x_3, \ldots, x_k}$ es la media de la respuesta. Se supone que su valor depende de los valores asumidos por los regresores $x_{t,\ 2}, \ldots, x_k$. El término β_0 representa un término independiente desconocido y $\beta_1, \beta_2, \ldots, \beta_k$ representan coeficientes desconocidos. Obsérvese que si k = 1, este modelo coincide con el modelo de regresión lineal simple de la Sección 11.2.

Ejemplo 11.6.1. Se sabe que en los mamíferos se puede alterar la toxicidad de diferentes tipos de fármacos, pesticidas y carcinógenos químicos induciendo la actividad enzimática del hígado. Se ha informado de un estudio para investigar este tipo de fenómeno en pollos realizado por M. Ehrich, C. Larson y J. Arnold, «Organophosphate Detoxification Related by Induced Hepatic Microsomal Enzymes in Chickens», *American Journal of Veterinary Research,* vol. 45,1983. Se utilizó el análisis de regresión para estudiar las relaciones entre la actividad enzimática inducida y la detoxificación del insecticida malatión. El hidroxitolueno butilatado (BHT) fue el inductor enzimático utilizado. Se emplearon como regresores cinco actividades enzimáticas y la respuesta fue el porcentaje de detoxificación del malatión. El modelo de regresión lineal múltiple expresa la idea de que el porcentaje medio de detoxificación del malatión depende del nivel de cada una de las cinco enzimas. Matemáticamente escribimos:

$$\mu_{Y|x_1, x_2, x_3, x_4, x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

Al igual que con la regresión lineal simple, nuestro problema es utilizar el método de los mínimos cuadrados para estimar los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Deseamos hallar los valores de estos parámetros que minimizan la suma de los cuadrados de los residuos, donde un residuo es la diferencia entre el dato puntual observado y el valor de la respuesta predicha por la

ecuación de regresión estimada. En el caso de la regresión lineal simple, pudimos desarrollar algunas ecuaciones simples para estimar la pendiente y el término independiente, estimando así la ecuación de regresión. Aquí esto no es posible. No existen expresiones algebraicas simples que puedan desarrollarse para estimar $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ cuando $k > 1$. La herramienta matemática requerida para obtener estas estimaciones es el álgebra matricial y las estimaciones se obtienen fácilmente únicamente mediante el ordenador. Una vez obtenidas las estimaciones, la ecuación de regresión se utiliza para el mismo fin que en el caso de la regresión lineal simple. Es decir, puede utilizarse para estimar la respuesta media para diferentes valores de los regresores o para predecir una respuesta individual para estos valores. El Ejemplo 11.6.2 ilustra la idea.

Ejemplo 11.6.2. Cuando se realizó el estudio descrito en el Ejemplo 11.6.1, se obtuvo esta ecuación con ayuda del ordenador:

$$\hat{\mu}_{Y|x_1, x_2, x_3, x_4, x_5} = 54.079 + 0.097x_1 + 0.034x_2 + 0.522x_3 - 2.655x_4 + 2.559x_5$$

Supóngase que deseamos estimar el porcentaje medio de detoxificación cuando $x_1 = 350$, $x_2 = 270$, $x_3 = 100$, $x_4 = 85$ y $x_5 = 102$. Esto se hace sustituyendo estos valores en la ecuación anterior. Obtenemos:

$$\begin{aligned}\hat{\mu}_{Y|x_1, x_2, x_3, x_4, x_5} &= 54.079 + 0.097(350) + 0.034(270) + 0.522(100) \\ &\quad -2.655(85) + 2.559(102) \\ &= 184.752\end{aligned}$$

Esto es también el porcentaje predicho de detoxificación de un pollo individual con estos niveles de las cinco enzimas. Obsérvese el impacto del signo algebraico de los coeficientes sobre la respuesta. Un coeficiente positivo implica que el aumento en el nivel enzimático tiende a aumentar el porcentaje de detoxificación. Un signo negativo significa lo opuesto. En todos los casos excepto en el enzima 4, a medida que aumenta el nivel enzimático aumenta el porcentaje de detoxificación. Con el enzima 4, el aumento del nivel tiende a inhibir la detoxificación.

Un valor r^2 también aparece en el estudio de regresión múltiple. Tiene el mismo significado aquí que en el caso de la regresión lineal simple. Es decir, r^2 da el porcentaje de variación de respuesta explicada por la asociación lineal con los regresores $x_1, x_2, \ldots, x_k$. En el Ejemplo 11.6.2 se determinó que r^2 era 0.976. Este valor es muy alto. Podemos decir que el 97.6 % de la variación de la respuesta puede explicarse por la asociación lineal entre la respuesta y los cinco niveles enzimáticos.

La regresión múltiple es un tema muy complejo. Normalmente, cuando se realiza un estudio de regresión múltiple, el experimentador intenta señalar todas las variables que cree que son factores importantes para explicar la respuesta. Estos regresores no son elegidos por el profesional de la estadística sino por el experto en la materia, el experimentador. A contif nuación se realiza un experimento que implica tomar decisiones sobre todos los regresores y sobre la respuesta. Una vez realizado esto, se efectúa un poco de trabajo detectivesco. El experimentador, junto con el experto en estadística, deben decidir cuáles de los regresores propuestos desempeñan el mejor trabajo en explicar la respuesta. ¿Deben utilizarse todos los regresores como se ha hecho en el Ejemplo 11.6.1, o algún subconjunto de regresores funciona igualmente bien o aún mejor? El objetivo es hallar el mejor conjunto de regresores. Existen varios criterios utilizados para decidir cuál es el mejor conjunto. De éstos, r^2 y MS_E ya le son familiares. Para el modelo final, es deseable tener un valor alto de r^2. Deseamos que explique un alto porcentaje de la variación observada en la respuesta. Puesto que MS_E es una medida de la variación no explicada en la respuesta, el modelo final tendrá una media dé

cuadrados pequeña debido al error. Sin embargo, se sabe que a medida que se añaden más y más regresores al modelo, r^2 siempre aumenta. ¿Significa esto que «más es mejor»? La respuesta a la pregunta es *no*. Existen otros criterios que son indicadores mucho mejores de la capacidad predictiva de una ecuación de regresión que r^2 o MS_E. Estos criterios se ven afectados al añadir regresores innecesarios al modelo. Se puede realmente dañar la capacidad predictiva del modelo incluyendo regresores sin importancia. Así pues, para escoger el mejor modelo, debemos considerar todos los criterios y hacer un juicio de valor. Deseamos tomar el conjunto menor de regresores que produce un MS_E razonablemente pequeño, un r^2 aceptablemente alto y que tiene un buen valor predictivo según lo medido por los demás criterios. Existen muchos textos que tratan esas ideas con detalle. Si usted tiene un proyecto de investigación en el que aparezca regresión múltiple, le sugerimos que consulte [12] e inmediatamente a un especialista en estadística para ayudar a diseñar su experimento y analizar sus datos.

EJERCICIOS 11.6

1. Utilizar la ecuación de regresión del Ejemplo 11.6.1 para estimar la respuesta media cuando $x_1 = 233, x_2 = 260, x_3 = 82, x_4 = 80, x_5 = 88$. Estimar el porcentaje de detoxificación para un pollo con estos niveles enzimáticos.
2. Al derivar la ecuación de regresión del Ejemplo 11.6.1, la respuesta observada cuando $x_1 = 233, x_2 = 260, x_3 = 82, x_4 = 80, x_5 = 88$, ha sido 152. Utilizar la estimación hallada en el Ejercicio 1 para encontrar el residuo en este punto.
3. Considerar la ecuación de regresión del Ejemplo 11.6.1. Supóngase que mantenemos los niveles enzimáticos para las enzimas 1, 2, 3 y 5 como se ha indicado en el Ejercicio 1, pero cambiamos la x_4 de 80 a 81. ¿Qué efecto tendrá esto sobre la respuesta media estimada?
4. Utilizar la idea sugerida en el Ejercicio 3 para hacer una afirmación general en relación a la interpretación de los coeficientes $b_1, b_2, \ldots, b_k$ en una ecuación de regresión múltiple.
5. Se realiza un estudio sobre la compactación mecánica de la arenisca en Alaska. El objeto del estudio es determinar factores que afectan la porosidad. La ecuación de regresión obtenida es

$$\widehat{GF} = 90 + 0.23P - 0.72M + 0.0018\,d$$

donde GF = medida de porosidad
d = profundidad de enterramiento del material, en metros
P = porcentaje de granos dúctiles (granos que pueden extraerse o pulverizarse)
M = contenido de la roca madre (la roca madre es el material natural en el que se encuentra la arenisca)

Los valores grandes de GF indican un alto grado de compactación y menos porosidad.

a) La r^2 registrada es 0.73. Explicar lo que ello significa.

b) Considerar dos ejemplos con valores idénticos para P y M. La muestra 2 debe tomarse a un metro de profundidad más que la muestra 1. ¿Cuál es la diferencia estimada en GF entre las dos muestras?. ¿Cuál sería la diferencia si la muestra 2 se tomara a 10 metros más de profundidad que la muestra 1?

c) Considérense dos muestras con valores P y d idénticos. Si el valor de la roca madre de la muestra 2 excede en 20 al de la muestra 1, ¿cuál es la diferencia estimada en GF entre las dos muestras?

(Ecuación hallada en Richard Smosna, «Compaction Law for Cretaceous Sandstones of Alaska's North Slope», *Journal of Sedimentary Petrology,* julio de 1989, págs. 572-583.)

6. La ecuación de regresión del Ejercicio 5 tiene tres regresores. Hay siete ecuaciones de regresión que pueden escribirse y que implican el uso de uno o más de estos regresores Escribir estas siete ecuaciones.
7. ¿Cuántas ecuaciones de regresión pueden escribirse implicando el uso de uno o más de los regresores mencionados en el Ejemplo 11.6.1?
8. En cada uno de estos casos, utilice sus conocimientos sobre la materia para sugerir regresores que puedan utilizarse para estimar la respuesta media a la variable considerada,
   a) Y = pérdida de peso alcanzada tras un régimen de adelgazamiento de seis meses.
   b) Y= tamaño del anillo de crecimiento, por estación de crecimiento, de un árbol seleccionado aleatoriamente.
   c) Y = peso de un bebé recién nacido.
   d) Y = altura de un ser humano adulto.
   e) Y= tiempo requerido para que un paciente se recupere totalmente de una cadera rota.
   f) Y = nivel de colesterol en un adulto sano.
   g) Y= salto vertical de un jugador universitario de baloncesto.
   h) Y = velocidad de un pájaro en vuelo.

HERRAMIENTAS COMPUTACIONALES

TI83

XXV. Nube de puntos

Puede dibujarse fácilmente la nube de puntos utilizando la calculadora TI83. Para ilustrar el método, se usarán los datos del Ejemplo 11.1.3, reproducidos en la Figura 11.3.

Tecla/Comando de la TI83	Propósito
1. STAT 1	1. Acede al editor de datos estadísticos.
2. 0.9 ENTER 1.5 ENTER ⋮ 2.7 ENTER	2. Introduce valores de x en la columna L1.
3. ▷ 3.1 ENTER 3.6 ENTER ⋮ 6.3 ENTER	3. Introduce valores de y en la columna L2.
4. WINDOW	4. Accede a la ventana para especificar opciones gráficas.
5. 0 ENTER	5. Establece 0 como valor mínimo de x, un valor ligeramente por debajo del menor de los valores de x.

6. 3 ENTER	6. Establece 3 como valor máximo de *x,* un valor ligeramente por encima del mayor de los valores de *x.*
7. ▽ 0 ENTER	7. Establece 0 como valor mínimo de y.
8. 7 ENTER	8. Establece 7 como valor máximo de y.
9. 2^{ND} $Y=$ ENTER cursor a ON ENTER ▽ ▽ ▽ ▽ ▷ ▷ ENTER GRAPH	9. Dibuja la nube de puntos.
10. TRACE	10. Permite el movimiento dentro del gráfico, de un punto a otro, mediante las teclas del cursor izquierda y derecha.

XXVI. Regresión lineal simple

Se utilizan los datos del Ejemplo 11.2.1 para ilustrar cómo se emplea la TI83 para generar una nube de puntos, ajustar una línea recta a los datos mediante el método de los mínimos cuadrados, dibujar la línea de regresión sobre la nube de puntos y, entonces, utilizar la línea de regresión estimada para realizar predicciones.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 12.8 ENTER 13.9 ENTER 17.9 ENTER	2. Introduce los datos *x* en la columna L1.
3. ▷ 110 ENTER 54 ENTER 28 ENTER	3. Introduce los datos y en la columna L2.

4. WINDOW 10 ENTER	4. Establece 10 como valor mínimo de *x.*
5. 20 ENTER	5. Establece 20 como valor máximo de *x.*
6. ∇ 10 ENTER	6. Establece 10 como valor mínimo de y.
7. 120 ENTER	7. Establecer 120 como valor máximo de y.
8. STAT ▷ 8 ENTER	8. Obtiene, a partir de los datos, una línea estimada de la forma $y = \alpha + bx$; se muestra el caso de $\alpha = 290.0704461$ y $b = -15.11057071$.
9. 2^{ND} $Y =$ ENTER cursor a ON ENTER ∇ ENTER ∇ ENTER ∇ ∇ ▷ ▷ ENTER GRAPH	9. Dibuja la nube de puntos.
10. $Y =$ VARS 5 ▷ ▷ 1 GRAPH	10. Inserta la línea estimada de regresión dentro de la nube de puntos.
11. VARS ▷ 4 1 1 ENTER 2^{ND} TRACE 1 14.5 ENTER	11. Estima la longitud media de la duración de la estación de cría cuando $x = 14.5$.

XXVII. Correlación

Para ilustrar el uso de la TI83 para el cálculo de *r* utilizaremos los datos del Ejemplo 11.3.3.

Tecla/Comando de la TI83	**Propósito**
1. STAT 1	1. Accede al editor de datos estadísticos.
2. 89 ENTER 90 ENTER 20 ENTER	2. Introduce los datos *x* en la columna L1.
3. ▷ 2 ENTER 3 ENTER 14 ENTER	3. Introduce los datos y en la columna L2.
4. STAT ▷ 8 ENTER VARS 5 ▷ ▷ 7 ENTER	4. Calcula y muestra el valor de *r* (-0.339079882).

Paquete estadístico SAS

XIII. Nube de puntos

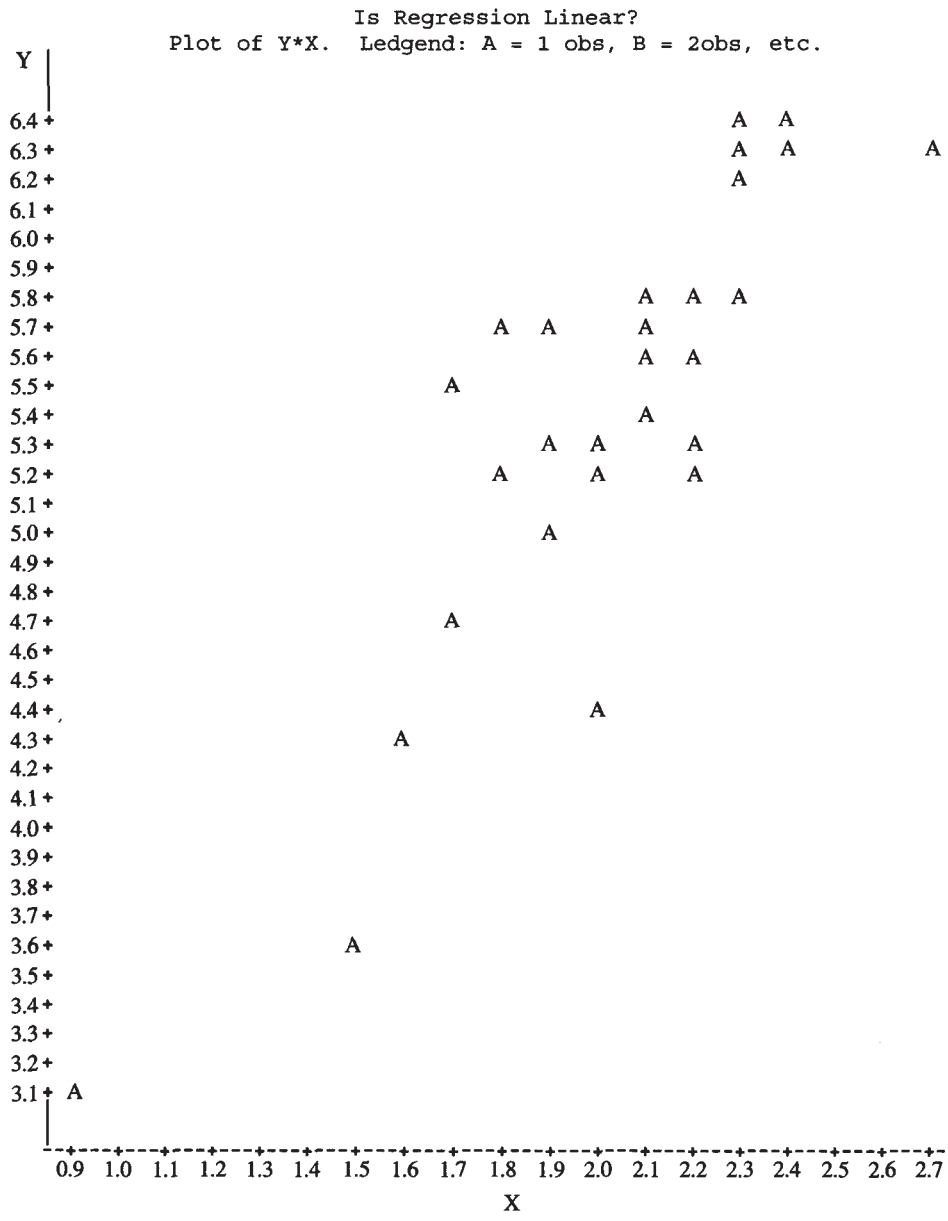
Se utilizan los datos del Ejemplo 11.1.3 para mostrar el código SAS necesario para dibujar una nube de puntos.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Definir las especificaciones de impresión.
DATA LIMPET;	Nombrar el conjunto de datos.
INPUT X Y;	Nombrar las variables.
LINES;	Señalar fin.

```
0.9     3.1
1.5     3.6
1.6     4.3                          Líneas de datos.

2.7     6.3
                                     Señal de fin de datos.
PROC PLOT;                           Dibujar datos con Y en el
PLOT Y*X;                            eje horizontal.
TITLE 'ES REGRESIÓN LINEAL?';        Titular la salida.
```

A continuación, se muestra el gráfico obtenido con este código.



XIV. Regresión lineal simple

En este ejemplo, se presenta el código SAS usado para generar la salida mostrada en el Ejemplo 11.2.3.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Definir las especificaciones de impresión.
DATA DUCKS	Nombrar el conjunto de datos.
INPUT PHOTOPRD LENGHT;	Nombrar las variables.
LINES;	Indicar que los datos vienen a continuación.
12.8 110 13.9 54 17.9 28 1.45	Líneas de datos. Pedir a SAS que prediga y cuando $x = 14.5$ Señalar el final de los datos.
PLOC PLOT; PLOT LENGHT*PHOTOPRD;	Generar la nube de puntos mostrada. en la Figura 11.10.
TITLE 'ES REGRESIÓN LINEAL?';	Titular la salida.
PROC GLM; MODEL LENGHT = PHOTOPRD/P	Generar el análisis de regresión; la variable de respuesta es la de la izquierda del signo igual en el modelo y el regresor es la de la derecha; P permite a SAS realizar prediciones basadas en la ecuación de regresión estimada.

XV. Correlación

A continuación se muestra el código SAS usado para generar la Figura 11.15. Los datos son los del Ejemplo 11.3.3.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NODATE;	Definir las especificaciones de impresión.
DATA OBESITY;	Nombrar el conjunto de datos.
INPUT PERCENT PAIN;	Nombrar las variables.
LINES;	Indicar que los datos vienen a continuación.
89 2 90 3 20 14	Líneas de datos. Señalar el final de los datos.
PLOC CORR;	Llamar al programa de correlación.
TITLE 'ESTÁN LA RESPUESTA AL DOLOR Y LA OBESIDAD CORRELADOS?';	Titular la salida.

12

Datos categóricos

En este capítulo nos ocuparemos de un tipo de análisis de datos caracterizado por el hecho de que cada una de las observaciones del conjunto de datos se puede decir que está incluida en una de varias categorías o «celdas» mutuamente excluyentes. El interés se centra en el número de observaciones que caen dentro de cada categoría. El problema estadístico consiste en determinar si las frecuencias de la categoría observada tienden a apoyar o a rechazar una hipótesis establecida.

Trataremos un problema que surge con frecuencia en la práctica: queremos presentar una prueba que nos permitirá decidir si existe o no asociación entre dos variables. La prueba se utiliza para responder a preguntas del tipo siguiente: ¿hay asociación entre el hábito de fumar y la hipertensión?, ¿entre obesidad y depresión?, ¿entre drogadicción por vía intravenosa y SIDA? Dado que preguntas de este tipo surgen con frecuencia, la prueba que se trata en este capítulo aparece citada a menudo en la bibliografía. Empezamos considerando el caso en que cada una de las variables se estudia a dos niveles.

12.1. TABLAS DE CONTINGENCIA 2 x 2

El diccionario de Webster define el término *contingencia* como «la cualidad o estado de tener conexión o relación íntima». El término *tabla de contingencia* se refiere a que las tablas construidas se usan para contrastar una asociación o relación entre dos variables. En esta sección estudiamos tablas de contingencia 2×2. Estas tablas aparecen cuando cada una de las dos variables se estudia a dos niveles. Las tablas tienen dos filas y dos columnas, que dan lugar a cuatro celdas o categorías. Cada observación del conjunto de datos cae exactamente en una celda. El análisis de datos se basa en el examen del número de observaciones que caen dentro de cada categoría. Ilustramos esta idea con el Ejemplo 12.1.1.

Ejemplo 12.1.1. En un estudio de una nueva vacuna de la hepatitis se utilizan 1083 voluntarios varones. De ellos, se eligen aleatoriamente 549 y son vacunados con el nuevo fármaco. Los restantes, 534, no son vacunados. Después de un cierto tiempo, se observó que 70 de los 534 voluntarios no vacunados contrajeron la hepatitis, mientras que solamente 11 de los 549 vacunados no la contrajeron. Estamos trabajando con dos características, el estatus vacunación y el

"¿Hola? ¿La FDA? Quisiera informar sobre una investigación que relaciona directamente el queso con la muerte de ratas".

estatus salud de cada sujeto. Cada voluntario fue vacunado o no. Del mismo modo, cada uno contrajo la hepatitis o no. Las dos características definen cuatro categorías:

Fue vacunado y contrajo la hepatitis
No fue vacunado y contrajo la hepatitis
Fue vacunado y no contrajo la hepatitis
No fue vacunado y no contrajo la hepatitis

Cada voluntario está situado en una categoría exactamente.

Puesto que nos estamos ocupando del número de observaciones que caen dentro de cada celda, necesitamos un convenio de notación para las frecuencias de éstas. También necesitamos otro para indicar el número de observaciones que caen dentro de cada nivel de cada una de las dos variables de clasificación. Utilizamos la siguiente notación:

n_{11} = número de observaciones dentro de la celda en la fila 1 y la columna 1

n_{12} = número de observaciones dentro de la celda en la fila 1 y la columna 2

n_{21} = número de observaciones dentro de la celda en la fila 2 y la columna 1

n_{22} = número de observaciones dentro de la celda en la fila 2 y la columna 2

$n_{1\cdot} = n_{11} + n_{12}$ = número de observaciones en la fila 1

$n_{2\cdot} = n_{21} + n_{22}$ = número de observaciones en la fila 2

$n_{\cdot 1} = n_{11} + n_{21}$ = número de observaciones en la columna 1

$n_{\cdot 2} = n_{12} + n_{22}$ = número de observaciones en la columna 2

n = número total de observaciones

En el Ejemplo 12.1.2 se describe la notación utilizada.

> **Ejemplo 12.1.2.** La Tabla 12.1. recoge los datos del Ejemplo 12.1.1 con la notación indicada. Obsérvese que n_1 y n_2 son totales de columnas que aparecen en los márgenes de la tabla 2 x 2. Se les llama totales de columna *marginales.* Análogamente, $n_{1.}$ y $n_{2.}$ son los totales de fila *marginales.*

La hipótesis nula a contrastar mediante una tabla de contingencia 2 x 2 es la de que «no hay asociación» entre las dos variables de clasificación. La alternativa es que hay asociación. La forma exacta de la hipótesis nula depende del diseño del experimento. Estudiamos dos propuestas experimentales diferentes que proporcionan las tablas de contingencia 2 x 2 :

1. Todos los totales marginales pueden variar sin restricciones.
2. Un grupo de totales marginales está establecido por el investigador, el otro puede variar sin restricciones.

En el caso 1, la prueba de no asociación se llama *prueba de independencia;* en el caso 2 se llama *prueba de homogeneidad.* Empezamos por considerar el primer contexto.

Prueba de independencia

En una prueba de independencia el único número que el investigador controla directamente es el tamaño total de la muestra. Se extrae una muestra de tamaño n de la población y cada objeto se clasifica según las dos variables que se estudian. Ni las frecuencias de cada celda ni los totales de fila y columna se conocen de antemano. Un ejemplo de estudio de este tipo es el Ejemplo 12.1.3.

> **Ejemplo 12.1.3.** Se realiza una investigación para determinar si hay alguna asociación aparente entre el peso de un muchacho y un éxito precoz en la escuela, a juicio de un psicólogo escolar. Se selecciona una muestra aleatoria consistente en 500 estudiantes de los grados 1 al 3. Se clasifica a cada muchacho de acuerdo con dos criterios, el peso y el éxito en la escuela. La tabla de contingencia generada es la que aparece en la Tabla 12.2. En este diseño, la única cantidad fija es n, el tamaño total de la muestra. Ambos totales marginales, fila y columna, son libres; el investigador no fija previamente ningún conjunto.

Para deducir un estadístico para la prueba de independencia, denotemos por A y B las dos variables que estamos estudiando. Cuando los dos conjuntos de marginales totales son aleatorios, la hipótesis nula de no asociación se enuncia en la forma

H_0: A y B son independientes

Tabla 12.1. La tabla de contingencia 2 x 2 para los datos del Ejemplo 12.1.1

	Vacunación		
Hepatitis	Sí	No	
Sí	$11 = n_{11}$	$70 = n_{12}$	$81 = n_{1.}$
No	$538 = n_{21}$	$464 = n_{22}$	$1002 = n_{2.}$
	$549 = n_{.1}$	$534 = n_{.2}$	$1083 = n$

Tabla 12.2. Una tabla de contingencia 2 x 2 utilizada para contrastar una asociación entre peso y precocidad en la escuela. Sólo el tamaño total de la muestra está fijado de antemano.

	Sobrepeso		
Éxito	Sí	No	
Sí			$n_{1\cdot}$ = ? (aleatorio)
No			$n_{2\cdot}$ = ? (aleatorio)
	$n_{\cdot 1}$ = ? (aleatorio)	$n_{\cdot 2}$ = ? (aleatorio)	n = 500 (fijo)

La alternativa es que hay una asociación entre *A y B,* o que *A* y *B,* no son independientes. Independencia significa que el conocimiento del nivel de clasificación de un objeto respecto a la característica A no tiene nada que ver con su nivel respecto a la característica *B.* Por ejemplo, en el Ejemplo 12.1.3 la hipótesis nula es que el sobrepeso es independiente de la precocidad en la escuela; saber que un niño es gordo no sirve para predecir éxito en la escuela. Para expresar esta idea matemáticamente utilizamos las probabilidades dadas en la Tabla 12.3.

Vimos en el Capítulo 3 que, para que dos sucesos sean independientes, la probabilidad de que ocurran ambos a la vez debe ser igual al producto de las probabilidades de que cada suceso ocurra individualmente. Obsérvese que p_{11} expresa la proporción de objetos con ambas características *A y B, p.1* la proporción con la característica *A, y* p_1 la proporción con 1a característica *B.* La definición anterior de independencia implica que para que *A y B* sean independientes

$$P[A \quad yB] = P[A]P[B]$$

o

$$p_{11} = p_{\cdot 1}\, p_{1\cdot}$$

La relación debe cumplirse para cada celda. Por tanto, la hipótesis nula de independencia se expresa matemáticamente como

$$H_0\colon p_{ij} = p_{i\cdot}\, p_{\cdot j} \qquad \begin{matrix} i = 1, 2 \\ j = 1, 2 \end{matrix}$$

Tabla 12.3. Tabla de proporciones asociadas a una tabla de contingencia 2 x 2 en la cual todos los totales marginales son aleatorios

	Con la característica *A*		
Con la característica *B*	Sí	No	
Sí	p_{11}	p_{12}	$p_{1\cdot}$
No	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	1

La alternativa es que $p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$ para algún i y j. Esta formulación de H_0, es útil para comprender el estadístico. En la práctica escribiremos simplemente

$$\boxed{\mathbf{H_0}\text{: } A \text{ y } B \text{ son independientes}}$$

al enunciar la hipótesis nula cuando los totales marginales son aleatorios.

El estadístico para contrastar H_0 se basa en una idea muy simple. Comparamos el número de observaciones en cada celda con el número esperado, si H_0 es cierta. Si estos números difieren poco, no hay razón para rechazar H_0; si hay una gran discrepancia entre los valores observados y esperados, entendemos esto como evidencia de que H_0 no es cierta. Sea E_{ij} el número esperado de observaciones en la celda ij bajo la hipótesis de que A y B son independientes. Como p_{ij} es la proporción (o porcentaje) teórica de observaciones en la celda ij, el número esperado se calcula multiplicando esta proporción por el número total de observaciones. Es decir,

$$E_{ij} = np_{ij}$$

Las proporciones p_{11}, p_{12}, p_{21} y p_{22} no se conocen y hay que estimarlas a partir de los datos bajo el supuesto de que la hipótesis nula es cierta.

¿Cómo se puede hacer esto? ¡Muy simple! Obsérvese por ejemplo que, si H_0 es cierta y las características A y B son independientes, entonces

$$p_{11} = p_{1\cdot}\, p_{\cdot 1}$$

Puesto que p_1 es la probabilidad de que una observación caiga en la fila 1, es lógico estimar p_1 mediante

$$\hat{p}_{1\cdot} = \frac{\text{número de elementos en la fila 1}}{\text{tamaño de la muestra}} = \frac{n_{1\cdot}}{n}$$

Análogamente, puesto que $p_{\cdot 1}$ es la probabilidad de que una observación caiga en la columna 1, estimamos $p_{\cdot 1}$ mediante

$$\hat{p}_{\cdot 1} = \frac{\text{número de elementos en la columna 1}}{\text{tamaño de la muestra}} = \frac{n_{\cdot 1}}{n}$$

Se tiene, por lo tanto,

$$\hat{p}_{11} = \hat{p}_{1\cdot}\, \hat{p}_{\cdot 1} = \frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n}$$

Esto, a su vez, implica que

$$\hat{E}_{11} = \hat{p}_{11}\, n = \frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n}\, n$$

Obsérvese que

$$\widehat{E}_{11} = \frac{n_{1\cdot} n_{\cdot 1}}{n} = \frac{\left(\begin{matrix}\text{total fila}\\ \text{marginal}\end{matrix}\right)\left(\begin{matrix}\text{total columna}\\ \text{marginal}\end{matrix}\right)}{\text{tamaño de la muestra}}$$

Un argumento semejante es válido para las probabilidades de otras celdas.

Concluimos así que para cada i y j,

$$\boxed{\widehat{E}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} = \frac{\left(\begin{matrix}\text{total fila}\\ \text{marginal}\end{matrix}\right)\left(\begin{matrix}\text{total columna}\\ \text{marginal}\end{matrix}\right)}{\text{tamaño de la muestra}}}$$

El cálculo de las frecuencias esperadas de las celdas se ilustra en el siguiente ejemplo:

Ejemplo 12.1.4. Cuando se realiza el experimento del Ejemplo 12.1.3 se hallan los datos de la Tabla 12.4.

Las frecuencias esperadas de las celdas, bajo H_0son

$$\widehat{E}_{11} = \frac{n_{1\cdot} n_{\cdot 1}}{n} = \frac{425(200)}{500} = 170$$

$$\widehat{E}_{12} = \frac{n_{1\cdot} n_{\cdot 2}}{n} = \frac{425(300)}{500} = 225$$

$$\widehat{E}_{21} = \frac{n_{2\cdot} n_{\cdot 1}}{n} = \frac{75(200)}{500} = 30$$

$$\widehat{E}_{22} = \frac{n_{2\cdot} n_{\cdot 2}}{n} = \frac{75(300)}{500} = 45$$

Resumimos la situación en la Tabla 12.5. Obsérvese que hay algunas diferencias entre lo que se espera si H_0 es cierta (listado entre paréntesis) y lo que realmente se observa. La pregunta es la siguiente: ¿son estas diferencias demasiado grandes como para que se deban únicamente al azar?

Tabla 12.4. Datos utilizados para contrastar asociación entre obesidad y precocidad en la escuela

	Sobrepeso		
Éxito	Sí	No	
Sí	$n_{11} = 162$	$n_{12} = 263$	$n_{1\cdot} = 425$
No	$n_{21} = 38$	$n_{22} = 37$	$n_{2\cdot} = 75$
	$n_{\cdot 1} = 200$	$n_{\cdot 2} = 300$	$n = 500$

Tabla 12.5. Datos utilizados para contrastar asociación entre obesidad y precocidad en la escuela. Las frecuencias esperadas aparecen entre paréntesis.

	Sobrepeso		
Éxito	**Sí**	No	
Sí	162 (170)	263 (255)	425
No	38 (30)	37 (45)	75
	200	300	500

Para responderla necesitamos un estadístico cuya distribución de probabilidad sea conocida bajo la hipótesis de que H_0, es cierta. El estadístico en cuestión es

$$X_1^2 = \sum_{\text{todas las celdas}} \frac{(n_{ij} - \widehat{E}_{ij})^2}{\widehat{E}_{ij}}$$

Obsérvese que el numerador de este estadístico compara las frecuencias observadas con las frecuencias esperadas de cada celda, hallando la diferencia entre las dos. Las diferencias se elevan al cuadrado para que al sumar las negativas no cancelen las positivas. Si hay un buen ajuste, las diferencias serán pequeñas, el numerador será pequeño y el valor del estadístico será pequeño. En este caso, no se rechaza H_0. Si existen grandes discrepancias entre las frecuencias observadas y las esperadas, las diferencias serán grandes, el numerador será grande y el valor observado del estadístico será grande. En este caso, se rechaza H_0. Como el valor P de la prueba se halla por medio de la distribución ji-cuadrado con un grado de libertad y la prueba se basa en la apreciación del grado de ajuste entre las frecuencias esperadas y observadas, en el supuesto de que H_0 es cierta, esta prueba se llama prueba de la *ji-cuadrado de bondad de ajuste.* Se ilustra su utilización completando el análisis de los datos del Ejemplo 12.1.3.

Ejemplo 12.1.5. Considérese las frecuencias observadas y las esperadas estimadas dadas en la Tabla 12.5. El valor observado del estadístico es

$$\sum_{\text{todas las celdas}} \frac{(n_{ij} - \widehat{E}_{ij})^2}{\widehat{E}_{ij}} = \frac{(162 - 170)^2}{170} + \frac{(263 - 255)^2}{255} + \frac{(38 - 30)^2}{30} + \frac{(37 - 45)^2}{45} = 4.18$$

El valor P es

$$P = P[X_1^2 \geq 4.18]$$

A partir de la tabla de la ji-cuadrado vemos que

$$P[X_1^2 \geq 3.84] = 0.05 \quad \text{y} \quad P[X_1^2 \geq 5.02] = 0.025$$

Como 4.18 está entre 3.84 y 5.02, el valor P está entre 0.025 y 0.05. El valor P es pequeño. Podemos rechazar H_0 y concluir que obesidad y precocidad en la escuela no son independientes.

Para que las pruebas ji-cuadrado de bondad de ajuste sean válidas, es necesario que las muestras sean grandes. Esto lleva de manera natural a la pregunta «¿cuánto de grandes?». Hay varias opiniones como respuesta a esta pregunta. Sin embargo, habitualmente se considera que n es suficientemente grande cuando ninguna frecuencia esperada es menor que 1 y no más del 20 % son menores que 5. En el caso de una tabla 2 x 2, estas cifras orientativas se alcanzan sólo cuando ninguna frecuencia esperada es menor que 5. Si esto no se cumple, puede utilizarse una prueba para pequeñas muestras llamada prueba exacta de Fisher [2] para contrastar independencia.

Prueba de homogeneidad

Consideramos ahora la segunda configuración experimental, en la que un conjunto de marginales totales está fijado por el investigador, mientras que el otro es aleatorio. Esta situación aparece cuando hay dos poblaciones en estudio. Estamos interesados en una característica particular, y queremos responder a la pregunta «¿es igual la proporción de objetos con la misma característica en las dos poblaciones?». Si no hay asociación entre la característica y la población a la cual pertenece un determinado individuo, entonces la proporción con la característica debería ser la misma en cada caso. Si hay asociación, las proporciones deberían ser diferentes. El Ejemplo 12.1.6 es un ejemplo de este tipo de problema.

Ejemplo 12.1.6. Un gran número de personas que viven en una sección determinada de una comunidad han estado expuestas durante los últimos diez años a la radiactividad procedente de un vertedero en el que se almacenan desechos atómicos. Se realiza una investigación para descubrir si hay alguna asociación aparente entre la exposición y el desarrollo de una cierta enfermedad de la sangre. Para llevar a cabo el experimento se eligen muestras aleatorias de 300 personas de la comunidad que han estado expuestas al peligro y 320 no expuestas. Estamos trabajando con muestras de las dos poblaciones, los expuestos a la radiactividad del vertedero y los no expuestos a ella. Se estudia cada sujeto para determinar si tiene la enfermedad. El experimento genera una tabla 2 x 2 de la forma de la Tabla 12.6. Obsérvese que los totales de fila marginales se han fijado en 300 y 320, ya que estos tamaños de muestras son

Tabla 12.6. Una tabla de contingencia 2 x 2 con totales de filas, que presentan muestras aleatorias de dos poblaciones, fijados por el investigador

	Tiene la enfermedad		
Expuesto a la radiactividad	**Sí**	**No**	
Sí			$n_{1.} = 300$ (fijado previamente al experimento)
No			$n_2 = 320$ (fijado previamente al experimento)
	$n_1 = ?$ (aleatorio)	$n_2 = ?$ (aleatorio)	$n = 620$

determinados previamente por el investigador. Los totales columna marginales son libres; es decir, son variables aleatorias cuyos valores numéricos sólo se conocen al final del experimento.

Si no hay asociación entre exposición y desarrollo de la enfermedad, la proporción de personas con la enfermedad debería ser la misma en las dos poblaciones. Si hay asociación, estas proporciones podrían ser distintas.

Las proporciones dadas en la Tabla 12.7 se utilizan para expresar la hipótesis nula y comprender el estadístico utilizado para contrastar H_0. La hipótesis nula de no asociación entre población (variable *B)* y característica (variable A) es

$$\boxed{H_0: P_{11} - P_{21}}$$

Obsérvese que si H_0 es cierta, también es cierto que $p_{12} = p_{22}$ En el Ejemplo 12.1.6 esta hipótesis nula toma la forma

H_0: proporción de personas con la enfermedad entre los expuestos a la radiactividad = proporción de personas con la enfermedad entre los no expuestos.

En la práctica, podemos expresar la hipótesis nula como

H_0: la proporción con la característica es la misma en cada población.

La prueba que se está llevando a cabo es una prueba de homogeneidad. La palabra *homogéneo* significa «de igual naturaleza». Estamos realizando un contraste para ver si las dos poblaciones de las que se extrajeron muestras son iguales en el sentido de que la proporción de objetos con la característica es la misma en cada población.

Para comprender la lógica subyacente al cálculo de esperanzas, obsérvese que la frecuencia esperada de observaciones de la población 1 que poseen la característica se halla multiplicando el tamaño de la muestra de la población 1 por la probabilidad teórica de tener la característica, es decir, $E_{11} = n_1.p_{11}$. De igual modo, $E_{21} = n_2.p_{21}$ .Así, para estimar E_{11} y E_{2l} a partir de la tabla de contingencia, sólo nos hace falta saber estimar p_{11} y p_{2l}. Esto no es difícil. Si H_0 es cierta, $p_{11} = p_{2l}$. Denotamos porp esta proporción poblacional común. Además, si la proporción de objetos con la característica es la misma para ambas poblaciones, la proporción

Tabla 12.7. Tabla de proporciones asociadas a una tabla de contingencia 2 x 2 en la cual los totales de las filas están fijados

	Tienen la característica		
Población	Sí	**No**	
1	p_{11}	$p_{12} = 1 - p_{11}$	$n_{1\cdot}$ (fijo)
2	p_{21}	$p_{22} = 1 - p_{21}$	$n_{2\cdot}$ (fijo)
	$n_{\cdot 1}$ (aleatorio)	$n_{\cdot 2}$ (aleatorio)	n

global de objetos en las dos poblaciones combinadas será también p. Un estimador lógico para la proporción global de objetos con la característica es

$$\hat{p} \quad \frac{\text{número de elementos en la columna 1}}{\text{tamaño de la muestra global}} \quad \frac{n_{\cdot 1}}{n}$$

Como estamos suponiendo que $p_{11} = p_{21} = p$, podemos utilizar también $\hat{p}$ como estimador de P_{11} y P_{21}. Sustituyendo, obtenemos las frecuencias esperadas de las celdas bajo la hipótesis de que es cierta H_0:

$$\hat{E}_{11} = n_{1\cdot}\,\hat{p}_{11} = n_{1\cdot}\,\frac{n_{\cdot 1}}{n} = \frac{\begin{pmatrix}\text{total fila}\\ \text{marginal}\end{pmatrix}\begin{pmatrix}\text{total columna}\\ \text{marginal}\end{pmatrix}}{\text{tamaño de la muestra}}$$

$$\hat{E}_{21} = n_{2\cdot}\,\hat{p}_{21} = n_{2\cdot}\,\frac{n_{\cdot 1}}{n} = \frac{\begin{pmatrix}\text{total fila}\\ \text{marginal}\end{pmatrix}\begin{pmatrix}\text{total columna}\\ \text{marginal}\end{pmatrix}}{\text{tamaño de la muestra}}$$

De la misma manera se calculan $\hat{E}_{12}$ y $\hat{E}_{22}$. Obsérvese que estas esperanzas se hallarían igual que las obtenidas para la prueba de independencia. Desde este punto de vista, la prueba de homogeneidad es idéntica a la de independencia.

Ejemplo 12.1.7. La Tabla 12.8 da las frecuencias observadas y esperadas (entre paréntesis) para el experimento descrito en el Ejemplo 12.1.6. El valor observado del estadístico del contraste es

$$\sum_{\text{todas las celdas}} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = 0.62$$

Como $P = P[X_1^2 \geq 0.62] > 0.10$, no podemos rechazar H_0. No hay evidencia de asociación entre enfermedad sanguínea y exposición a esta fuente de radiactividad.

Tabla 12.8. Datos utilizados para contrastar asociación entre las poblaciones de los expuestos a la radiactividad y los no expuestos a ella

	Tiene la enfermedad		
Expuesto a la radiactividad	Sí	No	
Sí	52 (48.39)	248 (251.61)	300 (fijo)
No	48 (51.61)	272 (268.39)	320 (fijo)
	100	520	620

EJERCICIOS 12.1

1. Utilizar los datos del Ejemplo 12.1.1 para determinar si el hecho de contraer hepatitis es independiente de haber sido vacunado contra la enfermedad.
2. Se realiza una investigación sobre una nueva vacuna contra la gripe. Se elige una muestra aleatoria de 900 individuos y se clasifica a cada uno de ellos según haya contraído la gripe durante el último año o no, y según haya sido o no vacunado. Se obtiene la información que se muestra en la Tabla 12.9.
   a) ¿Son fijos algunos de los totales marginales?
   b) Construir la hipótesis nula apropiada para contrastar la asociación entre variables.
   c) Hallar la frecuencia esperada para cada celda.
   d) ¿Puede rechazarse H_0? ¿Cuál es el valor P del contraste?
3. En un estudio para determinar la asociación, si hay alguna, entre la rubéola materna y las cataratas congénitas, se selecciona una muestra de 20 niños con este defecto y 25 niñas con antecedentes y edad semejantes que no lo presentan. Se entrevista a la madre de cada niño para determinar si tuvo o no la rubéola durante el embarazo. Se obtienen los datos que se muestran en la Tabla 12.10.
   a) ¿Son fijos algunos de los totales marginales?
   b) Construir la hipótesis nula apropiada para contrastar la asociación entre variables.
   c) Hallar la frecuencia esperada para cada celda.
   d) Contrastar la hipótesis nula.
4. Se ha creído durante mucho tiempo que las úlceras pépticas se debían tanto a factores genéticos como ambientales. Una investigación reciente considera la asociación entre el nivel de pepsinógeno I en sangre y la presencia de úlceras pépticas; se compararon 14 pacientes que tuvieron úlceras duodenales con 49 individuos que no la tuvieron. Se determinó el nivel de pepsinógeno I en sangre de cada uno y se clasificó como alto o bajo. En la Tabla 12.11 se muestran los datos que se obtuvieron.
   a) Completar la tabla.
   b) Construir la hipótesis nula apropiada para contrastar la asociación entre variables.
   c) Hallar el número esperado de observaciones para cada celda.
   d) Contrastar la hipótesis nula.
5. Se realiza un pequeño estudio piloto para determinar la asociación entre la aparición de leucemia y los antecedentes de alergia. Se selecciona una muestra de 19 pacientes con leucemia y 17 controles, y se determina la existencia o no de antecedentes de alergia. En la Tabla 12.12 se recogen los datos.
   a) ¿Son fijos algunos de los totales marginales?
   b) Construir la hipótesis nula apropiada para contrastar la asociación entre las variables de clasificación.
   c) Hallar la frecuencia esperada para cada celda.

Tabla 12.9.

	Contraída la gripe		
Vacunado	**Sí**	**No**	
Sí	**150**	200	
No	300	**250**	

Tabla 12.10.

	La madre tuvo rubéola		
Tiene cataratas congénitas	**Sí**	**No**	
Sí	**14**	**6**	
No	**10**	**15**	

d) ¿Hay prueba de una asociación entre el antecedente de alergia y la leucemia? Explicar la respuesta basándose en el valor *P* del contraste.

6. Se realiza un estudio sobre la asociación entre tipo de hospital y muerte en el hospital después de una operación de alto riesgo, durante el mes de julio. Fueron seleccionados para su estudio 139 pacientes con operaciones de alto riesgo en hospitales universitarios; otros 528 pacientes fueron escogidos de otros tipos de hospitales. De los pacientes tratados en hospitales universitarios, murieron 32. De la muestra extraída de los otros hospitales murieron 62.
   α) Construir una tabla de contingencia 2 x 2 que recoja estos datos.
   b) Contrastar la hipótesis nula de no asociación. ¿Se trata de una prueba de independencia o de homogeneidad?
   c) Estimar la probabilidad de muerte en el hospital de tales pacientes en cada tipo de hospital durante el mes de julio.
   d) Los datos proceden de un artículo titulado «No es bueno ponerse enfermo en julio» ¿Está de acuerdo? Explicarlo. (Mark Blumberg, «It's Not OK to Get Sick in July», *Journal of the American Medical Association,* agosto 1, 1990, pág. 573.)
7. Se elige una muestra de 245 pacientes de menos de 19 años atendidos en una clínica por alergia. Cada paciente se clasifica por la edad y por haber presentado o no alergia a los huevos. De 133 pacientes de más de 3 años, 30 eran alérgicos a los huevos, 32 de los 112 pacientes de 3 años o menos mostraron esta alergia.
   a) Construir una tabla de contingencia 2 x 2 que recoja estos datos.
   b) Contrastar la hipótesis nula de no asociación entre edad y alergia a los huevos.
   c) ¿Se trata de una prueba de independencia o de homogeneidad?
   (Basado en información obtenida de Alian Bock and F. M. Atkins, «Patterns of Food Hypersensitivity During Sixteen Years of Double-Blind, Placebo-Controlled Food Challenges», *The Journal of Pediatrics,* octubre 1990, págs. 561-567.)
8. La proteína TAT es una proteína producida por las células infectadas por el VIH-1. Se lleva a cabo un estudio para contrastar asociación entre presencia de anticuerpos TAT y

Tabla 12.11.

	Nivel de pepsinógeno		
Úlcera	**Alto**	**Bajo**	
Presente	**12**		**14**
Ausente		**31**	**49**

Tabla 12.12.

	Antecedentes de alergia	**Antecedentes de alergia**	
Control	5	12	
Paciente	17	2	

el sarcoma de Kaposi en pacientes de SIDA. Se analiza el suero de 297 pacientes seropositivos para el VIH-1 dentro del período de un mes desde la diagnosis del SIDA. Cada muestra se clasifica según contenga el sarcoma de Kaposi o no, y según contenga anticuerpos TAT o no. De 78 pacientes de la muestra que tenían sarcoma, 10 presentaban anticuerpos TAT; de los 219 pacientes sin sarcoma, 21 tenían anticuerpos TAT.

a) Construir una tabla de contingencia 2 x 2 que muestre estos datos.

b) Contrastar si hay asociación entre presencia de anticuerpo y presencia de sarcoma de Kaposi. ¿Se trata de una prueba de independencia o de homogeneidad?

(Basado en cifras dadas por Peter Reiss and Joseph Lange «Kaposi's Sarcoma and AIDS», *Nature,* agosto 30, 1990, pág. 801.)

9. En una tabla de contingencia 2 x 2, la prueba de homogeneidad es una comparación de dos proporciones. Recuerde que el contraste Zpara comparar dos proporciones ha sido expuesto en el Capítulo 8. Se puede probar que el estadístico ji-cuadrado definido aquí es el cuadrado del estadístico Z del contraste conjunto de dicho capítulo. Para comprobarlo, analice los datos del Ejercicio 12.1.7 por medio del método Z del Capítulo 8, eleve al cuadrado el valor Z obtenido y verifique que z^2 es igual al valor de x^2 hallado en el Ejercicio 12.1.7.

12.2. TABLAS DE CONTINGENCIA *rxc*

En esta sección ampliamos los métodos de la Sección 12.1 a situaciones en las que el número de filas o columnas en la tabla de contingencia es mayor que 2. El propósito del experimento es el mismo de antes, es decir, comprobar la asociación entre las variables de clasificación, comparando las frecuencias observadas de las celdas con las esperadas si no existe asociación. Se utilizan dos variables de clasificación A y *B*. Suponemos que hay *c* niveles relativos a la variable A y *r* niveles relativos a *B*. De este modo, la tabla de contingencia generada tiene *r* filas, *c* columnas, y *rc* celdas o categorías. La notación utilizada se detalla en la Tabla 12.13. Obsérvese que

n_{ij} = frecuencia observada en la (ij)-ésima celda

$n_{i.}$ = total fila marginal para la i-ésima fila, $i = 1, 2, ..., r$

$n_{.j}$ = total columna marginal para y-ésima columna, $j = 1, 2, ..., c$

Describimos la utilización de esta notación en el Ejemplo 12.2.1

Ejemplo 12.2.1. Se realiza un estudio para determinar si existe asociación entre el grupo sanguíneo y las úlceras duodenales. Se selecciona una muestra de 1301 pacientes y 6313 controles y se determina el grupo sanguíneo de cada uno. Entre los pacientes, 698 son del grupo 0; 472, del grupo A; 102, del grupo B, y el resto, del grupo AB. Entre los controles, las

Tabla 12.13. Tabla de frecuencias asociadas con una tabla de contingencia *r x c.* La tabla tiene r filas y *c* columnas

	Variable *A*					
Variable *B*	**1**	*2*	3	⋯	*c*	
1	n_{11}	n_{12}	n_{13}	⋯	n_{1c}	$n_{1\cdot}$
2	n_{21}	n_{22}	n_{23}	⋯	n_{2c}	$n_{2\cdot}$
⋮	⋯	⋯	⋯	⋯	⋯	⋯
r	n_{r1}	n_{r2}	n_{r3}	⋯	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	⋯	$n_{\cdot c}$	n

cifras son 2892,2625,570 y 226, respectivamente. Estos datos se disponen convenientemente en una tabla de contingencia 2 x 4 (dos filas y cuatro columnas) tal como se muestra en la Tabla 12.14.

Una vez más, hay dos configuraciones experimentales.

1. Todos los totales marginales pueden variar libremente.
2. Un conjunto de marginales totales está fijado por el investigador; el otro puede variar libremente.

Como en el caso de las tablas 2 x 2, la configuración 1 corresponde a una prueba de independencia, mientras que la configuración 2 conduce a una prueba de homogeneidad. La Tabla 12.15 muestra la notación para las proporciones teóricas asociadas a una tabla de contingencia *r* x *c*. Esta notación nos permite expresar simbólicamente la hipótesis nula de no asociación. Si todos los totales marginales son aleatorios, entonces la hipótesis nula de no asociación entre las variables de clasificación se expresa como

$$H_0: A \text{ y } B \text{ son independientes}$$

Estadísticamente, esto se indica de la forma siguiente.

$$H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j} \qquad \begin{matrix} i = 1, 2, ..., r \\ j = 1, 2, ..., c \end{matrix}$$

Tabla 12.14. Tabla de contingencia 2 x 4 que muestra los datos del Ejemplo 12.2.1

	Grupo sanguíneo				
	0	**A**	**B**	**AB**	
Pacientes	$n_{11} = 698$	$n_{12} = 472$	$n_{13} = 102$	$n_{14} = 29$	$n_{1\cdot} = 1301$ (fijo)
Controles	$n_{21} = 2892$	$n_{22} = 2625$	$n_{23} = 570$	$n_{24} = 226$	$n_{2\cdot} = 6313$ (fijo)
	$n_{\cdot 1} = 3590$	$n_{\cdot 2} = 3097$	$n_{\cdot 3} = 672$	$n_{\cdot 4} = 255$	$n = 7614$

Tabla 12.15. Proporciones asociadas a una tabla de contingencia r x c. En esta notación el primer subíndice denota el número de fila y el segundo el número de columna

	Variable *A*					
Variable *B*	**1**	**2**	**3**	⋯	*c*	
1	p_{11}	p_{12}	p_{13}	⋯	p_{1c}	
2	p_{21}	p_{22}	p_{23}	⋯	p_{2c}	
3	p_{31}	p_{32}	p_{33}	⋯	p_{3c}	
⋮	⋯	⋯	⋯	⋯	⋯	
r	p_{r1}	p_{r2}	p_{r3}	⋯	p_{rc}	

Como en el caso de la tabla 2 x 2 , esto expresa la idea de que, si A y B son independientes, la probabilidad p_{ij} de caer en la celda (y) es la probabilidad p_i de caer en la fila i multiplicada por la probabilidad p_j de caer en la columna/ Si $r = c = 2$, esto se reduce a la formulación presentada en la Sección 12.1.

Si los totales de fila están fijados, la hipótesis nula de no asociación se indica en la forma

H_0: el porcentaje entre los niveles de la variable A es el mismo en cada población

Estadísticamente, esta hipótesis se expresa en términos de las proporciones de la Tabla 12.15. Toma la forma

$$H_0: \begin{cases} p_{11} = p_{21} = \cdots = p_{r1} \text{ (las proporciones en la columna 1 son idénticas)} \\ p_{12} = p_{22} = \cdots = p_{r2} \text{ (las proporciones en la columna 2 son idénticas)} \\ \vdots \\ p_{1c} = p_{2c} = \cdots = p_{rc} \text{ (las proporciones en la columna } c \text{ son idénticas)} \end{cases}$$

De nuevo, si $r = c = 2$, esto se reduce a la formulación dada en la Sección 12.1.

El estadístico para contrastar la hipótesis nula de no asociación, en cualquiera de los dos diseños, es

$$X^2_{(r-1)(c-1)} = \sum_{\text{todas las celdas}} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

donde $\hat{E}_{ij}$ es la frecuencia esperada estimada en la celda (ij). Como antes,

$$\hat{E}_{ij} = \frac{\text{(total fila marginal) (total columna marginal)}}{\text{tamaño de la muestra}}$$

Obsérvese que el número de grados de libertad asociados con este estadístico ji-cuadrado es (r - l)(c - 1), donde *r* es el número de filas de la tabla y *c* el número de columnas. En una tabla 2 x 2 el número de grados de libertad es (2 - 1)(2 - 1) = 1, como se afirmó en la Sección 12,1.

La prueba rechazará H_0 para valores del estadístico demasiado grandes para ser atribuidos al azar. La prueba se aplica a muestras suficientemente grandes para que ninguna frecuencia esperada sea menor que 1 y no más del 20 % sean menores que 5. Ilustramos este procedimiento en el Ejemplo 12.2.2.

Ejemplo 12.2.2. Cuando se analizan los datos del Ejemplo 12.2.1, se obtienen las frecuencias observadas y esperadas que se recogen en la Tabla 12.16. El valor observado del estadístico es

$$\frac{(698 - 613.42)^2}{613.42} + \frac{(2892 - 2976.58)^2}{2976.58} + \cdots + \frac{(226 - 221.43)^2}{211.43} = 29.12$$

El número de grados de libertad es

$$(r-1)(c-1) = (2-1X4-1) = 3$$

En la tabla ji-cuadrado puede verse que

$$P = [X_3^2 \geq 29.12] < 0.005$$

La hipótesis nula de no asociación entre grupo sanguíneo y presencia de úlceras duodenales puede ser rechazada. En este estudio, el investigador fijó el número de pacientes en 1301 y el número de controles en 6313. Por tanto, la prueba de no asociación es una prueba de homogeneidad. Tratamos de ver si hay diferencia en la distribución de grupos sanguíneos entre pacientes y controles. Esta hipótesis se puede expresar simbólicamente como

$$\begin{cases} P_{11} = P_{21} & \text{(porcentaje del grupo 0 entre pacientes = porcentaje del grupo 0 entre controles)} \\ P_{12} = P_{22} & \text{(porcentaje del grupo A entre pacientes = porcentaje del grupo A entre controles)} \\ p_{13} = p_{23} & \text{(porcentaje del grupo B entre pacientes = porcentaje del grupo B entre controles)} \\ P_{14} = P_{24} & \text{(porcentaje del grupo AB entre pacientes = porcentaje del grupo AB entre controles)} \end{cases}$$

Tabla 12.16. Datos utilizados para contrastar la asociación entre grupo sanguíneo y presencia de úlcera duodenal. Los totales de filas están fijados por el investigador

	Grupo sanguíneo				
	0	**A**	**B**	AB	
Pacientes	698 (613.42)	472 (529.18)	102 (114.82)	29 (43.57)	1301 (fijo)
Controles	2892 (2976.58)	2625 (2567.82)	570 (557.18)	226 (211.43)	6313 (fijo)
	3590	3097	672	255	7614

Como se rechaza H_0, concluimos que hay diferencias en estos porcentajes. Para tener una idea intuitiva de qué diferencias existen entre pacientes y controles, observemos que la probabilidad estimada de que un paciente esté en el grupo sanguíneo 0 viene dada por:

$$\frac{\text{Número de pacientes en el grupo 0}}{\text{Número de pacientes}} \quad \frac{698}{1301}$$

Análogamente, la probabilidad de que un control esté en el grupo 0 es

$$\frac{\text{Número de controles en el grupo 0}}{\text{Número de controles}} = \frac{2892}{6313} = 0.46$$

Parece, pues, que la proporción de personas del grupo sanguíneo 0 no es la misma entre los pacientes que entre los controles. Es decir:

$$p_{11} \neq p_{21}$$

EJERCICIOS 12.2

1. Considérense los datos del Ejemplo 12.2.2. Estimar la proporción de pacientes y la proporción de controles que presentan el grupo sanguíneo A. Hágase lo mismo para los grupos sanguíneos B y AB. Comentar los resultados.
2. Para intentar convencer al público de que utilice equipos de seguridad en los automóviles, se elige de los archivos una muestra aleatoria de 1000 accidentes. Se clasifica cada accidente de acuerdo con el tipo de sujeción de seguridad utilizado por los ocupantes y la gravedad de las lesiones sufridas. Resultaron los datos que se muestran en la Tabla 12.17.
   a) ¿Hay algún total marginal fijado por el investigador?
   b) ¿Se trata de una prueba de independencia o de homogeneidad?
   c) Establecer y contrastar la hipótesis nula apropiada de no asociación. ¿Cuál es el valor *P?*
   d) Estimar la probabilidad de muerte cuando no se utiliza sujeción alguna, cuando se utiliza cinturón en el asiento y correajes, y cuando sólo se utiliza cinturón. Comente las implicaciones prácticas de sus averiguaciones.

Tabla 12.17. Datos utilizados para contrastar asociación entre el tipo de sujeción utilizado y la gravedad de las heridas sufridas en una muestra de accidentes de coche

	Tipo de sujeción			
Magnitud de la lesión	**Cinturón de seguridad**	**Cinturón de seguridad y correaje**	**Ninguna**	
Ninguna	**75**	60	65	
Menor	160	115	175	
Mayor	100	65	135	
Muerte	**15**	10	25	
				1000

3. Se realiza un estudio para investigar la asociación entre el color de las flores y la fragancia de las azaleas silvestres. Se observan doscientas plantas floridas seleccionadas aleatoriamente. Cada una de ellas se clasifica según el color y la presencia o ausencia de fragancia. Los datos se dan en la Tabla 12.18.
   α) ¿Hay algún total marginal fijado por el investigador?
   b) ¿Se trata de una prueba de independencia o de homogeneidad?
   c) Enunciar la hipótesis nula apropiada.
   d) ¿Puede rechazarse H_0? ¿Cuál es el valor *P* de la prueba? ¿Qué conclusiones pueden obtenerse de estos datos?
4. En un estudio sobre bocio se seleccionan muestras aleatorias de determinado tamaño de 10 estados. A las personas elegidas se las examina para ver si tienen bocio y se registra el número de casos hallados. Los datos obtenidos aparecen en la Tabla 12.19.
   a) La prueba de asociación entre procedencia geográfica y bocio, ¿es una prueba de independencia o de homogeneidad?
   b) Enunciar la hipótesis nula de interés.
   c) Contrastar H_0.
   d) Estimar la probabilidad de que un individuo de California, elegido aleatoriamente, tenga bocio. Estimar la probabilidad de que un individuo de Massachusetts, aleatoriamente seleccionado, tenga bocio.
5. Se realiza un estudio para considerar la asociación entre el nivel de dióxido sulfúrico del aire y el número medio de cloroplastos por célula en las hojas de los árboles de la zona. Se elige una muestra de 10 zonas de las que se sabe que tienen una alta concentración de dióxido sulfúrico, 10 que se sabe tienen un nivel normal del mismo y 10 que tienen una baja concentración. Dentro de cada zona se seleccionan aleatoriamente veinte árboless y se determina para cada árbol el número de cloroplastos por célula en las hojas. Sobre esta base se clasifica cada árbol según tenga un recuento bajo, normal o alto de cloroplastos. Se obtienen los datos que se muestran en la Tabla 12.20.
   α) Establecer la hipótesis nula apropiada.
   b) Contrastar la hipótesis nula.
   c) Estimar la probabilidad de que un árbol aleatoriamente seleccionado tenga un nivel bajo de cloroplastos, supuesto que es de una zona con una alta concentración de dióxido sulfúrico. Estimar la probabilidad de que un árbol aleatoriamente seleccionado de una zona de concentración normal de dióxido sulfúrico tenga un nivel bajo de cloroplastos. Estimar esta probabilidad para un árbol aleatoriamente seleccionado que crece en una zona de baja concentración de dióxido sulfúrico. Comentar las implicaciones prácticas de estas estimaciones.

Tabla 12.18. Datos utilizados para contrastar asociación entre color y fragancia de las azaleas

	Color de la flor			
Fragancia	Blanca	Rosa	Naranja	
Sí	12	60	58	
No	50	10	10	

Tabla 12.19. Datos utilizados para contrastar asociación entre procedencia geográfica y bocio

	Diagnóstico de bocio		
	Positivo	**Negativo**	
California	**36**		500
Kentucky	**17**		350
Louisiana	**12**		300
Massachusetts	**1**		300
Michigan	**4**		350
New York	**14**		500
South Carolina	**7**		200
Texas	**27**		500
Washington	**2**		200
West Virginia	**4**		200
			3400

6. Se realiza un estudio para averiguar los factores que influyen en la decisión de un médico al ordenar una transfusión a un paciente. Se seleccionó una muestra de 49 médicos responsables de hospitales. A cada médico se le preguntó sobre la frecuencia con que se han hecho transfusiones innecesarias debido a las sugerencias de otro médico. La misma pregunta se hizo con una muestra de 71 médicos residentes. Los datos se muestran en la Tabla 12.21.
   a) Hacer el contraste con la hipótesis nula de no asociación.
   b) ¿Es una prueba de independencia o de homogeneidad?

Tabla 12.20. Datos utilizados para contrastar asociación entre nivel de cloroplasto y exposición al dióxido de azufre

	Nivel de cloroplastos			
Nivel de SO_2	**Alto**	**Normal**	**Bajo**	
Alto	3	**4**	13	20
Normal	5	**10**	5	20
Bajo	7	**11**	2	20

Tabla 12.21. Datos utilizados para el contraste de no asociación entre tipo de médico y tendencia a autorizar transfusiones

	Frecuencia de transfusiones innecesarias					
Tipo de médico	**Muy frecuentemente (1 por semana)**	**Frecuentemente (1 cada 2 semanas)**	**Ocasionalmente (1 por mes)**	**Raramente (1 cada 2 meses)**	**Nunca**	
Responsable	1	1	3	31	13	49
Residente	2	13	28	23	5	71

c) Estimar la probabilidad de que un médico responsable nunca ordene una transfusión innecesaria; lo mismo para un médico residente.

(Datos encontrados en Susanne Salem-Schatz, Jerry Avorn y Stephen Soumerai, «Influence of Clinical Knowledge, Organizational Context and Practice Style on Transfusion Decision Making», *Journal of the American Medical Association,* julio 25, 1990, págs. 476-483.)

7. En un estudio dirigido a investigar el efecto de la presencia de una gran planta industrial sobre la población de invertebrados en un río que atraviesa la planta, se tomaron muestrias aguas arriba y aguas abajo de la planta. Se obtuvieron los siguientes datos:

	Especies						
	A	**B**	**C**	**D**	**E**	**F**	**G**
Aguas arriba	37	12	6	**18**	**7**	**6**	**0**
Aguas abajo	9	3	7	**0**	**0**	**6**	**3**

Contrastar la asociación entre la situación respecto a la planta de la zona del río y el libo de especies halladas en ella. Comentar la aplicabilidad del test de la ji-cuadrado a estos datos. (Basado en un estudio de Lawrence Scott Cook, Departamento de Biología, Universidad de Radford, 1994.)

HERRAMIENTAS COMPUTACIONALES

TI83

XXVIII. Contraste de la asociación entre dos variables

La calculadora TI83 puede contrastar la asociación entre dos variables. Para ello, los datos de la tabla de contingencia r x c s e introducen en una matriz; la calculadora construirá la tabla de las frecuencias esperadas de las celdas y realizará el contraste X^2 de asociación. Para ilustrarlo, usamos los datos del Ejemplo 12.2.1.

Tecla/Comando de la TI83	**Propósito**
1. MATRIX ▷ ▷ ENTER	1. Accede a la pantalla necesaria para introducir los datos de la tabla de contingencia.
2. 2 ENTER 4 ENTER	2. Indica que es una tabla de 2 x 4.
3. 698 ENTER 472 ENTER 102 ENTER 29 ENTER	3. Introduce los datos de la fila 1, de la Tabla 12.6.
4. 2892 ENTER 2625 ENTER 570 ENTER 226 ENTER	**4.** Introduce los datos de la fila *2,* de la Tabla 12.16.
5. STAT ◁ ALPHA C ▽ ▽ ENTER	5. Realiza y muestra el contraste X^2 de asociación.
6. MATRIX ▽ ENTER ENTER	6. Muestra la tabla de frecuencias esperadas de las celdas.

Paquete estadístico SAS

XVI. Contraste de la asociación entre dos variables

SAS puede contrastar la asociación entre dos variables utilizando la información de los datos sin procesar o las frecuencias observadas de la tabla de contingencia generada previamente. Mostramos esta última aplicación con los datos del Ejemplo 12.2.1.

Código SAS	**Propósito**
OPTIONS LS = 80 PS = 60 NO DATE;	Establece las especificaciones de impresión.
DATA ULCER;	Nombra el conjunto de datos.

```
INPUT ROW COLUMN
  OBSERVED;
LINES;
1   1    698
1   2    472
1   3    102
1   4     29
2   1   2892
2   2   2625
2   3    570
2   4    226

PROC FREQ;
TABLES ROW* COLUMN/
  NOCOL NOPERCENT
  EXPECTED CHISQ;
  WEIGHT = OBSERVED;
```

Nombra las variables.

Indica que los datos siguen a continuación.

Líneas de datos.

Señala el final de los datos.

Realiza el contraste X^2; muestra los porcentajes de cada fila y las frecuencias esperadas de las celdas.

El programa proporciona la salida siguiente:

The SλS System
TABLE OF ROW BY COLUMN

ROW / COLUMN Frequency Expected Row Pct	1	2	3	4	Total
1	698 613.42 53.65	472 529.18 36.28	102 114.82 7.84	29 43.572 2.23	1301
2	2892 2976.6 45.81	2625 2567.8 41.58	570 557.18 9.03	226 211.43 3.58	6313
Total	3590	3097	672	255	7614

STATISTICS FOR TABLE OF ROW BY COLUMN

Statatic	DF	Value	Prob
Chi-Square	3	29.122 ①	0.000 ②
Likelihood Ratio Chi-Square	3	29.559	0.000
Mantel-Haenszel Chi-Square	1	25.002	0.000
Phi Coefficient		0.062	
Contingency Coefficient		0.062	
Cramer's V		0.062	

Sample Size = 7614

Observe que las frecuencias esperadas de las celdas difiere ligeramente de la que se han dado en el texto debido a diferencias de redondeo. El valor observado del estadístico del contraste X^2 esta señalado con ①. El valor *P,* marcado con ②, es 0 a efectos prácticos.

Otros procedimientos y métodos alternativos de distribución libre

En este capítulo, presentamos algunos otros métodos que pueden resultar útiles en ocasiones, así como pruebas de contraste alternativas a las presentadas en capítulos anteriores.

Recordemos que en la mayor parte de los procedimientos estadísticos introducidos hasta ahora subyace la presunción de normalidad. Es decir, generalmente hemos supuesto que las muestras se extraen de poblaciones que, o bien están normalmente distribuidas, o bien están gobernadas por una distribución aproximadamente normal. Durante muchos años, después del descubrimiento de la curva normal, quienes hacían uso de la estadística creyeron que cualquier variable aleatoria seguía, prácticamente, una distribución normal o, al menos, una distribución que podía aproximarse bien por una distribución normal. A medida que se trataron más datos, se constató que esto no era cierto. No obstante, investigadores en campos de investigación muy diversos deseaban poder utilizar los importantes métodos estadísticos desarrollados por Fisher, Pearson, y «Student», que presuponen normalidad. Los profanos, que no comprendían la matemática subyacente en estas técnicas, pensaban que la «hipótesis» de normalidad no era importante, que era una ley de la naturaleza, o que se cumplía siempre por alguna sofisticada razón matemática. La mejor descripción de la situación se encuentra en las palabras de Lippman a Poincaré (1912) [2]:

> Todos creen en ella (la ley normal de los errores de medida); me dijo Lippman un día: «los que experimentan se figuran que es un teorema de matemáticas y los matemáticos que es un hecho experimental».

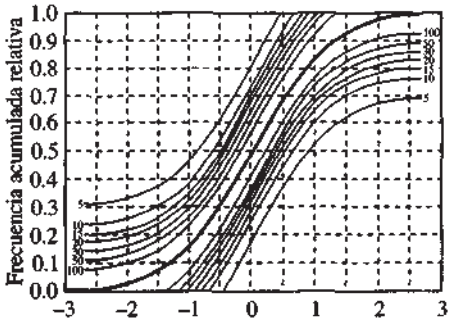
Hasta aquí, sólo hemos indicado cómo hacer una rudimentaria comprobación de la hipótesis de normalidad mediante el diagrama de tallos y hojas o el histograma. Si estos diagramas adoptan forma de campana, es razonable que la distribución sea normal. Ahora bien, si queremos salir de dudas debemos realizar una prueba de contraste para ver si hay evidencia estadística de que los datos proceden de una distribución que no es normal. Existen varios métodos para ello. Aquí consideraremos un método gráfico que es particularmente útil cuando el tamaño de la muestra es pequeño.

13.1. PRUEBAS DE NORMALIDAD: LA PRUEBA DE LILLIEFORS

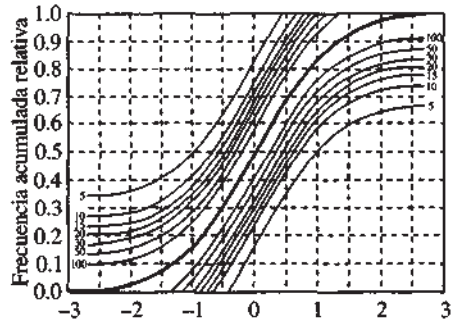
El método que consideramos aquí para detectar que no hay normalidad, llamado «prueba de normalidad de Lilliefors», fue desarrollado por H. W. Lilliefors a finales de los años 60. Aunque puede utilizarse para muestras grandes, es idóneo para muestras relativamente pequeñas. Básicamente, la prueba compara la frecuencia acumulada relativa observada de la muestra con la de la distribución normal típica. Esto se realiza dibujando el histograma de frecuencias relativas acumuladas observadas en un gráfico de Lilliefors. La Figura 13.1 proporciona los gráficos de Lilliefors necesarios para contrastar

H_0: los datos proceden de una distribución normal
H_1: los datos no proceden de una distribución normal

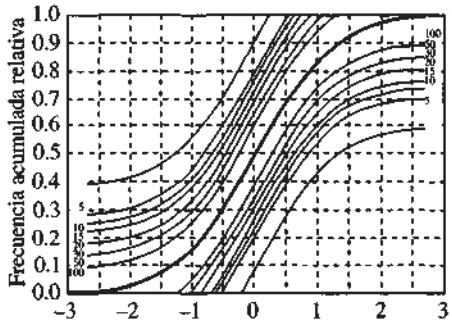
para varios niveles de significación. La curva de trazo grueso en el centro del gráfico es la gráfica de la función de la distribución normal típica. Las curvas situadas a ambos lados de



(a)



(b)



(c)

Figura 13.1. *a)* Cotas de Lilliefors al 90% para muestras normales ($\alpha = 0.1$); *b)* Cotas de Lilliefors al 95% para muestras normales ($\alpha = 0.05$); c) Cotas de Lilliefors al 99% para muestras normales ($\alpha = 0.01$). (Copyright 1982 by the American Statistical Association. Reproducido con autorización.)

ella representan las cotas de Lilliefors para los tamaños de muestras indicados. Si la frecuencia relativa acumulada está fuera de las cotas dadas para el tamaño de muestra especificado, entonces se rechaza H_0 y se concluye que los datos no proceden de una distribución normal. El Ejemplo 13.1.1 ilustra la utilización de esta técnica.

Ejemplo 13.1.1. Denotamos por X el crecimiento (en cm), en un período de 2 semanas, de 20 tejos cultivados en condiciones idénticas. Los datos obtenidos se muestran en la Tabla 13.1. Para estos datos, $x = 3.39$ y $s = 1.41$. Puesto que hemos de comparar la distribución de frecuencias relativas acumuladas con la de la distribución normal estándar, primero tipificamos estas observaciones restando x y dividiendo por s. Las observaciones tipificadas se ordenan entonces de menor a mayor y se halla la frecuencia relativa acumulada de cada observación. Los resultados aparecen en la Tabla 13.1. Para contrastar mediante estos datos

H_0: los datos proceden de una distribución normal

H_0: los datos no proceden de una distribución normal

al nivel $\alpha = 0.05$, dibujamos el histograma de frecuencias relativas acumuladas observadas sobre el gráfico de Lilliefors de la Figura 13.1b. El resultado se da en la Figura 13.2. Como el gráfico de frecuencias relativas acumuladas observadas no cae fuera de las bandas correspondientes al 20 (el tamaño de nuestra muestra) no podemos rechazar H_0. No tenemos evidencia de que los datos procedan de una distribución que no es normal.

Tabla 13.1. La columna 1 da el crecimiento, en cm, de 20 tejos, la columna 2 muestra los valores tipificados y la columna 3 es la distribución de frecuencias relativas acumuladas

Observación	Observación tipificada	Frecuencia relativa acumulada
1.1	-1.62	0.05
1.3	-1.48	0.10
1.4	-1.41	0.15
1.5	-1.34	0.20
1.9	-1.06	0.25
2.5	-0.63	0.30
2.6	-0.56	0.35
3.2	-0.13	0.40
3.5	0.08	0.45
3.7	0.22	0.50
3.7	0.22	0.50
3.9	0.36	0.60
4.1	0.50	0.65
4.2	0.57	0.70
4.2	0.57	0.70
4.4	0.72	0.80
4.6	0.86	0.85
4.8	1.00	0.90
4.9	1.07	0.95
6.2	1.99	1.00

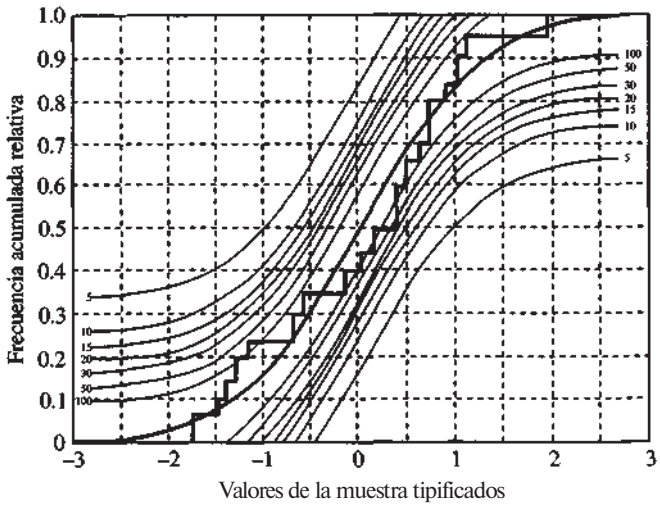


Figura 13.2. Prueba de normalidad del crecimiento de tejos, al nivel $\alpha = 0.05$.

Hay varias pruebas de tipo analítico para contrastar la hipótesis de normalidad. La mayoría de los paquetes estadísticos existentes en el mercado incluyen al menos una de ellas.

La pregunta natural a responder es: ¿qué hacer cuando no se da la presunción de normalidad? Se ha demostrado que el uso de la teoría normal conduce a comprobaciones aproximadas. En muchos casos, las aproximaciones son excelentes; en otros, resultan tan malas que se consideran inaceptables. De cualquier modo, utilizando la teoría normal en situaciones en las que no está justificada la presunción de normalidad, se llega a resultados poco fiables. Hay dos formas posibles de actuar. Primera: podemos intentar transformar los datos de tal manera que tenga consistencia la presunción de normalidad. En [5], [7] y [16] se describen métodos para hacerlo. Segunda: podemos desarrollar un cuerpo de métodos estadísticos que presupongan poco acerca de la distribución de la población muestreada. Tales métodos se llaman de *distribución libre.*

En las secciones siguientes describimos algunas de las técnicas de distribución libre más frecuentes. En particular, se incluyen procedimientos paralelos a los ya tratados. Así dispondremos de alternativas viables para muchos procedimientos de la teoría normal.

Los procesos estadísticos de distribución libre poseen algunas características atractivas. En particular, la deducción de la prueba estadística depende, en la mayoría de los casos, de métodos de cálculo como los descritos en los Capítulos 2 y 3. De este modo, es posible seguir de cerca la lógica de la prueba con gran facilidad. Para las pruebas de distribución libre, sólo se requieren, a menudo, pequeños cálculos que pueden realizarse muy rápidamente. Cuando los tamaños muestrales son pequeños ($n < 10$), es difícil detectar violaciones en los supuestos de la teoría normal, lo cual puede ser causa de importantes efectos negativos. Sin embargo, las pruebas de distribución libre tienen, para muestras pequeñas, una utilidad comparable a las de la teoría normal, incluso cuando se cumplen todos los supuestos requeridos por ésta. Si éste no es el caso, los procesos de distribución libre son, habitualmente, superiores. Por lo tanto, salvo que se den los supuestos clásicos, para muestras pequeñas, lo más aconsejable es elegir la prueba de distribución libre. Señalaremos sus ventajas más significativas. Varios de estos métodos se basan más en el análisis de los rangos que en las propias observaciones. Por ello, estas técnicas se utilizan más con datos de rangos que con observaciones o recuentos.

EJERCICIOS 13.1

Se está investigando un proceso nuevo de producción de pequeñas piezas de precisión. El proceso consiste en mezclar polvo fino de metal con una liga de plástico, inyectar la mezcla en un molde y finalmente eliminar el plástico con un disolvente. Los siguientes datos se han obtenido de piezas que deberían tener un diámetro de 1 pulgada y cuya desviación típica no debería exceder 0.0025 pulgadas.

1.0030	0.9997	0.9990	1.0054	0.9991
1.0041	0.9988	1.0026	1.0032	0.9943
1.0021	1.0028	1.0002	0.9984	0.9999

Para estos datos, $\bar{x} = 1.00084$ y $s = 0.00283$.

a) Utilizar el gráfico de Lilliefors de la Figura 13.3 para probar que estos datos no permiten rechazar la hipótesis de normalidad al nivel $\alpha = 0.05$.

b) Contrastar

$$H_0: \mu = 1$$
$$H_1: \mu \neq 1$$

al nivel $\alpha = 0.05$.

c) Contrastar

$$H_0: \sigma = 0.0025$$
$$H_1: \sigma > 0.0025$$

al nivel $\alpha = 0.05$.

2. Las piscinas cubiertas suelen tener condiciones acústicas deficientes. Se trata de diseñar una piscina donde un sonido de baja frecuencia tarde en apagarse un tiempo medio no mayor de 1.3 segundos, con una desviación típica de, a lo sumo, 0.6 segundos. Se realizan

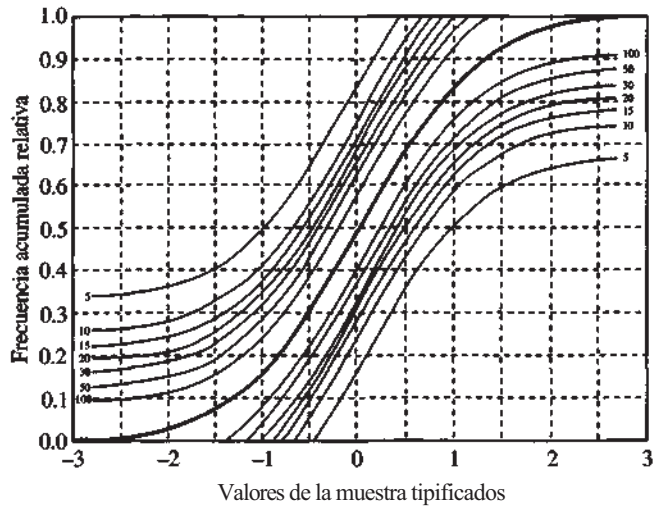


Figura 13.3. Cotas de Lilliefors al 95% para muestras normales.

simulaciones por ordenador de un diseño preliminar para ver si se logran estos niveles Se han obtenido los siguientes datos sobre el tiempo tardado por un sonido de baja frecuencia en apagarse. (R. Hughes y M. Johnson, «Diseño Acústico en Piscinas», *The Sound Engineering Magazine,* abril de 1983, págs. 34-36.)

1.8	3.7	5.0	5.3	6.1	0.5
2.8	5.6	5.9	2.7	3.8	5.9
4.6	0.3	2.5	1.3	4.4	4.6
5.3	4.3	3.9	2.1	2.3	7.1
6.6	7.9	3.6	2.7	3.3	3.3

Para estos datos, $\bar{x} = 3.97$ *y* $s = 1.89$.

a) Utilizar el gráfico de Lilliefors de la Figura 13.4 para probar que estos datos no permiten rechazar la hipótesis de normalidad al nivel $\alpha = 0.01$.

b) Contrastar

$$H_0: \mu = 1.3$$
$$H_1: \mu > 1.3$$

al nivel $\alpha = 0.01$.

c) Contrastar

$$H_0: \sigma = 0.6$$
$$H_1: \sigma > 0.6$$

al nivel $\alpha = 0.01$. ¿Puede decirse que se alcanzan los objetivos del proyecto?

3. Un problema habitual cuando se trabaja con ordenadores es la incompatibilidad. Se está ensayando un nuevo transformador de frecuencias de muestreo. Éste toma frecuencias de muestreo de 30 a 52 kHz, longitudes de palabra de 14 a 18 bits, y les da formato nuevo la frecuencia de muestreo de salida. Se piensa que el error de transformación tiene una desviación típica de menos de 150 picosegundos. Los siguientes datos se han obtenido

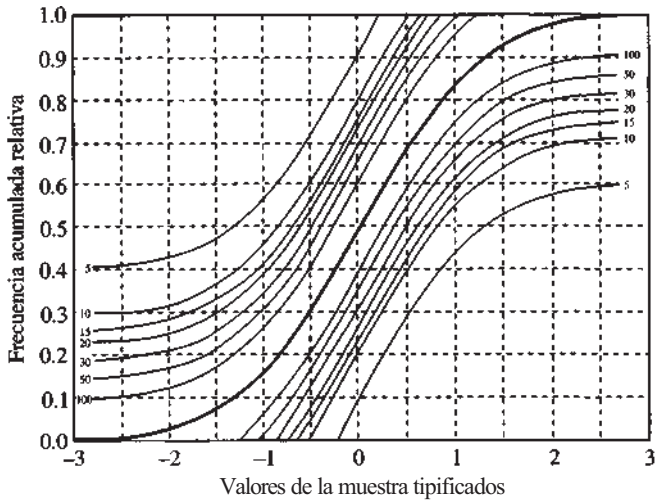


Figura 13.4. Cotas de Lilliefors al 99% para muestras normales.

del error de muestreo cometido en 20 pruebas del dispositivo. (K. Pohlmann, «The Compatibility Solution», *The Sound Engineering Magazine,* abril 1983, págs. 12-14.)

133.2	-11.5	-126.1	17.9	139.4
-81.7	314.8	147.1	-70.4	104.3
56.9	44.4	1.9	-4.7	96.1
-57.3	-43.8	-95.5	-1.2	9.9

Para estos datos, $x = 28.69$ y $s = 104.93$.

a) Utilizar el gráfico de Lilliefors de la Figura 13.5 para probar que estos datos no permiten rechazar la hipótesis de normalidad al nivel $\alpha = 0.1$.

b) Contrastar

$$H_0: \mu = 0$$
$$H_1: \mu \neq 0$$

al nivel $\alpha = 0.1$.

c) Contrastar

$$H_0: \sigma = 150$$
$$H_1: \sigma < 150$$

al nivel $\alpha = 0.1$. ¿Puede decirse que el transformador es tan preciso como se había afirmado?

13.2. CONTRASTES DE POSICIÓN: UNA MUESTRA

En los procedimientos basados en la teoría normal, la medida habitual del centro de posición de la distribución de una variable aleatoria es la media. En la mayoría de los contrastes de

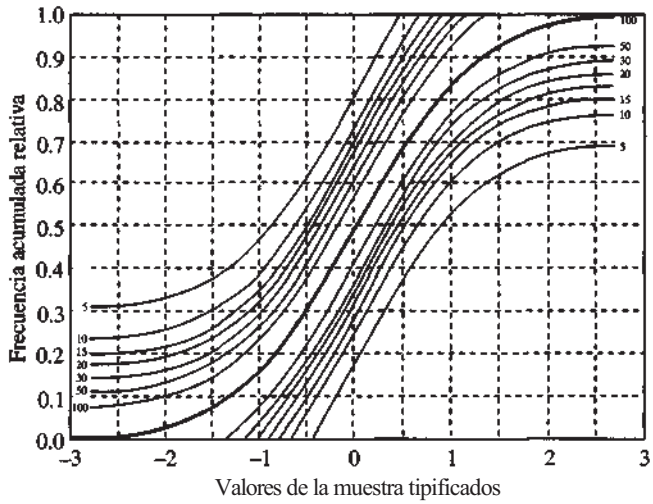


Figura 13.5. Cotas de Lilliefors al 90% para muestras normales.

distribución libre, el centro de posición lo determina la mediana de la variable aleatoria. La *mediana* de una variable aleatoria X se define como el número M tal que

$$P[X < M] \leq \tfrac{1}{2} \quad \text{y} \quad P[X \leq M] \geq \tfrac{1}{2}$$

Obsérvese que si X es continua, entonces

$$P[X < M] = P[X \leq M] = \tfrac{1}{2}$$

Es decir, para una variable aleatoria continua la mediana es el punto M tal que el 50 % de la veces X cae por debajo de M y el otro 50 % por encima.

Describiremos dos contrastes para la mediana de una variable aleatoria continua. El primero, llamado *contraste de los signos,* se basa en la distribución binomial. Sólo parte de un supuesto: que la variable aleatoria en estudio es continua.

Contraste de los signos para la mediana

Supongamos que X es una variable aleatoria continua con mediana M. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n de la distribución de X. Si M_0 designa el valor hipotético de la mediana, el contraste de hipótesis puede tener cualquiera de las tres formas habituales:

$H_0: M \leq M_0$	$H_0: M \geq M_0$	$H_0: M = M_0$
$H_1: M > M_0$	$H_1: M < M_0$	$H_1: M \neq M_0$
Cola a la derecha	Cola a la izquierda	Dos colas

Obsérvese que si H_0 es cierta ($M = M_0$), entonces cada una de las variables aleatorias continuas $X_1, X_2, \ldots, X_n$ tiene probabilidad $\frac{1}{2}$ de caer por debajo de M_0, probabilidad $\frac{1}{2}$ de caer por encima de M_0 y probabilidad 0 de tomar el valor M_0. Ello implica que cada una de las diferencias $X_1 - M_0, X_2 - M_0, \ldots, X_n - M_0$ tiene probabilidad $\frac{1}{2}$ de ser negativa, probabilidad $\frac{1}{2}$ de ser positiva y probabilidad 0 de tomar el valor 0. Sean N y N' el número de signos negativos y positivos obtenidos, respectivamente. Si H_0 es cierta, cada una de las variables aleatorias N y N' tiene una distribución binomial con parámetros n y $\frac{1}{2}$, y valor esperado $n\ \frac{1}{2}$. Es decir, si H_0 es cierta, esperamos que la mitad de los signos sean positivos y la otra mitad negativos.

Consideremos ahora el diagrama de la relación entre M y M_0 para un contraste con cola a la derecha, de la Figura 13.6. Claramente, si H_1 es cierta, debemos observar más signos positivos y menos negativos que los esperados. En este caso, usamos N, número de signos negativos observados, como estadístico. Rechazamos $H_0(M \leq M_0)$ en favor de $H_1(M > M_0)$ íi el valor observado de N es demasiado pequeño como para ser atribuido al azar, bajo el supuesto de que H_0 sea cierta. La situación es la inversa para un contraste con cola a la izquierda. Ahora, el estadístico es N', número de signos positivos observados. Una vez más se rechaza H_0 en favor de H, si el valor observado del estadístico es demasiado pequeño para que se deba al azar. El estadístico para un contraste de dos colas es el mínimo de los valores N y N', se rechaza H_0 para valores pequeños de este estadístico.

En el Ejemplo 13.2.1 utilizamos el contraste de los signos para la mediana.

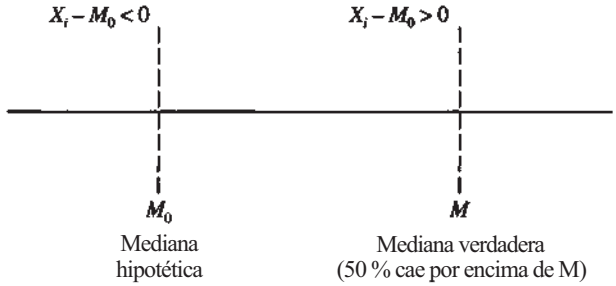


Figura 13.6. Relación entre M y M_0 (contraste con cola a la derecha: H_1: $M > M_0$).

Ejemplo 13.2.1. Un estudio sobre crecimiento basado en datos recogidos desde 1971 a 1974 indica que la estatura de los hombres entre los dieciocho y los veinticuatro años era de 177 cm. Se presume que debido a una mejor alimentación y atención sanitaria, la mediana de las estaturas de los jóvenes de este grupo de edades es generalmente superior a los 177 cm. Así pues, hemos de contrastar

$$H_0: M \leq 177 \qquad H_1: M > 177$$

Puesto que el contraste es con cola a la derecha, el estadístico es N, número de diferencias negativas que se obtiene al restar la mediana hipotética (177) de cada observación. Los resultados son los siguientes:

178(+)	167(-)	198(+)	189(+)
181(+)	171(-)	184(+)	171(-)
177(+)	181(+)	177(+)	179(+)
170(-)	183(+)	186(+)	180(+)
178(+)	186(+)	165(-)	193(+)

Se observan cinco signos negativos. El valor P del contraste viene dado por

$$P = P[N \leq 5 | p = \tfrac{1}{2}] = 0.0207$$

(Utilizar la tabla de la binomial, Tabla I del Apéndice B, con $n = 20$ y $p = \frac{1}{2}$.) Puesto que este valor es pequeño, rechazamos H_0, y concluimos que ha habido un aumento de la estatura media sobre la primitiva cifra de 177 centímetros.

Conviene hacer varias observaciones acerca del contraste de los signos para la mediana. El razonamiento subyacente al contraste es absolutamente lógico. Requiere muy poco cálculo y, por lo tanto, es rápido y fácil de aplicar. Presupone muy poco con respecto a la distribución de la población de la que se ha hecho el muestreo, imponiendo solamente que sea continua. Si, de hecho, la distribución es también normal, entonces media y mediana coincidirán. En este caso, el contraste de los signos está contrastando la misma hipótesis que el contraste T para una muestra que vimos con anterioridad (Capítulo 6).

A continuación nos referiremos al contraste de Wilcoxon de los rangos de signos para la mediana. Este contraste, desarrollado por F. Wilcoxon en 1945, contrasta la hipótesis nula de que una distribución continua es simétrica en torno a una mediana hipotética M_0.

Contraste de los rangos de signos de Wilcoxon

Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria de tamaño n de una distribución continua. Consideremos el conjunto de las diferencias $X_1 - M_0, X_2 - M_0, \ldots, X_n - M_0$, donde M_0 es la mediana hipotética de la distribución de la que se ha extraído la muestra. Si la hipótesis nula es cierta, entonces estas diferencias han sido tomadas de una distribución que es simétrica con respecto a

cero. Para aplicar el contraste, consideramos en primer lugar el conjunto de las n diferencias en valor absoluto, $|X_i - M_0|$, ordenadas de menor a mayor. Les asignamos rangos de 1 a n, de forma que el 1 corresponda a la más pequeña. A cada rango R_i se le asigna el signo algebraico de la diferencia correspondiente. Si la hipótesis nula es cierta y las diferencias son simétricas respecto a cero, entonces cada rango tiene exactamente la misma posibilidad de que se le asigne signo positivo o negativo.

Consideremos los estadísticos

$$W_+ = \sum_{\substack{\text{rangos} \\ \text{positivos}}} R_i \quad \text{y} \quad |W_-| = \left| \sum_{\substack{\text{rangos} \\ \text{negativos}}} R_i \right|$$

Si H_0 es cierta, deben ser de la misma magnitud. Si la auténtica mediana poblacional es en realidad mayor que la hipotética, entonces obtendremos más diferencias positivas, más rangos positivos, W_+ será mayor y $|W_-|$ menor de lo esperado. Si la verdadera media poblacional es más pequeña que la hipotética, estamos ante la situación inversa. Así que podremos utilizar como estadístico W, el menor de W_+ y $|W_-|$. Se ha tabulado la distribución de este estadístico para varios valores de α y n. Una de tales tablas es la Tabla XI del Apéndice B la tabla nos permite realizar contrastes unilaterales a los niveles α de 0.05, 0.025, 0.01 y 0.005. También pueden hacerse contrastes bilaterales al nivel 0.10, 0.05, 0.02 ó 0.01. La hipótesis nula se rechaza siempre para valores de W demasiado pequeños como para que se deban al azar. De modo que rechazamos H_0 al nivel α si el valor observado de W cae sobre o por debajo del punto crítico reseñado en la Tabla XI.

En el Ejemplo 13.2.2. se describe el modo de utilizar el contraste de los rangos de signos de Wilcoxon.

Ejemplo 13.2.2. En general, la mediana del tiempo de supervivencia de los pacientes de leucemia mieloblástica aguda con los que se logra una remisión completa, a partir de un tratamiento convencional, es de veintiún meses. Se está estudiando un nuevo procedimiento con el que se espera que aumente la mediana del tiempo de supervivencia. Se anota este tiempo para 10 pacientes a los que les ha sido aplicado:

24.1 25.8 20.5 20.9 27.3
21.5 20 1 28 9 19.2 26 3

Para contrastar por simetría respecto a 21, formamos primero el conjunto de las 10 diferencias que se obtienen restando 21 de cada una de las observaciones:

X_i	24.1	21.5	25.8	20.1	20.5	28.9	20.9	19.2	27.3	26.3
$X_i - 21$	3.1	0.5	4.8	-0.9	-0.5	7.9	-0.1	-1.8	6.3	5.3

Ordenamos, a continuación, de menor a mayor, los valores absolutos de las diferencias, asignándoles un rango de 1 a 10 y a cada uno el signo de la diferencia que le corresponda. Dado el supuesto de continuidad, no deben darse coincidencias. Sin embargo, en la práctica se dan. Por ejemplo, la diferencia 0.5 en valor absoluto aparece dos veces. Cuando esto ocurre, se asigna a todas las coincidencias contabilizadas la media de los rangos que les correspondería. La idea se materializa del modo siguiente:

$\lvert X_i - 21 \rvert$	0.1	0.5	0.5	0.9	1.8	3.1	4.8	5.3	6.3	7.9
Rango R_i	1	2.5	2.5	4	5	6	7	8	9	10
Rango con signo	-1	-2.5	2.5	-4	-5	6	7	8	9	10

Los valores observados de W_+ y $|W_-|$ son

$$W_+ = \sum_{\substack{\text{rangos} \\ \text{positivos}}} R_i = 2.5 + 6 + 7 + 8 + 9 + 10 = 42.5$$

y

$$|W_-| = \left| \sum_{\substack{\text{rangos} \\ \text{negativos}}} R_i \right| = |-1 + (-2.5) + (-4) + (-5)| = 12.5$$

Nuestro estadístico es W, el menor de los dos. Como se está realizando un contraste unilateral, entonces en la Tabla XI vemos que para una muestra de tamaño 10, $P[W \leq 11] = 0.05$. Puesto que el valor observado del estadístico es 12.5, $P = P[W \leq 12.5]$ es mayor que 0.05. De modo que no tenemos prueba suficiente, basándonos en esta muestra, de que el nuevo procedimiento haya aumentado la mediana del tiempo de supervivencia.

Hay varios detalles a destacar en relación con el contraste de los rangos de signos de Wilcoxon. Al igual que el contraste de los signos, supone que la distribución de la que se ha extraído la muestra es continua. Si la distribución es también simétrica, entonces el contraste lo es al mismo tiempo para la media y la mediana de la distribución, ya que en este caso los dos parámetros coinciden. En ocasiones, a partir de una experiencia anterior se puede saber si una población es simétrica, pero la mayor parte de las veces no es así. Cuando estamos en este último caso, es preferible el contraste de los signos como contraste de posición. Si hay pruebas de que la distribución es normal, entonces el contraste de los rangos de signos de Wilcoxon está contrastando la misma hipótesis que la contrastada por T para una muestra en la teoría normal.

EJERCICIOS 13.2

1. Recientes investigaciones sobre el ejercicio de la medicina en centros privados en los que no se atienden pacientes del seguro Medicaid, indican que la mediana de la duración de la visita por paciente es de veintidós minutos. Se cree que en centros con un elevado número de pacientes de Medicaid esta cifra es menor. Se obtuvieron los siguientes datos sobre las visitas de 20 pacientes aleatoriamente seleccionados:

Duración de la visita del paciente, en minutos			
21.6	13.4	20.4	16.4
23.5	26.8	24.8	19.3
23.4	9.4	16.8	21.9
24.9	15.6	20.1	16.2
18.7	18.1	19.1	18.9

A partir del contraste de los signos, ¿hay pruebas suficientes para concluir que, en los centros en los que se atiende a pacientes de Medicaid, la mediana de la duración de una visita por paciente es menor de veintidós minutos? Explicar la respuesta basándose en el número esperado de signos positivos bajo H_0 y el valor P del contraste.

2. Se realiza un estudio sobre nutrición en pacientes con insuficiencia respiratoria que requieren ventilación asistida. Una variable considerada es el índice de creatinina, que

es una medida del nivel proteínico del paciente. El hecho de que el índice tenga un valor inferior a 6 es indicativo de un grave déficit proteínico. Si la mediana de los valores de índice en este tipo de pacientes está por debajo de 6 se pondrá en funcionamiento un nuevo programa dietético para corregir el problema. A partir de una muestra aleatoria de 15 pacientes se obtuvieron los siguientes valores:

5.7	4.2	4.7	4.6
5.3	5.4	6.8	4.9
4.9	5.8	4.1	5.5
6.4	5.1	4.7	

Por medio del contraste de los signos, ¿hay pruebas de que la mediana de los índices esté por debajo de 6? Explicar la respuesta basándose en el número esperado de signos positivos bajo H_0 y el valor *P* del contraste.

3. Dos terrenos de pasto bien desarrollados contienen micorrizas que estimulan el creci miento de tréboles y césped. La mediana del número de esporas por gramo de tierra en buenos terrenos de pastos es 9. Se piensa que en áreas erosionadas la presencia de micorri zas está altamente reducida ¿Tienden a apoyar el argumento los datos siguientes, extraídos de 20 áreas erosionadas? Explicarlo basándose en el valor *P* del contraste de los signos

0.0l	0.12	0.28	0.54	2.7
0.02	0.15	0.30	0.92	2.7
0.06	0.16	0.32	1.52	8.24
0 08	0.24	0.48	1.64	9.3

4. Se lleva a cabo un estudio sobre llamadas de alarma entre ardillas terreras. Una variable considerada es la distancia máxima a la que es audible una llamada de alerta. Se cree que la mediana de las distancias máximas a las que la llamada es audible es de más de 87 metros. ¿Apoyan el argumento los datos siguientes? Explicar la respuesta basándose en el número esperado de signos negativos bajo H_0 y el valor *P* del contraste.

Distancia máxima audible, en metros			
90.8	79.4	94.4	96.7
91.9	94.3	95.1	84.5
85.2	89.7	82.0	88.2
88.6	95.6	89.4	87.3
98.5	87.1	82.1	86.7

5. Observemos que $W_+ + |W_-| = 1 + 2 + 3 + ... + n$. Del álgebra elemental sabemos que esta suma es $n(n + 1)/2$. Este resultado puede utilizarse para comprobar la precisión de la clasificación y el cálculo de los valores de $W_+ + |W_-|$ cuando se utiliza el contraste de Wilcoxon. Comprobar este resultado para los datos del Ejemplo 13.2.2, en que $n = 10$.
6. Si la hipótesis nula es que la distribución de la que procede la muestra es simétrica respecto de la mediana hipotética M_0, entonces el valor esperado de W_+ viene dado por

$$E[W_+] = \frac{n(n+1)}{4}$$

Hallar el valor esperado de W_+ para el Ejemplo 13.2.2 ($n = 10$).

7. Una cierta clase de escarabajos se reúne durante el día en grupos de centenares de individuos llamados *agregaciones,* que están compuestos por una única especie o por varias especies. Se cree que la mediana de las distancias entre agregaciones es menor de 0.8 km. Utilizar el contraste de los signos de Wilcoxon para comprobar este valor hipotético basándose en los siguientes datos. Supóngase simetría.

0.71	0.65	0.51	0.32	0.21
0.13	0.21	1.10	0.71	1.63
0.16	1.00	1.11	0.40	

8. Antiguas investigaciones sobre la especie de arañas denominada «viuda negra» indican que la mediana del tiempo que la araña permanece en tierra cuando desciende para enganchar un hilo, durante el tejido de la tela, es de menos de dieciséis segundos. Para comprobar esto, se las observa cuando están activamente ocupadas en tejer la tela. Se han obtenido los datos siguientes (supóngase simetría).

Tiempo en tierra, segundos				
18.9	9.3	10.8	20.0	11.9
11.7	19.8	11.1	15.5	9.8
15.9	13.1	23.7	10.4	11.9
25.2	13.5	8.9	12.5	19.9
21.3	8.5	17.5	18.9	13.4

Si la verdadera mediana es de dieciséis segundos, ¿cuál es el valor esperado de W_+? (Véase el Ejercicio 6.) ¿Cuál es el valor observado de W_+? ¿Tienden los datos a apoyar el argumento de que la mediana del tiempo empleado en tierra es menor de dieciséis segundos? Explicar la respuesta basándose en el valor P del contraste de Wilcoxon.

9. Se cree que la mediana de las edades en que comienza la diabetes es cuarenta y cinco años. Supóngase que sea simétrica la distribución de la variable edad de comienzo de la diabetes. Queremos contrastar

$$H_0: M = 45 \qquad \mathbf{H_1: M \neq 45}$$

al nivel $\alpha = 0.05$, utilizando el contraste de los rangos de signos de Wilcoxon basándose en una muestra de tamaño 30. De un estudio de 30 diabéticos aleatoriamente seleccionados se obtuvieron los siguientes datos:

35.5	30.5	40.1	59.8	47.3
44.5	48.9	36.8	52.4	36.6
39.8	42.1	39.3	26.2	55.6
33.3	40.3	65.4	60.9	45.1
51,4	46.8	42.6	45.6	52.2
51.3	38.0	42.8	27.1	43.5

¿Puede rechazarse H_0? ¿A qué tipo de error nos arriesgamos?

10. En 1970 se informó de que la mediana de los costes medios de una primera visita a la consulta de un médico era de 14.23 dólares. Si bien el coste de tal visita es ciertamente más alto en dólares actuales, teniendo en cuenta la inflación, puede realmente no haberse producido una subida en los precios con respecto al total de la economía. Los si-

guientes datos representan el coste actual de una primera visita a la consulta, referido a dólares de 1970. Supuesta la simetría, ¿indican los datos que la mediana de los costes es ahora mayor que la mediana de 1970? Explicar la respuesta basándose en el valor *P* del contraste de Wilcoxon.

Coste **actual, dólares de 1970**				
16.14	15.71	16.23	17.44	16.88
15.79	15.10	15.82	15.89	16.99
14.08	16.30	14.88	14.02	14.22
17.22	14.39	16.04	14.56	15.32
16.18	15.26	16.73	16.03	13.94

11. *Aproximación normal* a W_+. Para valores grandes de *n*, W_+ y $|W_-|$ tienen aproximadamente distribución normal con media $n(n + 1)/4$ y varianza

$$\frac{n(n + 1)(2n + 1)}{24}$$

Ello puede emplearse para aproximar los valores de *P* para valores de *n* que no aparecen en la Tabla XI del Apéndice B. Necesitamos solamente tipificar W_+ o $|W_-|$ y aproximar *P* utilizando la tabla normal tipificada.

a) Para una muestra de tamaño 70, hallar $P[W_+ \leq 1000]$.

b) Para una muestra de tamaño 80, hallar $P[|W_-| \leq 1500]$.

13.3. CONTRASTES DE POSICIÓN: DATOS EMPAREJADOS

Consideremos ahora dos contrastes de posición que utilizan observaciones emparejadas. La primera, la prueba de los signos para la mediana de las diferencias, es una extensión de la prueba de los signos anteriormente presentada.

Contraste de los signos para la mediana de las diferencias

Sean X e Y variables aleatorias continuas. Sea $(X_1, F,)$, (X_2, F_2), ..., (X_n, Y_n) una muestra aleatoria de tamaño *n* de la distribución de (X, *Y).* Consideremos el conjunto de las *n* diferencias continuas $X_1 - Y_1$, $X_2 - F_2$, ..., $X_n - Y_n$. La hipótesis nula es que la mediana de las diferencias es cero. Es decir, estamos contrastando

$$H_0: M_{X-Y} = 0$$

Si H_0 es cierta, cada una de las diferencias $X_i - Y_i$ tiene probabilidad $\frac{1}{2}$ de ser positiva, probabilidad $\frac{1}{2}$ de ser negativa y probabilidad 0 de tomar el valor 0. Sean *N* y *N'* el número de signos positivos y negativos obtenidos, respectivamente. Si H_0 es cierta, cada una de estas variables aleatorias está distribuida binomialmente con parámetros *n* y $\frac{1}{2}$. De este modo, si H_0 es cierta, esperamos que la mitad de las diferencias observadas sean positivas y la mitad negativas. Si la mediana de las diferencias es realmente positiva, entonces obtendremos muchos más sig-

nos positivos y muchos menos negativos. Si la mediana de las diferencias es negativa la situación se invierte. Podemos pues utilizar como estadístico el menor de entre N y N'. Rechazamos H_0 en favor de la alternativa apropiada si el valor observado del estadístico es demasiado pequeño para que pueda deberse al azar bajo el supuesto de que H_0 sea cierta.

El desarrollo del contraste se describe en el Ejemplo 13.3.1.

Ejemplo 13.3.1. En un estudio sobre empleo del captopril en pacientes hipertensos tratados con diuréticos, se utiliza una dosis de 6.25 mg. Se anota la tensión arterial sistólica de cada paciente antes de que reciba el fármaco *(X)* y setenta minutos después de que le haya sido administrado *(Y)*. Para determinar si el fármaco es eficaz en la disminución de la tensión arterial contrastamos

$$H_0: M_{X-Y} \leq 0 \qquad H_1: M_{X-Y} > 0$$

Si H_1 es cierta, esperaremos muy pocas diferencias negativas. El estadístico es, por tanto, N, el número de signos negativos obtenido. Se tienen los siguientes datos:

X (antes)	Y (después)	Signo de $X - Y$
175	140	+
179	143	+
165	135	+
170	133	+
160	162	−
180	150	+
177	182	−
178	139	+
173	140	+
176	141	+

El valor observado de N es 2. El valor P para el contraste viene dado por

$$P = P[N \leq 2 | p = \tfrac{1}{2}] = 0.0547$$

(Utilizar la Tabla I del Apéndice B con n = 10 y p = $\frac{1}{2}$.) Puesto que esta probabilidad es claramente pequeña, rechazamos H_0 y concluimos que el captopril es eficaz en la reducción de tensión arterial sistólica, en los pacientes hipertensos tratados con diuréticos.

Observemos que el contraste de los signos para la mediana de las diferencias sólo tiene un supuesto: que ambas, *XeY,* sean continuas. Si cada una de ellas es también normal, entonces $X - Y$ es normal. En este caso, la mediana de las diferencias es igual a la media de las diferencias, que, a su vez, es igual a la diferencia de las medias poblacionales. Es decir:

$$M_{X-Y} = \mu_{X-Y} = \mu_X - \mu_Y$$

De este modo, si X e Y son normales, contrastar que la hipótesis nula H_0: $M_{X\_Y} = 0$ es equivalente a contrastar la hipótesis nula de igualdad de medias poblacionales. Se trata de la misma hipótesis contrastada mediante el contraste T para datos emparejados en la teoría estadística normal.

El siguiente contraste que consideramos para su empleo con datos emparejados es el contraste de los rangos de signos de Wilcoxon para observaciones emparejadas. Se realiza exactamente de la misma forma que el contraste de los rangos de signos para la mediana, descrito en la Sección 13.2.

Contraste de los rangos de signos de Wilcoxon: datos emparejados

Sean X e Y variables aleatorias continuas. Sea $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ una muestra aleatoria de tamaño n de la distribución de (X, Y). Consideremos el conjunto de las n diferencias continuas $X_1 - Y_1, X_2 - Y_2, \ldots, X_n - Y_n$. La hipótesis nula es que estas diferencias han sido extraídas de una población que es simétrica con respecto a cero. El contraste se realiza ordenando los valores absolutos de estas diferencias de menor a mayor y asignándoles rangos de 1 a n (a las coincidencias contabilizadas se les asigna la media de los rangos que les corresponderían). A cada rango se le asocia el signo de la diferencia correspondiente y se calcula el valor de

$$W_+ = \sum_{\substack{\text{rangos} \\ \text{positivos}}} R_i \quad \text{y} \quad |W_-| = \left| \sum_{\substack{\text{rangos} \\ \text{negativos}}} R_i \right|$$

El estadístico es W, el menor de W_+ y $|W_-|$. Se rechaza la hipótesis H_0 en favor de la alternativa apropiada para valores de W que sean demasiado pequeños para que se deban al azar, basándose en la Tabla XI del Apéndice B.

El Ejemplo 13.3.2 describe la utilización del procedimiento de Wilcoxon con datos emparejados.

Ejemplo 13.3.2. La incorporación accidental de compuestos de bifenilo polibrominadou (PBB) en las pildoras de alto contenido proteínico en una vaquería, llevaron a la contaminación del ganado vacuno en algunos estados del medio oeste. Se lleva a cabo un estudio para determinar si la cocción reduce el nivel de **PBB** de la carne de los animales contaminados. Se realiza el experimento midiendo el nivel de **PBB** en un trozo de solomillo crudo (X), cociéndolo y midiendo después el nivel de **PBB** (Y). Si la cocción reduce el nivel de PBB esperaremos muy pocas diferencias negativas cuando la sustracción se hace en el orden $X - Y$. Nuestro estadístico es, por tanto, $|W_-|$. Concluiremos que el cocimiento efectivamente reduce el nivel de PBB si el valor observado del estadístico es demasiado *pequeño* para que pueda deberse al azar. Se obtuvieron los siguientes datos (en ppm) (supóngase simetría):

X **(crudo)**	Y **(cocido)**	$X - Y$ **(diferencia)**
0.19	0.15	0.04
0.20	0.10	0.10
0.01	0.02	-0.01
0.16	0.18	-0.02
0.15	0.10	0.05
0.27	0.04	0.23
0.08	0.01	0.07
0.23	0.15	0.08
0.07	0.04	0.03
0.10	0.10	0.00

Ordenemos ahora los valores absolutos de las diferencias de menor a mayor y asignémosles rangos de 1 a 10. Al asociar los signos a los rangos surge un problema. ¿Qué signo algebraico se asignaría a la diferencia 0, que no es ni positiva ni negativa? Teóricamente, a causa de la continuidad de X e Y, no deberían presentarse diferencias nulas. Sin embargo, en la práctica, como resultado de dificultades en las medidas pueden surgir ocasionalmente diferencias cero. Hay varias posibilidades para abordar el problema. Tomamos una aproximación conservadora: puesto que la hipótesis nula es que la población de diferencias es simétrica con respecto a 0, una diferencia nula tiende a apoyar H_0. Así, lógicamente asignaremos a la diferencia 0 el signo que menos contribuya a que H_0 sea rechazada. En este caso, le asignamos un

signo negativo, ya que esto aumentará el tamaño de $|W_-|$ y hará más difícil rechazar H_0. LOS resultados obtenidos son los siguientes:

$\|X_i - Y_i\|$	0	0.01	0.02	0.03	0.04	0.05	0.07	0.08	0.10	0.23
Rango R_i	1	2	3	4	5	6	7	8	9	10
Rango con signo	-1	-2	-3	4	5	6	7	8	9	10

El valor observado de |W_| es:

$$|W_-| = \left| \sum_{\substack{\text{rangos} \\ \text{negativos}}} R_i \right| = |-1 + (-2) + (-3)| = 6$$

De la Tabla XI del Apéndice B se obtiene que el valor P $(P[|W_-| \leq 6])$ está entre 0.01 (punto crítico 5) y 0.025 (punto crítico 8). Puesto que este valor es pequeño, rechazamos H_0 y concluimos que la cocción tiende a reducir el nivel de PBB en la carne contaminada.

De nuevo, es importante observar la relación entre este contraste y otros anteriores. Si se sabe que la población de diferencias es simétrica, entonces se está contrastando la hipótesis nula de que la mediana de las diferencias es cero. Ésta es la misma hipótesis que la contrastada mediante el contraste de los signos. Si no se sabe si la población es simétrica y se quiere contrastar una hipótesis acerca de la mediana, entonces es preferible el contraste de los signos al de Wilcoxon. Si hay pruebas de que la población de diferencias está normalmente distribuida, entonces se está contrastando H_0: $\mu_x = \mu_y$, la misma hipótesis que la contrastada mediante el contraste T para datos emparejados de la teoría normal.

EJERCICIOS 13.3

1. Se desarrolla un programa, de veinte semanas de duración, de entrenamiento físico para mujeres. Una variable estudiada es la capacidad máxima de oxígeno admitida por el sujeto. Se mide mientras está utilizando una banda de marcha, tanto antes *(X)* como después *(Y)* del período de ejercicio. Se espera que el ejercicio aumentará el valor de esta variable para la mayor parte de los individuos.
   a) La hipótesis de investigación es H_1: $M_{X-Y} < 0$. ¿Cuál es el estadístico para detectar esta situación?
   b) Se obtuvieron los siguientes datos:

Capacidad máxima de oxígeno, litros/minuto		
Sujeto	Antes *(X)*	Después *(Y)*
1	1.98	2.26
2	1.57	1.83
3	1.89	2.31
4	1.42	1.79
5	1.73	1.65
6	1.95	2.26
7	1.69	2.10
8	1.92	2.15
9	1.96	1.54
10	1.94	1.87

Utilizar el contraste de los signos para contrastar.

$$H_0: M_{X-Y} \geq 0 \quad H_1: M_{X-Y} < 0$$

¿Se puede concluir que el ejercicio tiende a aumentar la capacidad máxima de oxígeno? Explicarlo basándose en el valor *P* del contraste.

2. Se realiza un estudio sobre el tiempo que tardan en reaccionar jugadores de hockey indios. El propósito es comparar el tiempo de reacción visual con el tiempo de reacción auditiva. El tiempo de reacción visual se mide anotando el tiempo que se necesita para responder a una señal luminosa; el tiempo de reacción auditiva es el tiempo que se necesita para responder al chasquido de un interruptor eléctrico. Se anotaron los siguientes tiempos para 15 individuos:

Tiempo de reacción, milisegundos			Tiempo de reacción, milisegundos		
Sujeto	Visual	Auditiva	Sujeto	Visual	Auditiva
1	165.75	162.32	9	195.76	207.34
2	207.57	211.84	10	182.82	198.44
3	240.21	202.65	11	164.37	177.82
4	180.50	166.14	12	232.54	142.28
5	205.89	239.14	13	197.55	187.09
6	192.96	201.51	14	196.58	164.42
7	233.16	184.88	15	216.09	161.39
8	215.86	170.48			

¿Hay pruebas basadas en el contraste de los signos de que la reacción visual tiende a ser más lenta que la reacción auditiva? Explicarlo sobre la base del valor *P* del contraste

3. Se realiza un estudio para determinar los efectos de eliminar una obstrucción renal en pacientes cuya función renal está deteriorada, a causa de una metástasis maligna avanza da de origen no urológico. Se mide la tensión arterial de cada paciente antes *(X)* y de pues de la operación *(Y),* y se obtienen los siguientes resultados:

Presión arterial, mm Hg		
Paciente	Antes *(X)*	Después *(Y)*
1	150	90
2	132	102
3	130	80
4	116	82
5	107	90
6	100	94
7	101	84
8	96	93
9	90	89
10	78	85

Basándonos en el contraste de los signos, ¿se puede concluir que la intervención quirúrgica tiende a disminuir la tensión arterial? Explicarlo basándose en el valor *P* del contraste.

4. Se realiza un estudio sobre los efectos del ejercicio físico en pacientes con enfermedad coronaria, midiendo el máximo de oxígeno consumido por cada paciente, antes de comenzar con el entrenamiento *(X)*. Después de seis meses de hacer ejercicio con bicicleta tres veces por semana, se midió nuevamente el oxígeno consumido por cada persona *(Y)*. Se obtuvieron los siguientes datos (supóngase la simetría):

Máximo de oxígeno admitido, mL/(kg)(min)		
Paciente	Antes *(X)*	Después *(Y)*
1	46.98	40.96
2	23.98	26.21
3	48.25	57.25
4	41.24	38.83
5	42.90	52.17
6	42.45	54.02
7	23.00	24.58
8	30.39	51.51
9	33.80	31.62
10	47.41	54.83

Basándose en el contraste de los rangos de signos de Wilcoxon, ¿se puede concluir que, a nivel $\alpha = 0.05$, el ejercicio tiende a aumentar el máximo de oxígeno admitido por los pacientes?

5. Se lleva a cabo un estudio sobre el comportamiento de cortejo de pinzones cebra domesticados, para determinar el efecto del color del pico de la hembra sobre el número de tipos de canto utilizados por el macho durante el cortejo. Se cree que el pico rojo, un signo de madurez, irá unido a más tipos de canto que el pico negro, que se da en los pájaros jóvenes. Diez machos maduros se presentan por separado a una hembra en celo de pico rojo y a una hembra en celo de pico negro. Se observa en cada caso el comportamiento de los machos. La variable medida es el número medio de tipos de canto, sobre tres períodos de observación de diez minutos, con cada hembra:

Número medio de tipos de cantos		
Número del pájaro	Pico rojo *(X)*	Pico nego *(Y)*
1	11.24	2.19
2	12.21	1.69
3	11.7	9.84
4	14.09	13.98
5	15.7	12.66
6	16.9	7.9
7	17.08	14.78
8	11.17	15.66
9	15.18	11.06
10	14.72	20.08

a) Supóngase la simetría y utilícese el contraste de los rangos de signos de Wilcoxon para contrastar

$$H_0: M_{x\_y} < 0 \quad H_0: M_{x\_y} > 0$$

al nivel $\alpha = 0.05$.

b) Si no suponemos la simetría, entonces la hipótesis

$$H_0: M_{X-Y} \leq 0 \quad H_1: M_{X-Y} > 0$$

se contrasta mejor con el contraste de los signos. Si se utiliza éste, ¿puede rechazarse H_0 al nivel $\alpha = 0.05$?

6. En un estudio sobre la hipertensión moderada en pacientes con edades entre los veintiún y los veintinueve años, se utilizan dos grupos. Un grupo experimental recibe clortalidona más reserpina; un segundo grupo recibe un placebo. Se determina el nivel total de colesterol de cada paciente al comienzo del estudio y, de nuevo, tras un año de tratamiento. Se obtienen los siguientes datos:

Nivel total de colesterol (mg/dL)

Clortalidona y reserpina		Placebo	
Antes del tratamiento *(X)*	Después del tratamiento (Y)	Antes del tratamiento *(X)*	Después del tratamiento *(Y)*
192.3	172.1	180.5	182.3
178.6	164.1	170.1	170.9
185.7	171.4	174.1	170.2
175.3	152.9	180.4	178.3
183.9	163.9	175.4	175.8
182.6	170.7	188.4	186.1
180.9	165.3	182.8	185.1
184.2	164.2	181.0	177.6

Supuesta la simetría, utilizar el contraste de los rangos de signos de Wilcoxon para contrastar

$$H_0: M_{X-Y} \leq 0 \quad H_1: M_{X-Y} > 0$$

al nivel $\alpha = 0.05$ para cada grupo.

13.4. CONTRASTES DE POSICIÓN: DATOS NO ASOCIADOS

En esta sección analizamos un contraste de distribución libre que puede utilizarse para comparar la posición de dos poblaciones continuas, basado en muestras independientes de tamaños *m y n* extraídas de aquellas poblaciones. Se llama *contraste de la suma de los rangos de Wilcoxon.*

Contraste de la suma de los rangos de Wilcoxon

Sean X e Y variables aleatorias continuas. Sean $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$ muestras aleatorias independientes, de tamaños m y n, de las distribuciones de X e Y, respectivamente. Su-

pongamos que $m \leq n$. Esto es, supongamos que las X representan la muestra más pequeña. La hipótesis nula es que las poblaciones X e Y son idénticas. Queremos contrastar estas hipótesis con un contraste que es especialmente idóneo para rechazar H_0 si las poblaciones difieren en posición. Las $m + n$ observaciones se funden para formar una única muestra. Las observaciones se ordenan linealmente y se les atribuye un rango de 1 a $m + n$, conservando su identidad de grupo. Se asigna a las coincidencias que aparezcan la media de los rangos que les corresponderían, como en los contrastes de Wilcoxon.

El estadístico es W_m, la suma de los rangos asociados con las observaciones que originalmente constituyeron la muestra menor (valores X). La lógica que está detrás de esta elección del estadístico es la siguiente: si la población X está situada por debajo de la población Y, entonces los rangos menores tenderán a asociarse con los valores X. Ello producirá un valor pequeño para W_m. Si es cierto lo contrario (la población X está situada por encima de la población Y), entonces los rangos mayores se encontrarán entre las X, dando lugar a un gran valor de W_m. De este modo, rechazaremos H_0 si el valor observado de W_m fuera demasiado pequeño o demasiado grande para que se debiera al azar. La Tabla XII del Apéndice B da las probabilidades para valores seleccionados de m y n. Indicamos la forma de utilizar esta tabla en el Ejemplo 13.4.1.

Ejemplo 13.4.1. En un estudio sobre el hábito de fumar y sus efectos sobre las pautas del sueño, una de las variables importantes es el tiempo que se tarda en quedarse dormido. Se extrae una muestra, de tamaño 12, de la población de fumadores y otra independiente, de tamaño 15, de la población de no fumadores. Se obtienen los siguientes datos:

Tiempo que tardan en dormirse, minutos

Fumadores *(S)*		**No fumadores** *(N)*	
69.3	52.7	28.6	30.6
56.0	34.4	25.1	31.8
22.1	60.2	26.4	41.6
47.6	43.8	34.9	21.1
53.2		29.8	36.0
48.1		28.4	37.9
23.2		38.5	13.9
13.8		30.2	

¿Indican estos datos que los fumadores tienden a tardar más tiempo en quedarse dormidos que los no fumadores?

Para responder a esta pregunta, fundimos las dos muestras, ordenamos las observaciones de menor a mayor conservando su identidad de grupo y les atribuimos un rango de 1 a 27:

Observación	13 8	13.9	21. 1	22. 1	23.2	25. 1	26.4	28.4	28.6
Grupo	***S***	***N***	***N***	***S***	***S***	***N***	***N***	***N***	***N***
Rango	1	***2***	3	4	5	6	***1***	8	***9***

Observación	29.8	30.2	30.6	31.8	34.4	34.9	36.0	37.9	38.5
Grupo	***N***	***N***	***N***	***N***	***S***	***N***	***N***	***N***	***N***
Rango	10	11	12	13	14	15	16	17	18

Observación	41.6	43.8	47.6	48.1	52.7	53.2	56.0	60.2	69.3
Grupo	*N*	*S*	*S*	*S*	*S*	5	*S*	*S*	*S*
Rango	19	20	21	22	23	24	25	26	27

Puesto que la muestra de la población de fumadores *(m* = 12) es de menor tamaño que la de no fumadores (n = 15), el estadístico del contraste W_m es la suma de los rangos asociados a los fumadores. Como sospechamos que los fumadores tardan más tiempo en quedarse dormidos que los no fumadores, rechazamos la hipótesis nula de que no existe diferencia entre los dos grupos si el valor observado de W_m es demasiado grande para que se deba al azar. Para estos datos

$$W_m = 1 + 4 + 5 + 14 + 20 + 21 + 22 + 23 + 24 + 25 + 26 + 27 = 212$$

Volvamos ahora a la Tabla XII del Apéndice B, entrando con $m = 12$ y $n = m + 3 = 15$. El punto crítico para un contraste con cola a la derecha, y $\alpha = 0.05$, es 202. Como $212 > 202$, rechazamos H_0 y concluimos que los fumadores tienden a tardar más tiempo en quedarse dormidos que los no fumadores.

Si ambas poblaciones X e Y se suponen normales, entonces el contraste de la suma de rangos de Wilcoxon contrasta la misma hipótesis que el contraste T para datos emparejados en la teoría normal.

Muchos otros contrastes para distribuciones libres son equivalentes al contraste de la suma de rangos de Wilcoxon. La alternativa más conocida es el contraste de Mann-Whitney. Depende también de las observaciones X e Y linealmente ordenadas. El estadístico en este caso es U, número de veces que un valor X precede a un valor Y. Si la población X está situada por encima de la población Y, entonces U será grande; si es cierto lo contrario, U será pequeño. También se ha tabulado la distribución de probabilidad de U para tamaños muestrales seleccionados. Puesto que el contraste es equivalente al de Wilcoxon, no es necesario insistir en su manejo.

EJERCICIOS 13.4

1. Se ha demostrado que el factor de crecimiento NGF es una proteína que influye en el desarrollo y mantenimiento de las neuronas simpáticas periféricas. Un enfoque del estudio del NGF consiste en privar al animal del factor y estudiar el efecto de su ausencia en distintos tipos de células. En este estudio, consideramos dicho efecto sobre el contenido proteínico total de las raíces de los ganglios dorsales de las ratas. Se comparan dos grupos de ratas: las nacidas de hembras deficientes en NGF (en el útero) y las nacidas de hembras normales, pero amamantadas por hembras deficientes en NGF (en la leche). Se obtienen los siguientes datos:

Contenido proteínico total, en miligramos de proteína por raíz de ganglio dorsal

En el útero (U)		**En la leche (M)**	
0.12	0.09	0.19	0.20
0.19	0.13	0.21	0.22
0.17	0.21	0.21	
0.20		0.23	

¿Indican estos datos al nivel $\alpha = 0.05$ que el contenido proteínico total tiende a ser menor entre las ratas privadas de NGF en el útero que entre las privadas del factor en la leche?

2. Los bifenilos policlorados (PCB) son contaminantes del medio ambiente en todo el mundo, de origen industrial, que están relacionados con el DDT. En Estados Unidos van siendo eliminados, pero permanecen en el medio durante muchos años. Se realiza un experimento para estudiar los efectos del PCB en la capacidad reproductiva de las lechuzas chillonas. El propósito es comparar el espesor de la cascara de los huevos producidos por las aves expuestas al PCB con los de las aves no expuestas al contaminante. Se cree que las cascaras del primer grupo serán menos consistentes que las del último. ¿Sostienen los siguientes datos esta hipótesis de investigación? Explicarlo.

Espesor de la cáscara, mm			
Expuestos al PCB *(E)*		Libres del PCB *(F)*	
0.21	0.226	0.22	0.27
0.223	0.215	0.265	0.18
0.25	0.24	0.217	0.187
0.19	0.136	0.20	0.256
0.20		0.23	

3. En un estudio sobre las características de los pacientes que sufrieron infartos de miocardio se compara el volumen cardíaco de aquellos para los que la duración del dolor fue menor de ocho horas con el de los pacientes cuyo dolor duró ocho horas o más. Se obtuvieron los siguientes datos:

Volumen cardíaco, mL			
Menos de 8 horas (<)		8 horas o más (≥)	
793.4	760.5	979.1	940.7
906.5	856.6	797.0	1009.9
604.1	899.1	961.8	1330.3
646.8	806.8	1100.6	909.3
688.1	968.1	843.6	812.4
		739.4	850.0
		1335.8	818.9

¿Apoyan estos datos la hipótesis que se investiga de que el volumen cardíaco de aquellos cuya experiencia de dolor fue menor de ocho horas tiende a ser menor que el de aquellos para los que fue de ocho horas o más?

4. La anemia de células falciformes (drepanocitosis) es una enfermedad asociada con una deficiente eliminación de potasio a través de la orina. Se realiza un estudio para comparar las respuestas de los sujetos con hemoglobina normal con las de los pacientes con drepanocitosis, frente a una dosis oral (0.75 meq/kg de peso corporal) de cloruro potásico (KC1). Antes de que los pacientes reciban la dosis de KC1 no se detectan diferencias en el pH de la orina. Se obtuvieron los siguientes datos al final del estudio. ¿Indican que hay diferencia en la forma en que estos grupos responden a una dosis oral de KC1, en relación con la variable pH de la orina?

PH en la orina			
Normal (N)		**Células en casco *(S)***	
6.6	**5.9**	**5.7**	**5.2**
6.1	**5.4**	**5.6**	**5.6**
6.2	**5.7**	**5.3**	**5.9**
5.8	**4.7**	**5.4**	**6.0**
		4.8	

5. *Aproximación normal a W_m.* Para muestras grandes, W_m está, aproximadamente, normalmente distribuida con media $\mu = m(m + n + 1)/2$ y varianza $\sigma^2 = mn(m + n + 1)/12$. Esto puede utilizarse para contrastar hipótesis que reclaman el empleo del procedimiento de Wilcoxon para tamaños de muestras no incluidas en la Tabla XII del Apéndice B. Necesitamos solamente tipificar W_m y utilizar la tabla normal tipificada (Tabla III del Apéndice B) para aproximar el valor P del contraste.
   a) Sean $m = 30$ y $n = 60$. Hallar $P[W_m \leq 1350]$.
   b) Sean m = 40 y $n = 50$. Hallar $P[W_m \geq 2000]$.

13.5. CONTRASTE DE POSICIÓN DE KRUSKAL-WALLIS PARA k-MUESTRAS: DATOS NO ASOCIADOS

La idea de utilizar sumas de rangos para comprobar dos poblaciones basadas en muestras aleatorias independientes extraídas de poblaciones, puede extenderse a más de dos poblaciones. El contraste resultante fue desarrollado por W. H. Kruskal y W. A. Wallis, en 1952

Contraste para k-muestras de Kruskal-Wallis

Supongamos que de k poblaciones continuas se extraen muestras aleatorias independientes de tamaños $n_1, n_2, n_3, \ldots n_k$, respectivamente. Queremos contrastar la hipótesis nula de que estas poblaciones son idénticas mediante un contraste que es especialmente sensible a diferencias de posición. Para hacerlo, las $n_1 + n_2 + n_3 + \cdots + n_k = N$ observaciones se agrupan y se ordenan de menor a mayor. De esa forma, reciben un rango de 1 a N, asignándose a las coincidencias contabilizadas la media de los rangos que les corresponderían, como en los procedimientos de Wilcoxon.

Sea R_i, $i = 1, 2, \ldots, k$, la suma de los rangos asociada con las observaciones extraídas de la i-ésima población. Si la hipótesis nula de no diferencia entre las poblaciones es cierta, entonces los rangos más altos se dispersarán aleatoriamente a través de las k muestras; si una o más poblaciones se sitúan por encima de las otras, entonces en las muestras que se extraigan de estas poblaciones los rangos más altos estarán agrupados. De este modo, si H_0 es cierta, el rango medio asociado con cada grupo será de tamaño moderado; de lo contrario se disparará el valor de una o más de estas medias de rangos. El estadístico de Kruskal-Wallis viene dado por

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

donde $\bar{R}_i = R_i/n_i$ es el rango medio de las observaciones extraídas de la i-ésima población. Utilizando los métodos de los Capítulos 2 y 4 podemos demostrar que, si la hipótesis nula es cierta, entonces $E[\bar{R}_i] = (N + 1)/2$. Así, el estadístico de Kruskal-Wallis compara esencial-

mente los rangos promedio observados para las k muestras con los esperados bajo H_0. Si hay una discrepancia considerable, entonces H será grande. Ello implica que deberá rechazarse H_0 para valores grandes de H. Si H_0 es cierta, H sigue aproximadamente una distribución ji-cuadrado con k - 1 grados de libertad. Por tanto, los valores P para el contraste se hallan en la Tabla VIII del Apéndice B. Como en otros casos, hay una forma para H que es aritméticamente equivalente y mucho más fácil de manejar para el cálculo; esta es

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Será la que utilizaremos en la práctica.

Ejemplo 13.5.1. Para determinar el efecto de la hemodiálisis sobre el tamaño del hígado se estudian tres poblaciones: controles normales, pacientes renales no dializados y pacientes dializados. Se obtienen muestras aleatorias de cada población y se utilizan aparatos de medida para determinar el área del hígado (en centímetros cuadrados) para cada individuo. Se obtienen los siguientes datos (el rango de cada observación viene dado entre paréntesis):

I (controles normales)		II (pacientes no dializados)		**III** (pacientes dializados)	
206.9 (14)	143.8 (2)	194.6(11)	143.0 (1)	288.0 (21)	249.0 (19)
150.0 (5)	192.6 (10)	145.6 (3)	170.0 (6)	269.2 (20)	346.1 (23)
197.3 (12)		174.9 (8)		288.3 (22)	216.6 (16)
173.2 (7)		187.5 (9)		357.5 (24)	202.6 (13)
147.2 (4)		223.4 (17)		229.2 (18)	213.5 (15)

Las sumas de los rangos son

$$R_1 = 14 + 5 + 12 + \cdots + 10 = 54$$
$$R_2 = 11 + 3 + 8 + \cdots + 6 = 55$$
$$R_3 = 21 + 20 + 22 + \cdots + 15 = 191$$

Obsérvese que $n_1 = n_2 = 7$, $n_3 = 10$ y $N = 24$. El valor observado del estadístico de Kruskal-Wallis para estos datos es

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{24(25)} \left(\frac{54^2}{7} + \frac{55^2}{7} + \frac{191^2}{10} \right) - 3(25) = 14.94$$

El número de grados de libertad asociados con el estadístico ji-cuadrado es k - 1 = 2. De la Tabla VIII se obtiene

$$P[X_2^2 \geq 14.94] < 0.005$$

Como este valor P es pequeño, podemos concluir que existen diferencias en el tamaño del hígado entre las tres poblaciones.

El contraste de Kraskal-Wallis supone poblaciones continuas. En todo caso, puede ser aplicado sin reserva a datos que originalmente son rangos. Si las poblaciones muestreadas son normales con varianzas iguales, la hipótesis que está siendo contrastada es la misma que la de la clasificación de una vía, diseño completamente aleatorio, en el análisis de la varianza.

EJERCICIOS 13.5

1. El déficit de vitamina A es, en muchos lugares, un conocido problema de salud pública. Un aporte dietético inadecuado de vitamina A es el factor más importante entre los responsables del déficit. La fuente principal de vitamina A es el β caroteno, derivado de los vegetales. Se ha demostrado que, añadiendo vegetales de hoja verde a la dieta, se obtiene un aumento de suero en las concentraciones de vitamina A. Se realiza un estudio para determinar si la adición de grasa a la dieta produce algún beneficio. Un grupo de 30 niños, para cada uno de los cuales la concentración de vitamina A en suero es menor de 20 mg de vitamina A/100 mL de suero sanguíneo, se divide aleatoriamente en tres subgrupos. Cada subgrupo recibe diariamente 40 g de espinacas, pero el contenido en grasa de la dieta varía. Al final del experimento, se obtienen los siguientes datos sobre la concentración en suero de vitamina A.

I (no se le añadió grasa)	II (añadidos 5 g de grasa)	III (añadidos 10 g de grasa)
18.1	29.1	26.6
16.5	15.8	16.1
21.0	20.4	18.8
18.7	23.5	25.0
7.4	18.5	21.8
12.4	21.3	15.4
16.1	23.1	19.9
17.9	23.8	15.5
	20.1	21.1
	11.9	25.5

   a) Calcular la suma de los rangos en cada grupo.
   b) Contrastar la hipótesis nula de que el contenido de grasa de la dieta no tiene efecto sobre la concentración de vitamina A en suero.
2. La ureasa es una enzima conocida por producir amoníaco en el tracto gastrointestinal. Se sabe que el amoníaco es nocivo para los pacientes con enfermedades hepáticas. Se realiza un estudio para comparar la concentración de ureasa en los jugos gástricos de cinco poblaciones: controles normales, I; pacientes con obstrucción extrahepática de la vena porta, II; pacientes con tumores hepáticos amebianos, III; pacientes con hepatitis vírica, IV; y pacientes con hipertensión portal idiopática, V, respectivamente. Se obtienen los siguientes datos (en miligramos por mililitro):

I	II	III	IV	V
261.1	221.9	201.4	600.9	160.6
186.2	188.7	146.1	301.2	135.0
239.1	167.6	96.8	607.9	455.1
243.3	224.9	173.9	283.3	402.3
296.8	178.8	280.8	193.3	457.9
270.5	147.9	100.3	159.4	559.6

Basándose en los datos y en el contraste de Kruskal-Wallis, ¿se puede afirmar que estas poblaciones difieren con respecto a la concentración gástrica de ureasa?

3. Pueden presentarse envenenamientos debidos a contactos prolongados con residuos de pesticida agrícolas. La absorción de pesticida a través de la piel varía según la región anatómica del cuerpo. La palma de la mano es especialmente sensible. Se realiza un estudio para comparar el tiempo de exposición de la mano derecha en cinco tipos diferentes de trabajadores agrícolas. Se recogen los datos filmando a los trabajadores durante un período de diecisiete minutos y determinando para cada uno cuánto tiempo está la mano derecha en contacto con las plantas. Se obtienen los siguientes:

Recolectores de tabaco	**Empaquetadores de algodón**	**Conserveros de melocotón**	**Recolectores de arándanos**	**Empaquetadores de maíz dulce**
14.4	11.7	15.8	16.1	14.3
12.9	13.8	16.2	16.0	14.5
12.0	14.2	15.9	15.9	14.8
13.9	10.3	16.7	16.4	14.9
13.3	7.0	16.4	16.6	14.0

Utilizando el contraste de Kruskal-Wallis, ¿se puede pretender que difiere el tiempo de exposición de la mano derecha entre estos grupos de trabajadores? Explicar la respuesta basándose en el valor *P* del contraste.

4. Puesto que el hígado es el principal lugar para el metabolismo de los fármacos, se espera que los pacientes con enfermedades hepáticas tengan dificultades en su eliminación. Uno de tales fármacos es la fenilbutazona. Se realiza un estudio de la respuesta del sistema a este fármaco. Se estudian tres grupos: controles normales, pacientes con cirrosis hepática y pacientes con hepatitis crónica activa. A cada individuo se le suministran por vía oral 19 mg de fenilbutazona/kg de peso corporal. Basándose en los análisis de sangre, se determina para cada uno el tiempo de máxima concentración en plasma (en horas). Se obtienen estos datos:

Normal	Cirrosis	Hepatitis
4.0	22.6	16.6
30.6	14.4	12.1
26.8	26.3	7.2
37.9	13.8	6.6
13.7	17.4	12.5
49.0		15.1
		6.7
		20.0

Basándose en el contraste de Kruskal-Wallis, ¿se puede concluir que las tres poblaciones difieren respecto del tiempo de máxima concentración en plasma de la fenilbutazona? Explicar la respuesta basándose en el valor *P* del contraste.

5. Las glándulas nasales supraorbitarias (glándulas de la sal) tienen una importante función en las aves marinas: ayudan a excretar cloruro sódico cuando las condiciones del medio fuerzan al ave a consumir más sal de lo normal. Se realiza un estudio para determinar el

papel de estas glándulas en la excreción de plomo, un contaminante común del medio ambiente. Se estudian tres grupos de ánades: controles normales, I; ánades alimentados a la fuerza con una dosis de plomo comercial de perdigones, II; y ánades alimentados con perdigones de plomo y $CaNa_2EDTA$, III. Se obtuvieron los siguientes datos sobre la concentración de plomo (en microgramos de plomo por gramo de tejido) en las glándulas nasales:

I	II	III
1.4	11.1	5.0
1.0	10.3	8.2
0.9	10.2	4.9
0.7	9.7	3.2
0.5	7.7	4.4
1.2	10.1	3.1
3.4	11.6	5.1
1.3	13.3	2.9

Utilizar el contraste de Kruskal-Wallis para determinar si existen diferencias en la concentración de plomo en las glándulas nasales entre las tres poblaciones.

13.6. CONTRASTE DE POSICIÓN DE FRIEDMAN PARA *k*-MUESTRAS: DATOS ASOCIADOS

En esta sección introducimos un contraste de distribución libre para k poblaciones idénticas cuando las observaciones de las k poblaciones están asociadas. El contraste, desarrollado por M. Friedman, en 1937, es especialmente sensible a diferencias de posición.

Contraste de Friedman

Supongamos que estamos interesados en comparar los efectos de k tratamientos. Se cree que hay una variable que, si bien no es de interés directo, puede interferir nuestra capacidad para detectar diferencias reales entre los k tratamientos. Queremos controlar esta variable extraña mediante la construcción de bloques. Esto es, dividimos las unidades experimentales en b grupos, o «bloques», cada uno de tamaño k, siendo los miembros de un mismo bloque tan iguales como sea posible respecto a la variable extraña. Asignamos aleatoriamente los k tratamientos a las unidades de los bloques. Para contrastar la hipótesis nula de efectos idénticos de los tratamientos, le asignamos a las observaciones en cada bloque un rango de 1 a k (de menor a mayor), asociando a las coincidencias contabilizadas la media de los rangos que les corresponderían. A continuación calculamos R_i, $i = 1,2,..., k$, la suma de los rangos asociados con cada uno de los k tratamientos. Si H_0 es cierta, cada una de estas sumas de rango deberá tomar un valor moderado; de otra forma habría diferencias sustanciales entre ellos. Puede demostrarse que el valor esperado de cada suma de rangos bajo H_0 es $b(k + 1)/2$.

El *estadístico de Friedman S* se utiliza para comparar las sumas de rangos observadas con el valor esperado:

$$S = \sum_{i=1}^{k} \left[R_i - \frac{b(k+1)}{2} \right]^2$$

Obsérvese que, si H_0 es cierta, cada una de las diferencias $R_i - b(k+1)/2$ deberá ser pequeña, proporcionando un valor también pequeño de 5; de lo contrario, S será grande. La distribución de probabilidades de *S* ha sido tabulada para un número limitado de valores de *k* (número de tratamientos) y de *b* (número de bloques). En todo caso, se ha visto que el estadístico

$$\frac{12S}{bk(k+1)}$$

sigue aproximadamente una distribución ji-cuadrado con $k - 1$ grados de libertad, si H_0 es cierta. Este estadístico nos sirve como estadístico del contraste rechazándose H_0 si alcanza valores grandes. Los valores de *P* se leen en la Tabla VIII del Apéndice B.

En el Ejemplo 13.6.1 se describe la utilización del contraste de Friedman

Ejemplo 13.6.1. Recientes investigaciones han demostrado que cantidades de petróleo bruto, incluso del orden del microlitro, aplicadas a la superficie de los huevos fecundados de diversas especies avícolas, pueden dar como resultado el deterioro del embrión. Aparte de un alto contenido de hidrocarburos aromáticos, ciertos crudos de petróleo contienen altas concentraciones de níquel y vanadio. Se estudian los efectos de estos elementos sobre el desarrollo embrional del pato silvestre. Puesto que patos de distintas nidadas tienen distintos antecedentes genéticos, el factor se controla mediante la construcción de bloques. Se utilizan en el experimento seis nidadas. De cada nidada se seleccionan aleatoriamente cuatro huevos y a cada uno de ellos se le asigna aleatoriamente uno de estos cuatro tratamientos:

I. No se le aplica crudo de petróleo (control).
II. Se le trata con 1 μL de crudo de petróleo.
III. Se le trata con 1 μL de crudo de petróleo con 700 ppm de vanadio.
IV. Se le trata con 1 μL de crudo de petróleo con 700 ppm de níquel.

Se deja que los huevos se desarrollen durante dieciocho días; al final de este tiempo se determina el peso en gramos de cada uno de ellos. Se obtienen los siguientes datos (el rango de cada observación *dentro de su bloque* viene dado entre paréntesis):

	Tratamiento			
Bloque (nidada)	**I**	**II**	**III**	**IV**
1	16.77 (4)	14.34 (2)	16.08 (3)	14.29 (1)
2	15.61 (4)	11.92(1)	13.22 (2)	13.95 (3)
3	14.46 (4)	14.45 (3)	11.72 (1)	13.59 (2)
4	13.08 (3)	16.11 (4)	10.18 (1)	12.22 (2)
5	14.47 (4)	12.85 (1)	13.52 (3)	13.22 (2)
6	12.01 (2)	16.13 (4)	12.68 (3)	10.73 (1)

$$R_1 = 4 + 4 + 4 + 3 + 4 + 2 = 21$$
$$R_2 = 2 + 1 + 3 + 4 + 1 + 4 = 15$$
$$R_3 = 3 + 2 + 1 + 1 + 3 + 3 = 13$$
$$R_4 = 1 + 3 + 2 + 2 + 2 + 1 = 11$$

Obsérvese que, si H_0 es cierta, las sumas de rangos deberán ser próximas al valor esperado de

$$\frac{b(k+1)}{2} = \frac{6(4+1)}{2} = 15$$

Parece que hay alguna desviación de este valor. ¿Es la desviación lo suficientemente grande como para concluir que hay diferencias entre los cuatro tratamientos? Para responder calculamos el valor del estadístico S de Friedman y lo utilizamos para hallar el valor del contraste

$$X^2_{k-1} = \frac{12S}{bk(k+1)}$$

Para estos datos

$$S = \sum_{i=1}^{k} \left[R_i - \frac{b(k+1)}{2} \right]^2$$

$$= \sum_{i=1}^{4} \left[R_i - \frac{6(4+1)}{2} \right]^2$$

$$= (21-15)^2 + (15-15)^2 + (13-15)^2 + (11-15)^2$$

$$= 36 + 0 + 4 + 16 = 56$$

El valor observado del estadístico $X^2_{k-1} = X^2_3$ es

$$\frac{12S}{bk(k+1)} = \frac{12(56)}{6(4)(5)} = 5.6$$

En la Tabla VIII del Apéndice B vemos que el valor P está entre 0.25 (punto crítico, 4.11) y 0.1 (punto crítico, 6.25). Puesto que es grande, no podemos concluir que hay diferencias entre los tratamientos.

El contraste de Friedman supone poblaciones continuas. Sin embargo, puede utilizarse para analizar datos que originalmente son rangos. Es, para distribuciones libres, lo análogo a la teoría normal del diseño de bloques completos aleatorizados.

EJERCICIOS 13.6

1. Durante los últimos años, un incremento en el número de casos sospechosos de envenenamiento con pentaclorofenol (PCP) en las granjas de animales ha atraído la atención de los veterinarios. Se realiza un estudio para considerar los efectos de variar las cantidades de PCP sobre el peso total ganado por los cerdos (en kilogramos). Las carnadas son utilizadas como bloques para controlar los efectos de las diferencias naturales entre animales de diferentes familias. Se utilizan los siguientes tratamientos.

   I. Alimentado con lactosa (controles).
   II. Alimentado con lactosa y 5 mg de PCP purificado/kg de peso.

III. Alimentado con lactosa y 10 mg de PCP purificado/kg de peso.
IV. Alimentado con lactosa y 15 mg de PCP purificado/kg de peso.

Se obtienen los siguientes datos sobre el peso total ganado al final del período experimental:

	Tratamiento			
Bloque (carnada)	I	II	III	IV
1	8.9	6.6	5.6	4.2
2	7.2	6.9	7.3	6.9
3	3.1	6.2	7.2	4.1
4	7.1	8.3	6.3	5.8
5	6.7	6.4	5.9	9.4
6	5.3	6.7	8.0	7.9
7	2.4	5.5	6.1	3.1
8	5.7	9.2	9.6	4.2

Si el nivel de PCP no afecta al total del peso ganado, ¿a qué valor estará próxima cada suma de rangos? Utilizar el contraste de Friedman para comparar los efectos de estos cuatro tratamientos sobre el total de peso ganado.

2. Se realiza una investigación para comparar la concentración de mercurio en el encéfalo, la musculatura y los tejidos oculares de truchas expuestas a dosis subletales (0.3 unidades tóxicas) de metilo de mercurio. Se utilizan 12 truchas en el experimento; cada una de ellas se considera como un bloque. Se obtienen los siguientes datos sobre la concentración (en microgramos de mercurio por gramo de tejido):

	Tejido				**Tejido**		
Bloque (trucha)	Encéfalo	**Musculatura**	Ojo	**Bloque (trucha)**	**Encéfalo**	**Musculatura**	**Ojo**
1	1.65	0.98	0.49	**7**	1.22	1.24	0.43
2	1.37	1.17	0.40	**8**	1.66	1.01	0.57
3	1.48	1.05	0.44	**9**	1.49	0.86	0.87
4	1.40	1.45	0.55	10	1.67	1.13	0.52
5	1.61	0.96	0.43	11	1.31	1.18	0.46
6	1.59	1.00	0.39	**12**	1.55	1.17	0.45

Si no hay diferencias en las concentraciones de mercurio en estos tres tipos de tejidos, ¿a qué valor estaría próxima cada suma de rangos? Utilizar el contraste de Friedman para comparar los efectos de 0.3 unidades tóxicas de metilo de mercurio sobre la concentración de mercurio en el encéfalo, la musculatura y el tejido ocular de la trucha.

3. El administrador de un laboratorio está considerando la compra de un aparato para analizar muestras de sangre. En el mercado hay cinco de tales aparatos. Se le pide a cada uno de ocho técnicos que, después de probar los aparatos, les asignen un rango de acuerdo

con el orden de preferencia, dándole el rango 1 al preferido. Se obtienen los siguientes; datos:

	Aparato				
Bloque (técnico)	**I**	**II**	**III**	**IV**	**V**
1	**1**	**3**	**4**	**2**	**5**
2	**4**	**5**	**1**	**2**	**3**
3	**4**	**1**	**3**	**5**	**2**
4	**4**	**1**	**5**	**2**	**3**
5	**1**	**3**	**2**	**5**	**4**
6	**1**	**2**	**3**	**4**	**5**
7	**5**	**1**	**3**	**2**	**4**
8	**5**	**1**	**4**	**3**	**2**

Si no hay preferencia clara, ¿a qué valor estará próxima cada suma de rangos? Utilizar el contraste de Friedman para determinar si los técnicos perciben diferencias entre los aparatos.

13.7. CORRELACIÓN

En el Capítulo 11 comentamos el problema de medir el grado de asociación lineal entre dos variables aleatorias X e Y. Se realiza mediante el coeficiente de correlación momento-producto de Pearson. Ahora introducimos una medida de asociación lineal, que es particularmente útil cuando, o bien los datos son rangos, o bien el conjunto de datos es lo bastante pequeño como para que se puedan distribuir rangos con facilidad. El método fue introducido por G. Spearman, en 1904.

Coeficiente de correlación de rangos de Spearman

Consideremos un conjunto $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ de n observaciones emparejadas de las variables aleatorias X e Y. Para estimar la correlación entre X e Y, primero asociamos rangos de 1 a n, de menor a mayor, a las observaciones sobre X. Después hacemos lo mismo sobre Y. De este modo, generamos un conjunto de n rangos emparejados, que designamos por $(r_{x_1}, r_{y_1}), (r_{x_2}, r_{y_2}), (r_{x_3}, r_{y_3}), \ldots, (r_{x_n}, r_{y_n})$. De nuevo, asignamos a las coincidencias contabilizadas la media de los rangos que les corresponderían. La correlación estimada entre X e Y se halla calculando el coeficiente de correlación de Pearson para el conjunto de rangos emparejados. Esto es, definimos el *coeficiente de correlación de los rangos, r_s, de Spearman* por

$$r_s = \frac{n\sum r_x r_y - \sum r_x \sum r_y}{\sqrt{[n\sum r_x^2 - (\sum r_x)^2][n\sum r_y^2 - (\sum r_y)^2]}}$$

El procedimiento proporciona resultados que son ligeramente diferentes de los del método de Pearson. En todo caso, para muestras de tamaño grande hay una cierta concordancia. Además, si *no hay coincidencias,* el coeficiente de Spearman viene dado por

$$r_2 = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

donde $d_i = r_{x_i} - r_{y_i}$ es la diferencia entre los rangos de X e Y. De este modo, si no hay coincidencias, es más fácil calcular el coeficiente de Spearman que el de Pearson. La interpretación es la misma en ambos casos. Es decir, valores próximos a 1 indican una correlación positiva fuerte; valores grandes de X tienden a ser asociados con valores grandes de Y. Valores próximos a -1 indican una correlación negativa fuerte; valores grandes de X tienden a asociarse con valores pequeños de Y. Valores próximos a 0 indican asociación no lineal. Esto sugiere que no se utilice el procedimiento de Spearman si hay un gran número de coincidencias.

Describimos el procedimiento de Spearman en el Ejemplo 13.7.1.

Ejemplo 13.7.1. Se realiza un estudio para determinar la asociación lineal entre la concentración de nicotina en sangre de un individuo y el contenido en nicotina de un cigarrillo. Se obtuvieron los siguientes datos (los rangos están entre paréntesis):

X (concentración de nicotina en sangre, nmol/litro)	Y (contenido de nicotina por cigarrillo, mg)
185.7 (2)	1.51 (8)
197.3 (5)	0.96 (3)
204.2 (8)	1.21 (6)
199.9 (7)	1.66 (10)
199.1 (6)	1.11 (4)
192.8 (3)	0.84 (2)
207.4 (9)	1.14 (5)
183.0 (1)	1.28 (7)
234.1 (10)	1.53 (9)
196.5 (4)	0.76 (1)

Para estos datos,

$$\sum r_x = \sum r_y = 55 \qquad \sum r_x^2 = \sum r_y^2 = 385$$

$$\sum r_x r_y = 2(8) + 5(3) + 8(6) + \cdots + 4(1) = 325$$

Calculando r_s directamente, obtenemos

$$r_s = \frac{n\sum r_x r_y - \sum r_x \sum r_y}{\sqrt{[n\sum r_x^2 - (\sum r_x)^2][n\sum r_y^2 - (\sum r_y)^2]}}$$

$$= \frac{10(325) - 55(55)}{\sqrt{[10(385) - 55^2][10(385) - 55^2]}}$$

$$= 0.27$$

Puesto que no hay coincidencias, puede calcularse r_s mediante la fórmula abreviada

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6[(2-8)^2 + (5-3)^2 + (8-6)^2 + \cdots + (4-1)^2]}{10(10^2-1)}$$

$$= 1 - \frac{6(120)}{10(99)} = 0.27$$

Ya que el coeficiente de correlación de Spearman no difiere mucho del de Pearson, podemos aproximar el coeficiente de determinación r^2 mediante r_s^2. En este caso $r_s^2 = 0.07$. De este modo, el 7 %, aproximadamente, de la variación en la concentración de nicotina en sangre, puede atribuirse a asociación lineal con la nicotina producida por el cigarrillo. Puesto que este valor es pequeño interpretamos r_s como indicativo de una muy pequeña correlación positiva entre *X e Y.*

EJERCICIOS 13.7

1. La venoclisis continua de un fármaco a velocidad constante se supone que mantiene su concentración en suero a un nivel constante y predecible. Este parece ser el caso respecio a un único paciente. Sin embargo, aparecen problemas porque la misma dosis de fármaco no produce necesariamente la misma concentración en suero en diferentes pacientes. Para estudiar este fenómeno en pacientes con meningitis purulenta grave, tratados con ampicilina, se somete a 24 pacientes a la acción del fármaco con velocidad constante. Se cree que hay una relación entre la concentración de ampicilina en suero y la tasa de aclaramiento de la creatinina. Se obtienen estos datos:

x (aclaramiento de creatinina, mL/min)	*y* (concentración de ampicilina en suero, mg/litro)	*x* (aclaramiento de creatinina, mL/min)	*y* (concentración de ampicilina en suero, mg/litro)
69.0	60.0	112.6	12.8
70.0	38.5	116.5	35.0
81.0	92.0	120.2	13.3
85.0	19.2	120.0	59.6
85.1	41.0	122.5	47.8
84.0	69.2	130.1	24.6
95.0	42.7	135.0	18.0
96.1	78.1	132.3	48.5
100.0	50.0	155.7	21.2
107.0	20.0	170.0	23.0
107.5	28.3	121.0	7.4
110.0	15.0	125.0	10.0

a) Asignar rangos del 1 al 24 a cada uno de los conjuntos de observaciones.
b) Hallar $\sum r_x$, $\sum r_y$, $\sum r_x^2$, $\sum r_y^2$ y $\sum r_x r_y$. Utilizar estos valores para hallar r_s a partir de su definición.
c) Puesto que no hay coincidencias puede utilizarse la fórmula abreviada. Comprobar la respuesta dada al apartado *b* calculando r_s mediante esta fórmula.

d) La correlación estimada por el coeficiente de Pearson es -0.49. Comparar este valor con el de Spearman.

e*)* ¿Parece que hay una fuerte asociación lineal entre *X* e Y? Explicar la respuesta basándose en el valor estimado de r^2_s.

2. Se estudian pacientes con osteoporosis, una enfermedad de los huesos que produce una disminución de la masa ósea. Estos pacientes son tratados con un fragmento de hormona paratiroidea humana. Se espera que exista una fuerte relación entre la tasa de aumento de calcio durante el tratamiento y el volumen óseo final; de esta forma, puede utilizarse esta variable en el futuro para predecir el volumen óseo final. Se obtienen los siguientes resultados:

x (tasa de aumento de calcio, mmol/24h)	*y* (volumen óseo trabecular final, %)	*x* (tasa de aumento de calcio, mmol/24h)	*y* (volumen óseo trabecular final, %)
2.5	**8.0**	**15.0**	**14.9**
3.0	**12.1**	**15.2**	**9.7**
4.0	**9.1**	**17.0**	**15.0**
4.6	**10.0**	**20.5**	**17.0**
6.1	**11.0**	**24.6**	**12.2**
7.5	**5.2**	**28.3**	**47.2**
11.0	**10.1**	**30.0**	**31.4**
12.3	**18.0**		

Hallar r_s. ¿Puede afirmarse que hay una fuerte correlación positiva entre *X* e F? Explicarlo sobre la base del valor aproximado de r^2_s.

3. Una característica de la visión es la «persecución suave», o capacidad de los ojos para seguir objetos que se mueven lentamente a través del campo visual. Se realiza un estudio para explorar la relación entre la velocidad máxima alcanzada por el ojo durante la persecución suave y el contenido de alcohol en sangre. Se utilizan doce individuos. Se determina para cada uno de ellos la velocidad máxima de persecución suave. Se pide después a estos individuos que beban whisky con agua o ginebra con tónica hasta que ellos mismos se consideren incapaces de conducir. En este momento, se mide nuevamente la velocidad máxima de persecución suave y se determina el contenido de alcohol en sangre del individuo. Se obtienen los datos siguientes:

x (alcohol en sangre, mg/dL)	*y* (velocidad de persecución suave, incrementos en %)	*x* (alcohol en sangre, mg/dL)	*y* (velocidad de persecución suave, incrementos en %)
20	**2**	**85**	**19**
45	**11**	**86**	**30**
68	**30**	**108**	**38**
68	**50**	**110**	**10**
70	**4**	**120**	**51**
75	**6**	**150**	**60**

Hallar r_s e interpretar su valor en términos prácticos.

4. Para marcar babuinos que se utilizarán en futuros estudios, hay que inmovilizar temporalmente a los animales. Un fármaco utilizado con este fin es el clorhidrato de fenciclidi-

na. Se realiza un estudio para determinar: 1) la relación entre la dosis administrada y el tiempo transcurrido hasta la inmovilización completa. 2) la relación entre la dosis administrada y el tiempo transcurrido desde la inmovilización completa hasta que se ven movimientos grandes. Se obtienen los siguientes resultados:

x (dosis, mg de fármaco/kg de peso corporal)	y (tiempo hasta la inmovilización, min)	z (tiempo hasta la recuperación, min)
1.21	7.6	100.2
1.36	8.2	100.1
1.78	8.7	100.5
1.10	7.9	90.4
1.57	8.0	97.7
1.49	7.4	87.8
1.59	7.7	79.5
1.02	8.5	119.8

a) Hallar e interpretar el coeficiente de correlación de Spearman entre X e Y.
b) Hallar e interpretar el coeficiente de correlación de Spearman entre X y Z.

13.8. CONTRASTE DE BARTLETT DE IGUALDAD DE VARIANZAS

Como se indicó en el Capítulo 10, uno de los supuestos para aplicar el contraste F de igualdad de medias en un problema de clasificación de una vía es que las varianzas poblacionales sean iguales. Una manera de prevenir que este supuesto no se cumpla es tomar muestras del mismo tamaño siempre que sea posible. No obstante, si hay diferencias en los tamaños de las muestras, es aconsejable contrastar

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_1: \sigma_i^2 \neq \sigma_j^2 \quad \text{para algún } i \text{ y } j \text{ (al menos hay dos varianzas diferentes)}$$

Si H_0 no se rechaza, comparamos medias utilizando el contraste F descrito en el Capítulo 10); si se rechaza H_0 aplicamos el contraste de Kruskal-Wallis dado en la Sección 13.5.

El contraste más utilizado para contrastar la hipótesis de igualdad de varianzas es el llamado *contraste de Bartlett.* Se puede demostrar que el estadístico que se emplea en este contraste sigue aproximadamente una distribución ji-cuadrado con k - 1 grados de libertad cuando las muestras proceden de poblaciones normales.

Para llevar a cabo la prueba de Bartlett calculamos las varianzas muestrales $S_1^2, S_2^2, \ldots, S_k^2$ para cada una de las k muestras. Calculamos también el error cuadrático medio, estimador de σ^2 bajo el supuesto de que H_0 es cierta. En este contexto, es conveniente calcular MS_E directamente a partir de las varianzas muestrales individuales por medio de la fórmula

$$MS_E = S_p^2 = \sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{N - k}$$

A continuación, formamos el estadístico Q definido por

$$Q = (N-k)\log_{10} S_p^2 - \sum_{i=1}^{k} (n_i - 1)\log_{10} S_i^2$$

El valor observado de este estadístico es grande cuando las varianzas muestrales S_i^2, $i = 1, 2, \ldots, k$ son considerablemente distintas; es cercano a 0 cuando estas varianzas muestrales toman valores próximos. El estadístico de Bartlett está definido por

$$B = \frac{2.3026Q}{h}$$

donde

$$h = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N-k}\right)$$

Daremos un ejemplo para mostrar cómo se utiliza el contraste de Bartlett.

Ejemplo 13.8.1. Volvamos al problema descrito en el Ejemplo 10.1.1, en el que se realiza un estudio para comparar la eficacia de tres programas de tratamiento del acné. Se han obtenido los siguientes datos sobre el porcentaje de mejoría en lesiones de acné por paciente, al cabo de 16 semanas de tratamiento.

Tratamiento					
I		II		III	
48.6	50.8	68.0	71.9	67.5	61.4
49.4	47.1	67.0	71.5	62.5	67.4
50.1	52.5	70.1	69.9	64.2	65.4
49.8	49.0	64.5	68.9	62.5	63.2
50.6	46.7	68.0	67.8	63.9	61.2
		68.3	68.9	64.8	60.5
				62.3	

Puesto que los tamaños de las muestras son distintos, lo mejor es contrastar igualdad de varianzas y decidir, a la vista de los resultados del contraste, si hay que analizar los datos mediante un contraste de Kruskal-Wallis. Para estos datos

$$n_1 = 10 \qquad n_2 = 12 \qquad n_3 = 13 \qquad N = 35$$
$$s_1^2 = 3.000 \qquad s_2^2 = 4.002 \qquad s_3^2 = 4.938 \qquad k = 3$$

Estas varianzas muestrales se utilizan para calcular MS_E, de la siguiente manera:

$$MS_E = S_p^2 = \sum_{i=1}^{3} \frac{(n_i - 1)S_i^2}{N-k}$$
$$= \frac{9(3.000) + 11(4.002) + 12(4.938)}{35 - 3}$$
$$= 4.071$$

El estadístico Q viene dado por

$$Q = (N - k) \log_{10} S_p^2 - \sum_{i=1}^{3} (n_i - 1) \log_{10} S_i^2$$

Los logaritmos se hallan mediante una calculadora y están dados por

$$\log_{10} S_p^2 = \log_{10} (4.071) = 0.6097$$
$$\log_{10} S_1^2 = \log_{10} (3.000) = 0.4771$$
$$\log_{10} S_2^2 = \log_{10} (4.002) = 0.6023$$
$$\log_{10} S_3^2 = \log_{10} (4.938) = 0.6936$$

El valor observado de Q es

$$Q = (35 - 3)(0.6097) - [(9)(0.4771) + 11(0.6023) + 12(0.6936)]$$
$$= 0.268$$

El estadístico de Bartlett está dado por

$$B = \frac{2.3026Q}{h}$$

donde

$$h = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N - k} \right)$$

En este caso

$$h = 1 + \frac{1}{3(2)} \left(\frac{1}{9} + \frac{1}{11} + \frac{1}{12} - \frac{1}{32} \right)$$
$$= 1.0424$$

y el valor observado del estadístico de Bartlett es

$$B = \frac{2.3026(0.268)}{1.0424} = 0.5919$$

Como $k = 3$, el valor P se halla teniendo en cuenta que $\mathbf{X}_{k-1}^2 = \mathbf{X}_2^2$. En la tabla de la ji-cuadrado vemos que

$$P = P[X_2^2 \geq 0.5919] > 0.25$$

Al ser este valor de P grande, no podemos rechazar H_0. NO tenemos evidencia de que las varianzas de las tres poblaciones sean distintas. En este caso, es razonable comparar medias con los contrastes F de análisis de la varianza de clasificación simple, como se hizo en el Capítulo 10.

EJERCICIOS 13.8

1. Después de un grave vertido accidental de una planta de una industria química cercana a un río, se llevó a cabo un estudio para determinar si ciertas especies de pescado capturado en el río difieren en cuanto a las cantidades de producto químico absorbido. Si se encon-

trasen diferencias, posiblemente se recomendaría tomar precauciones en el destinado al consumo humano. Se midieron, en partes por millón, muestras de capturas de las tres especies principales. Los datos obtenidos se dan a continuación.

Especies		
A	*B*	C
18.1	29.1	26.6
16.5	15.8	16.1
21.0	20.4	18.8
18.7	23.5	25.0
7.4	18.5	21.8
12.4	21.3	15.4
16.1	23.1	19.9
17.9	23.8	15.5
	20.1	21.1
	11.9	25.5

a) Utilícese el contraste de Bartlett para contrastar H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$.
b) Sobre la base del resultado de esta prueba, contrástese la igualdad de posición por medio del contraste *F* o el contraste de Kruskal-Wallis.

2. Utilícense los datos del Ejercicio 3 de la Sección 10.1 para contrastar $\mathbf{H_0}$: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = = \sigma_4^2 = \sigma_5^2$ ¿Era apropiado aquel primer análisis?

13.9. APROXIMACIONES NORMALES

Vimos en la Sección 4.5 que, para valores de *n* grandes, la distribución de Poisson es una buena aproximación a la binomial. La curva normal puede utilizarse para calcular probabilidades asociadas con cualquier variable. La justificación teórica de cualquier método de aproximación se basa en el Teorema central del límite, presentado en el Capítulo 6. El argumento que damos aquí se basa sólo en la intuición y en la evidencia empírica.

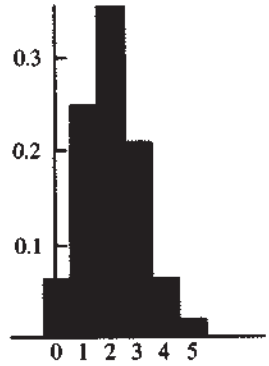
Para ver cómo las probabilidades binomiales pueden ser aproximadas razonablemente, consideramos una sucesión de variables binomiales. En particular, cuatro variables binomiales, cada una con probabilidad de éxito 0.4 y con valores de *n* de 5, 10, 15 y 20. En la Figura 13.7a a *d* se da, para cada una de estas variables, la densidad de probabilidad obtenida de la Tabla I del Apéndice B, y una representación de ella.

En la Figura 1*3.7d* se puede apreciar una característica: no es difícil imaginar una campana de Gauss que se ajusta mucho al diagrama de barras que se muestra. Esto sugiere que las probabilidades binomiales representadas por una o más barras en el diagrama pueden aproximarse razonablemente bien por un área cuidadosamente seleccionada bajo una curva normal escogida apropiadamente. ¿Cuál de las infinitas curvas normales es apropiada? El sentido común dice que la variable normal escogida debería tener igual media y varianza que la variable binomial que aproxima. El Teorema 13.9.1, que se ofrece sin demostración, resume estas ideas.

> **Teorema 13.9.1. Aproximación normal de** la **distribución binomial.** Sea *X* una variable binomial con parámetros *n* y *p.* Para *n* grande, *X* es aproximadamente normal con media *np* y varianza *np(l - p).*

$n = 5$, $p = 0.4$

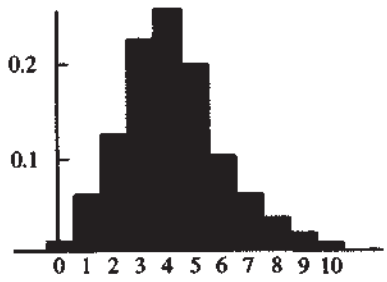
x	f(x)
0	0.0778
1	0.2592
2	0.3456
3	0.2304
4	0.0768
5	0.0102



(a)

$n = 10$, $p = 0.4$

x	f(x)
0	0.0060
1	0.0404
2	0.1209
3	0.2150
4	0.2508
5	0.2007
6	0.1114
7	0.0425
8	0.0106
9	0.0016
10	0.0001



(b)

$n = 15$, $p = 0.4$

x	f(x)	x	f(x)
0	0.0005	8	0.1181
1	0.0047	9	0.0612
2	0.0219	10	0.0245
3	0.0634	11	0.0074
4	0.1268	12	0.0016
5	0.1859	13	0.0003
6	0.2066	14	~0
7	0.1771	15	~0

(c)

$n = 20$, $p = 0.4$

x	f(x)	x	f(x)
0	~0	11	0.0710
1	0.0005	12	0.0355
2	0.0031	13	0.0145
3	0.0124	14	0.0049
4	0.0350	15	0.0013
5	0.0746	16	0.0003
6	0.1244	17	~0
7	0.1659	18	~0
8	0.1797	19	~0
9	0.1597	20	~0
10	0.1172		

(d)

Figura 13.7. Densidad para Xbinomial: *a) n* = 5, *p* = 0.4; *b) n* = 10, p = 0.4; *c) n* = 15, p = 0.4; *d) n* = 20, p = 0.4

Es cierto que el Teorema 13.9.1 es un poco ambiguo en el sentido de que la expresión *«n grande»* no está bien definida. En sentido matemático estricto, *grande* significa cuando n tiende a infinito. En la práctica, la aproximación es aceptable para valores de n y p tales que, o bien $p \leq 0.5$ y $np > 5$, o $p > 0.5$ y $n(1-p) > 5$.

Ejemplo 13.9.1. Se realiza un estudio para investigar la relación entre el hecho de que la madre fume durante el embarazo y los defectos de nacimiento en los niños. De las madres que participaron en el estudio, el 40 % fuma y el 60 % no. Cuando nacieron los bebés, 20 de ellos tenían un cierto defecto de nacimiento. Denotemos por X el número de niños cuyas madres fumaron durante el embarazo. Si no hay relación entre el hecho de que la madre fume durante el embarazo y el bebé nazca con defecto, entonces X es binomial con $n = 20$ y $p = 0.4$. ¿Cuál es la probabilidad de que 12 o más de los niños afectados tuvieran madres fumadoras?

Para responder a esta cuestión, hemos de hallar $P[X \geq 12]$, donde X es binomial con parámetros $n = 20$ y $p = 0.4$. Esta probabilidad puede leerse en la Tabla I del Apéndice B.

$$\begin{aligned} P[X \geq 12] &= 1 - P[X \leq 11] \\ &= 1 - 0.9435 \\ &= 0.0565 \end{aligned}$$

Obsérvese que como $p = 0.4 \leq 0.5$ *y* $np = (20)(0.4) = 8 > 5$, la aproximación normal debería dar un resultado próximo a 0.0565. Gráficamente, estamos tratando una variable aleatoria Y con media $np = 20(0.4) = 8$ y varianza $np(1-p) = 20(0.4)(0.6) = 4.8$. La probabilidad exacta de 0.0565 está dada por la suma de las áreas de las barras centradas en 12, 13,14,15,16,17, 18, 19 y 20, como se muestra en la Figura 13.8. La probabilidad aproximada está dada por el área bajo la curva normal sobre 11.5. Es decir,

$$P[X \geq 12] \cong P[Y \geq 11.5]$$

El número 0.5 se llama *corrección del medio punto* por continuidad. Se ha sustraído de 12 en la aproximación porque, de lo contrario, la mitad del área de la barra centrada en 12 hubiera sido ignorada, dando lugar a un error de cálculo innecesario. A partir de este punto, el cálculo es mecánico.

$$\begin{aligned} P[X \geq 12] &\cong P[Y \geq 11.5] \\ &= P\left[\frac{Y-8}{\sqrt{4.8}} \geq \frac{11.5-8}{\sqrt{4.8}}\right] \\ &= P[Z \geq 1.59] \\ &= 1 - P[Z \leq 1.59] \\ &= 1 - 0.9441 = 0.0559 \end{aligned}$$

Obsérvese que, incluso cuando n es tan pequeño como 20, el valor aproximado 0.0559 y el valor exacto 0.0565 son muy próximos. En la práctica, por supuesto, no se aproxima una probabilidad que puede encontrarse directamente en una tabla binomial. Aquí lo hemos hecho con fines pedagógicos.

La aproximación normal de la distribución de Poisson se maneja análogamente a la binomial. Su base teórica es también el Teorema central del límite. Es razonablemente precisa para valores grandes de los parámetros λs de Poisson y emplea el factor de corrección del medio punto para compensar el hecho de aproximar una distribución discreta por una curva continua. El procedimiento se basa en el Teorema 13.9.2.

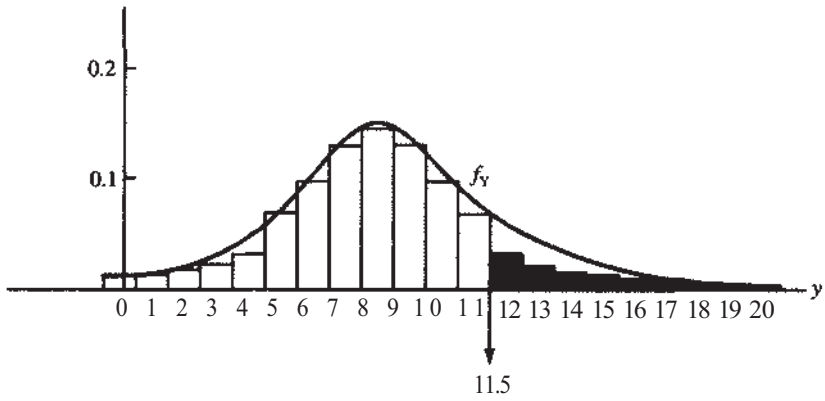


Figura 13.8. $P[X \geq 12]$ = áreas de las barras sombreadas $\cong$ área bajo la curva más allá de 11.5.

Teorema 13.9.2. Aproximación normal de la distribución de Poisson. Sea X una variable de Poisson con parámetro λs. Entonces, para valores grandes de λs, X es aproximadamente normal con media λs y varianza λs.

Ejemplo 13.9.2. Un hombre adulto sano tiene un promedio de 5 400 000 hematíes por mm^3 de sangre. Se examina una gota de 1/10 000 mm^3. ¿Cuál es la probabilidad de que el número de hematíes, X, en la gota esté entre 500 y 580?

La variable X es de Poisson con parámetro

$$\lambda s = 5\,400\,000 \frac{1}{10\,000} = 540$$

Por el Teorema 13.9.2, X es aproximadamente normal con media y varianza iguales a 540. Así, usando el factor de corrección del medio punto, obtenemos

$$\begin{aligned} P[500 \leq X \leq 580] &\cong P[499.5 \leq Y \leq 580.5] \\ &= P\left[\frac{499.5 - 540}{\sqrt{540}} \leq Z \leq \frac{580.5 - 540}{\sqrt{540}}\right] \\ &= P[-1.74 \leq Z \leq 1.74] \\ &= 0.9591 - 0.0409 = 0.9182 \end{aligned}$$

EJERCICIOS 13.9

1. Sea X binomial con $n = 20$ y $p = 0.3$. Utilizar la aproximación normal a la binomial para aproximar cada una de las probabilidades siguientes. Después, comparar los resultados con los valores obtenidos de la Tabla I del Apéndice B.
   a) $P[X \leq 3]$
   b) $P[3 \leq X \leq 6]$
   c) $P[X \geq 4]$
   d) $P[X = 4]$
2. Se ha calculado que el 10 % de los seres humanos tiene algún tipo de alergia. Se elige aleatoriamente un centenar de individuos y se les entrevista. Hallar la probabilidad de que al menos 12 tengan algún tipo de alergia. Hallar la probabilidad de que a lo sumo 8 tengan alergia.

3. La probabilidad de muerte como consecuencia del uso de pildoras anticonceptivas es de 3/100 000. De 1 000 000 de mujeres que utilizan este método de control de la natalidad, ¿cuántas muertes se esperan por esta causa? ¿Cuál es la probabilidad de que haya un máximo de 25 muertes por uso de la pildora? ¿Cuál es la probabilidad de que el número de muertes por esta causa esté entre 25 y 35, ambos inclusive?
4. Una prueba de laboratorio para detectar heroína en muestras de sangre tiene un 92 % de precisión. Si se analizan 72 muestras en un mes, cuál es la probabilidad de que
   a) 60 ó menos sean evaluadas correctamente.
   b) Menos de 60 sean evaluadas correctamente.
   c) Exactamente 60 sean evaluadas correctamente.
5. Uno de cada 400 bebés nacidos en un gran hospital de la ciudad padece la enfermedad genética fenilcetonuria (PKU). De los próximos 2000 bebés que nazcan en este hospital, ¿cuál es la probabilidad de que al menos uno tenga PKU?
6. Supóngase que la probabilidad de que crías de rata nazcan macho o hembra es la misma, en los laboratorios de una gran empresa de producción de animales de laboratorio. ¿Cuál es la probabilidad de que de 1000 animales nacidos, 445 ó más sean hembras?
7. Sea X una variable de Poisson con parámetros $\lambda s = 100$. Utilizar la aproximación normal para encontrar las siguientes probabilidades:
   a) $P[X \geq 95]$ *c)* $P[90 < X < 110]$
   b) $P[X \leq 80]$ *d)* $P[X = 99]$
8. El número medio de aviones que llegan o salen del aeropuerto de O'Hará es de uno cada 40 segundos. ¿Cuál es la probabilidad de que haya al menos 75 vuelos durante una hora elegida aleatoriamente? ¿Cuál es la probabilidad de que se den menos de 100 vuelos en una hora?
9. Un medio contiene 20 paramecios mortales (véase Ejercicio 4.5.7). ¿Cuál es la probabilidad de que en un período de 2 horas y media ninguno de ellos emita una partícula mortal? ¿Cuál es la probabilidad de que entre 5 y 10, inclusive, de ellos emitan al menos una partícula mortal? (Las partículas mortales son emitidas a una tasa de 1 cada 5 horas.)

13.10. UN CONTRASTE SOBRE PROPORCIONES PARA PEQUEÑAS MUESTRAS

En el Capítulo 8 se presentó un método para contrastar $\mathbf{H_0}$: $p = p_0$. El contraste se dedujo suponiendo que los tamaños de las muestras eran suficientemente grandes para aplicar el Teorema central del límite. Si se contrasta esta hipótesis sobre proporciones y el conjunto de datos de que se dispone es pequeño, puede utilizarse la distribución binomial para calcular los valores P. El estadístico de la prueba es X, número de observaciones de la muestra que poseen la característica.

Ejemplo 13.10.1. Se está investigando un nuevo fármaco para el tratamiento del cáncer de piel. Se espera que sea eficaz sobre una mayoría de los pacientes a los que se administra. La compañía que prepara el fármaco desea obtener evidencia estadística que apoye tal afirmación. Denotemos por p la proporción de pacientes en los cuales el fármaco será efectivo. Puesto que planteamos el contraste para rechazar la hipótesis nula, la alternativa aquí es que $p > 0.5$. Esto implica automáticamente que la hipótesis nula es la negación de H_1 a saber, que $p \leq 0.5$. Así, las dos hipótesis que intervienen son

$$\mathbf{H_0}: p \leq 0.5$$
$$\mathbf{H_1}: p > 0.5$$

Para contrastar H_0 se selecciona una muestra aleatoria de 20 pacientes, a los cuales se administra el fármaco. El estadístico que utilizaremos es *X,* número de pacientes en los cuales el fármaco es eficaz. Si la hipótesis nula es cierta, entonces *X* es binomial con $n = 20$, $p = 0.5$ y $E[X] = np = 10$. Así, si H_0 es cierta, esperamos que el fármaco sea eficaz en unos 10 pacientes; si H_1 es cierta y la tasa de eficacia es mayor que 0.5, entonces esperamos ver más de 10 personas mejoradas por el fármaco. Supóngase que cuando se realiza el experimento observamos 14 pacientes en los cuales el fármaco es efectivo. ¿Es este número tan alto como para rechazar H_0? Para decidirlo, calculemos el valor *P* de la prueba. En este caso

$$P = P[X \geq 14 | p = 0.5]$$

En la tabla binomial vemos que

$$\begin{aligned} P[X \geq 14 | p = 0.5] &= 1 - P[X \leq 13 | p = 0.5] \\ &= 1 - 0.9423 \\ &= 0.0577 \end{aligned}$$

Como este valor *P* es pequeño, rechazamos H_0 y concluimos que el fármaco es efectivo en una mayoría de pacientes.

EJERCICIOS 13.10

1. Un responsable de sanidad pública cree que más del 70% de los niños de menos de 3 años tratados en una cierta clínica recibe menos de la dosis diaria de vitamina A recomendada. Para apoyar esta afirmación, toma una muestra aleatoria de 15 niños y determina la dosis media de vitamina A en cada uno de ellos. El estadístico de la prueba es *X,* número de niños que reciben menos del promedio diario de 0.6 mg. La hipótesis que se contrasta es

$$H_0 : p \leq 0.7$$
$$H_1 : p > 0.7$$

donde *p* es la proporción de niños que reciben menos de la dosis diaria recomendada.
   a) Si H_0 es cierta y $p = 0.7$, ¿cuánto vale *E[X]7*
   b) Si cuando se realiza la prueba aparecen 12 niños que reciben menos del mínimo diario, ¿se rechazará H_0? Explicarlo sobre la base del valor *P* observado en el contraste.
2. Se ha descubierto un nuevo método para injertar naranjos. Se cree que la proporción de injertos que fallen, *p,* será menor que la tasa actual de 0.2. Para verificar esta afirmación, se utiliza el método nuevo sobre 20 árboles escogidos al azar, y el número de injertos que fallan, *X,* es el estadístico de prueba.
   a) Establecer las hipótesis nula y alternativa apropiadas para llevar a cabo el contraste.
   b) Si H_0 es cierta, ¿cuántos fallos se espera encontrar?
   c) Cuando se realizó el contraste, no hubo fallos. ¿Qué conclusión práctica debe extraerse? Explicarlo sobre la base del valor *P* del contraste.
3. Un vertido químico tuvo lugar en el río Xavier, y hay biólogos preocupados por la acumulación tóxica de la sustancia química en los tejidos de los peces de la zona inferior del río. Medidas previas mostraron que sólo un 20 % de los peces contiene más de 1.5 mg de la sustancia por kg de peso corporal. Se recoge una muestra de 16 peces después del vertido. El estadístico que hay que utilizar para detectar los efectos del vertido es *X,* número de peces que exceden el límite de 1.5 mg.

a) Establecer las hipótesis nula y alternativa apropiadas para detectar una situación en la cual la proporción p de peces que exceden el límite aceptado de 1.5 mg por kg de peso supera el 20 %.

b) Si H_0 es cierta, ¿cuántos de los 16 peces recogidos se espera que excedan el límite?

c) Si el valor observado de X es demasiado grande para ser atribuido al azar, debería rechazarse H_0. Suponer que de 16 peces examinados, se encuentra que 5 exceden el nivel de 1.5 mg. ¿Debe rechazarse H_0? Explicarlo a partir del valor P del contraste.

4. En este problema, se verá la interrelación que existe entre α, β, potencia y tamaño de la muestra. En cada caso predeterminamos un conjunto de valores que deciden el rechazo de H_0. ES decir, predeterminamos una región crítica. Al hacer esto, prefijamos α. Desde el apartado *a* hasta *d,* α se mantiene constante mientras que n puede aumentar. Se trata de poner de manifiesto el efecto que esto tiene en la potencia.

Va a probarse un procedimiento quirúrgico nuevo. El procedimiento actual es eficaz en el 50 % de los pacientes sobre los que se practica. Si puede demostrarse que el procedimiento nuevo es eficaz en el 60 % de los casos, será ampliamente utilizado. Así, queremos contrastar

$$H_0: p = 0.5$$
$$H_1: p = 0.6$$

El estadístico es X, número de pacientes en los cuales el procedimiento nuevo es eficaz. Así, si H_0 es cierta, X es binomial con parámetros n y $p = 0.5$, $E[X] = n(0.5)$ y Var $[X] = n(0.5)(0.5)$; si H_1 es cierta, X es binomial con parámetros n y $p = 0.6$, $E[X] = n(0.6)$ y Var $[X] = n(0.6)(0.4)$.

a) Para una muestra de tamaño 10, rechazamos H_0 si y sólo si X pertenece al conjunto $\{8, 9, 10\}$. Es decir, la región crítica es $\{8, 9, 10\}$. Hallar para este contraste. Hallar la potencia de este contraste. Esto es, hallar la probabilidad de que $X = 8$, $X = 9$ o $X = 10$ si $p = 0.6$. Con una muestra tan pequeña, ¿es posible detectar un porcentaje de efectividad del 0.6, manteniendo un nivel de significación de aproximadamente 0.05?

b) Para una muestra de tamaño 15, tomamos como región crítica $\{11, 12, 13, 14, 15\}$. ¿Cuál es el valor de α para esta prueba? ¿Cuál es la potencia si $p = 0.6$? ¿Es esta muestra suficientemente grande para poder detectar un p de 0.6, manteniendo un nivel de aproximadamente 0.05?

c) Si n se aumenta a 20, ¿qué región crítica hay que tomar para obtener un nivel de 0.05 aproximadamente? ¿Cuál es la potencia de esta prueba si $p = 0.6$?

d) Debería ser obvio que para distinguir entre $p = 0.5$ y $p = 0.6$ con $\alpha = 0.05$, n ha de ser mayor que 20. Utilizar la aproximación normal a la distribución binomial para probar que para una muestra de tamaño 100, $P[X > 59] = 0.05$. Probar que cuando $p = 0.6$, la potencia de esta prueba es aproximadamente 0.5. Es decir, hallar $\boldsymbol{P[X \geq 59]}$ si X es binomial con $n = 100$ y $p = 0.6$. (Utilizar la aproximación normal a la distribución binomial, explicada en la Sección 13.9, para calcular P.)

APÉNDICE 

Notación sumatoria y reglas para la esperanza matemática y la varianza

NOTACIÓN SUMATORIA

En la mayor parte de los procesos estadísticos, es necesario manipular conjuntos de observaciones numéricas. Con el fin de facilitar la simplificación de las operaciones, se ha desarrollado una notación abreviada. Para designar adición, se emplea la letra griega sigma (Σ). A continuación se detalla el uso de esta notación.

Ejemplo A.1. Consideremos el siguiente conjunto de observaciones:

$$x_1 = 4 \qquad x_3 = 2 \qquad x_5 = -3$$
$$x_2 = 1 \qquad x_4 = 5$$

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 \qquad \text{(sumar las } x\text{)}$$
$$= 4 + 1 + 2 + 5 + (-3) = 9$$

$$\sum_{i=1}^{5} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 \qquad \text{(sumar los cuadrados de las } x\text{)}$$
$$= 4^2 + 1^2 + 2^2 + 5^2 + (-3)^2$$
$$= 16 + 1 + 4 + 25 + 9 = 55$$

$$\sum_{i=1}^{5} 2x_i = 2x_1 + 2x_2 + 2x_3 + 2x_4 + 2x_5 \qquad \text{(multiplicar cada } x \text{ por 2 y después sumar)}$$
$$= 2(4) + 2(1) + 2(2) + 2(5) + 2(-3)$$
$$= 8 + 2 + 4 + 10 + (-6) = 18$$

$$\sum_{i=1}^{5} (x_i - 1) = (4 - 1) + (1 - 1) + (2 - 1) \qquad \text{(restar 1 de cada observación y después sumar)}$$
$$+ (5 - 1) + (-3 - 1)$$
$$= 3 + 0 + 1 + 4 - 4 = 4$$

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 4 + 1 + 2 = 7$$

(sumar las tres primeras observaciones)

$$\sum_{i=3}^{5} x_i^3 = x_3^3 + x_4^3 + x_5^3 = 2^3 + 5^3 + (-3)^3 = 8 + 125 - 27 = 106$$

(sumar los cubos de las tre últimas observaciones)

En numerosas aplicaciones estadísticas, se utilizan todas las observaciones en los cálculos Cuando esto ocurre, se omiten tanto los subíndices como los límites del signo sumatorio. En este caso, escribiremos $\sum x = 9$, $\sum x^2 = 55$, $\sum 2x = 18$, y $\sum (x - 1) = 4$. En los dos último cálculos se podrían no haber suprimido los límites del sumatorio.

Práctica 1. Consideremos el conjunto de observaciones:

$$y_1 = 3 \quad y_3 = 1 \quad y_5 = 2$$
$$y_2 = 6 \quad y_4 = -1 \quad y_6 = -4$$

Calcular $\sum y$, $\sum y^2$, $\sum 3y$, $\sum (y + 2)$, $\sum_{i=1}^{3} y_i$, $\sum_{i=3}^{6} y_i^2$, $\sum_{i=2}^{5} 2y_i$. (Las soluciones se dan al final de este apéndice.)

Reglas para el sumatorio. Las siguientes reglas permiten simplificar expresiones complejas y calcularlas rápidamente:

Regla 1. $\sum_{i=1}^{n} c = nc$ (c es cualquier número real)

Regla 2. $\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$ (las constantes pueden sacarse como factores fuera del sumatorio)

Regla 3. $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$ (las sumas pueden dividirse y calcularse por separado)

Ejemplo A.2. Consideremos los dos conjuntos de datos:

$$x_1 = 2 \quad y_1 = 2$$
$$x_2 = 3 \quad y_2 = -1$$
$$x_3 = 5 \quad y_3 = 4$$
$$x_4 = 2 \quad y_4 = 6$$
$$\sum_{i=1}^{4} x_i = 12 \quad \sum_{i=1}^{4} y_i = 11$$

Consideremos la expresión

$$\sum_{i=1}^{4} (2x_i - 3y_i + 3)$$

Para simplificar esta expresión, pueden utilizarse las reglas para el sumatorio, de la forma siguiente:

$$\sum_{i=1}^{4}(2x_i - 3y_i + 3) = \sum_{i=1}^{4} 2x_i + \sum_{i=1}^{4}(-3)y_i + \sum_{i=1}^{4} 3 \qquad \text{(regla 3)}$$

$$= 2\sum_{i=1}^{4} x_i - 3\sum_{i=1}^{4} y_i + \sum_{i=1}^{4} 3 \qquad \text{(regla 2)}$$

$$= 2(12) - 3(11) + 4(3) \qquad \text{(regla 1 y sutitución)}$$

$$= 24 - 33 + 12 = 3$$

Práctica 2. Consideremos los dos conjuntos de datos:

$$x_1 = 1 \qquad y_1 = -2$$
$$x_2 = 4 \qquad y_2 = 4$$
$$x_3 = -3 \qquad y_3 = 5$$

Calcular $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum(2x + y)$, $\sum(3x - 2y)$, $\sum(3x + 2y - 1)$ y $\sum(x + 3y + 4)$.

REGLAS PARA LA ESPERANZA MATEMÁTICA Y LA VARIANZA

Hay tres reglas para la esperanza matemática y otras tres para la varianza que son útiles en algunas demostraciones de resultados estadísticos expuestos en el texto. Estas reglas se dan e ilustran aquí. Nuestro primer cometido será demostrar que $E[\bar{X}] = \mu$ y que $\text{Var } \bar{X} = \sigma^2/n$

Reglas para la esperanza. Sean X e Y variables aleatorias, y c un número real cualquiera.

1. $E[c] = c$ esperanza matemática de cualquier constante es esa constante.)
2. $E[cX] = cE[X]$ s constantes pueden extraerse del argumento.)
3. $E[X + Y] = E[X] + E[Y]$ (La esperanza matemática de una suma es igual a la suma de las esperanzas.)

Ejemplo A.3. Sean X e Y variables aleatorias tal que $E[X] = 3$ y $E[Y] = -2$. Entonces:

$$E[X + Y] = E[X] + E[Y] = 3 + (-2) = 1$$
$$E[2X + Y] = E[2X] + E[Y]$$
$$= 2E[X] + E[Y]$$
$$= 2(3) + (-2) = 4$$
$$E[X - 3Y + 1] = E[X] + E[-3Y] + E[1]$$
$$= E[X] + (-3)E[Y] + E[1]$$
$$= 3 + (-3)(-2) + 1$$
$$= 10$$

Reglas para la varianza. Sean X e Y variables aleatorias, y c un número real cualquiera. Entonces:

1. Var c = 0.
2. Var cX = c^2 Var X.
3. Si X e Y son independientes, entonces Var$(X + Y)$ = Var X + Var Y. (Dos variables son independientes si el valor que una toma no tiene influencia en el valor que toma la otra.)

Ejemplo A.4. Sean X e Y variables aleatorias independientes, con $\mu_X = 2$, $\mu_Y = 6$, $\sigma_X^2 = 9$ y $\sigma_Y^2 = 3$. Entonces:

$$\begin{aligned} E[2X - 3Y - 6] &= 2E[X] - 3E[Y] - 6 \\ &= 2(2) - 3(6) - 6 \\ &= -20 \end{aligned}$$

$$\begin{aligned} \text{Var}[2X - 3Y - 6] &= \text{Var } 2X + \text{Var}[-3Y] + \text{Var}[-6] \\ &= 4 \text{ Var } X + 9 \text{ Var } Y + 0 \\ &= 4(9) + 9(3) \\ &= 63 \end{aligned}$$

Decir que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria simple de una distribución X, significa que cada una de ellas tiene exactamente la misma media y la misma varianza que X, así como la misma forma. Por ejemplo, si X se distribuye normalmente con media 10 y val rianza 16, cada variable $X_1, X_2, \ldots, X_n$ se distribuye también normalmente con media 10 y varianza 16. Además, el término *muestra aleatoria* implica que estas variables aleatorias son independientes. Teniendo en cuenta estas ideas, es posible utilizar las reglas para la esperanza matemática y la varianza para verificar las afirmaciones respecto a $\bar{X}$ hechas en la Sección 6.2.

Ejemplo A.5. Supongamos que $\bar{X}$ es la media muestral de una muestra aleatoria de tamaño n, extraída de una distribución con media μ y varianza σ^2. Ya que $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$,

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$$

Por la regla 2 para la esperanza matemática, la constante \/n puede sacarse fuera obteniéndose

$$E[\bar{X}] = \left(\frac{1}{n}\right) E[X_1 + X_2 + \cdots + X_n]$$

La regla 3 para la esperanza matemática permite reescribir estas expresiones como

$$E[\bar{X}] = \left(\frac{1}{n}\right) \{E[X_1] + E[X_2] + \cdots + E[X_n]\}$$

Sin embargo, $E[X]$ y μ son símbolos intercambiables para la esperanza matemática de X y ya que cada una de las variables $X_1, X_2, \ldots, X_n$ tiene la misma media que X, pueden sustituirse $E[X_1], E[X_2], \ldots, E[X_n]$ por μ. Podemos concluir que

$$\begin{aligned} E[\bar{X}] &= \frac{1}{n} \underbrace{(\mu + \mu + \cdots + \mu)}_{n \text{ términos}} \\ &= \frac{1}{n} (n\mu) \\ &= \mu \end{aligned}$$

como se afirmó en la Sección 6.2. Consideremos ahora Var $\bar{X}$.

$$\text{Var } \bar{X} = \text{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right]$$

$$= \left(\frac{1}{n}\right)^2 \text{Var}(X_1 + X_2 + \cdots + X_n)$$

$$= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \cdots + \sigma^2)$$

$$= \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

Obsérvese que, puesto que σ^2 es constante, cuando n crece de tamaño, σ^2/n se hace más pequeño. Así, para n grande, Var $\bar{X}$ es bastante pequeña. Esto significa que para n grande, la mayoría de las medias muestrales estarán cerca de la media poblacional μ, como se pretende.

Práctica 3. Sean X e Y independientes de forma que $\mu_X = 2$, $\mu_Y = 6$, $\sigma_X^2 = 9$, $\sigma_Y^2 = 16$. Hallar los valores numéricos de los apartados *a* al *f*.

a) σ_X, σ_Y
b) $E[X^2]$, $E[Y^2]$
c) $E[X + 2Y]$, $\text{Var}[X + 2Y]$
d) $E[3X - 2Y - 2]$, $\text{Var}[3X - 2Y - 2]$
e) $E[(X - 2)/3]$, $\text{Var}[(X - 2)/3]$
f) $E[(Y - 6)/4]$, $\text{Var}[(Y - 6)/4]$
g) Los resultados de los apartados *e* y/no coinciden. Intente generalizar la pauta observada en ellos.

Práctica 4. Si se suma 5 a cada valor de una variable aleatoria X para formar otra nueva variable Y, ¿cuál es la relación entre la media de X y la media de Y? ¿Y entre la varianza de X y de F? ¿Y entre la desviación típica de X y de F?

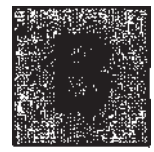
Práctica 5. Si se multiplica cada valor de X por 10 para formar una variable Y, ¿cuál es la relación entre la media de X y la media de Y? ¿Y entre la varianza de X y de Y? ¿Y entre la desviación típica de X y de Y?

Práctica 6. Utilice las reglas de la esperanza matemática y las de la varianza para probar el Teorema 5.3.1. Supóngase que $(X - \mu)/o$ es normal.

Soluciones de los problemas de prácticas

1. 7, 67, 21, 19, 10, 22, 16
2. 2,7,26,45,11,-8,17,35

APÉNDICE



Tablas estadísticas

Tabla I. Distribución binomial acumulada

$$F_X(t) = P[X \le t] = \sum_{x \le t} \binom{n}{x} p^x (1-p)^{n-x}$$

		P										
n	***t***	0.1	**0.2**	**0.25**	0.3	0.4	**0.5**	0.6	**0.7**	**0.75**	0.8	0.9
5	0	0.5905	0.3277	0.2373	0.1681	0.0778	0.0312	0.0102	0.0024	0.0010	0.0003	0.0000
	1	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0156	0.0067	0.0005
	2	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.1035	0.0579	0.0086
	3	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.3672	0.2627	0.0815
	4	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.7627	0.6723	0.4095
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	0.5314	0.2621	0.1780	0.1176	0.0467	0.0156	0.0041	0.0007	0.0002	0.0001	0.0000
	1	0.8857	0.6554	0.5339	0.4202	0.2333	0.1094	0.0410	0.0109	0.0046	0.0016	0.0001
	2	0.9841	0.9011	0.8306	0.7443	0.5443	0.3437	0.1792	0.0705	0.0376	0.0170	0.0013
	3	0.9987	0.9830	0.9624	0.9295	0.8208	0.6562	0.4557	0.2557	0.1694	0.0989	0.0159
	4	0.9999	0.9984	0.9954	0.9891	0.9590	0.8906	0.7667	0.5798	0.4661	0.3446	0.1143
	5	1.0000	0.9999	0.9998	0.9993	0.9959	0.9844	0.9533	0.8824	0.8220	0.7379	0.4686
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	0.4783	0.2097	0.1335	0.0824	0.0280	0.0078	0.0016	0.0002	0.0001	0.0000	0.0000
	1	0.8503	0.5767	0.4449	0.3294	0.1586	0.0625	0.0188	0.0038	0.0013	0.0004	0.0000
	2	0.9743	0.8520	0.7564	0.6471	0.4199	0.2266	0.0963	0.0288	0.0129	0.0047	0.0002
	3	0.9973	0.9667	0.9294	0.8740	0.7102	0.5000	0.2898	0.1260	0.0706	0.0333	0.0027
	4	0.9998	0.9953	0.9871	0.9712	0.9037	0.7734	0.5801	0.3529	0.2436	0.1480	0.0257
	5	1.0000	0.9996	0.9987	0.9962	0.9812	0.9375	0.8414	0.6706	0.5551	0.4233	0.1497
	6	1.0000	1.0000	0.9999	0.9998	0.9984	0.9922	0.9720	0.9176	0.8665	0.7903	0.5217
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	0	0.4305	0.1678	0.1001	0.0576	0.0168	0.0039	0.0007	0.0001	0.0000	0.0000	0.0000
	1	0.8131	0.5033	0.3671	0.2553	0.1064	0.0352	0.0085	0.0013	0.0004	0.0001	0.0000
	2	0.9619	0.7969	0.6785	0.5518	0.3154	0.1445	0.0498	0.0113	0.0042	0.0012	0.0000
	3	0.9950	0.9437	0.8862	0.8059	0.5941	0.3633	0.1737	0.0580	0.0273	0.0104	0.0004
	4	0.9996	0.9896	0.9727	0.9420	0.8263	0.6367	0.4059	0.1941	0.1138	0.0563	0.0050
	5	1.0000	0.9988	0.9958	0.9887	0.9502	0.8555	0.6846	0.4482	0.3215	0.2031	0.0381

Tabla I. Distribución binomial acumulada *(continuación)*

		P										
n	***t***	0.1	**0.2**	**0.25**	***03***	0.4	**0.5**	0.6	**0.7**	**0.75**	0.8	0.9
	6	1.0000	0.9999	0.9996	0.9987	0.9915	0.9648	0.8936	0.7447	0.6329	0.4967	0.1869
	7	1.0000	1.0000	1.0000	0.9999	0.9993	0.9961	0.9832	0.9424	0.8999	0.8322	0.5695
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	0.3874	0.1342	0.0751	0.0404	0.0101	0.0020	0.0003	0.0000	0.0000	0.0000	0.0000
	1	0.7748	0.4362	0.3003	0.1960	0.0705	0.0195	0.0038	0.0004	0.0001	0.0000	0.0000
	2	0.9470	0.7382	0.6007	0.4628	0.2318	0.0898	0.0250	0.0043	0.0013	0.0003	0.0000
	3	0.9917	0.9144	0.8343	0.7297	0.4826	0.2539	0.0994	0.0253	0.0100	0.0031	0.0001
	4	0.9991	0.9804	0.9511	0.9012	0.7334	0.5000	0.2666	0.0988	0.0489	0.0196	0.0009
	5	0.9999	0.9969	0.9900	0.9747	0.9006	0.7461	0.5174	0.2703	0.1657	0.0856	0.0083
	6	1.0000	0.9997	0.9987	0.9957	0.9750	0.9102	0.7682	0.5372	0.3993	0.2618	0.0530
	7	1.0000	1.0000	0.9999	0.9996	0.9962	0.9805	0.9295	0.8040	0.6997	0.5638	0.2252
	8	1.0000	1.0000	1.0000	1.0000	0.9997	0.9980	0.9899	0.9596	0.9249	0.8658	0.6126
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000	0.0000
	2	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0004	0.0001	0.0000
	3	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0035	0.0009	0.0000
	4	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0473	0.0197	0.0064	0.0001
	5	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0781	0.0328	0.0016
	6	1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.2241	0.1209	0.0128
	7	1.0000	0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.4744	0.3222	0.0702
	8	1.0000	1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.7560	0.6242	0.2639
	9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9990	0.9940	0.9718	0.9437	0.8926	0.6513
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0	0.3138	0.0859	0.0422	0.0198	0.0036	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.6974	0.3221	0.1971	0.1130	0.0302	0.0059	0.0007	0.0000	0.0000	0.0000	0.0000
	2	0.9104	0.6174	0.4552	0.3127	0.1189	0.0327	0.0059	0.0006	0.0001	0.0000	0.0000
	3	0.9815	0.8389	0.7133	0.5696	0.2963	0.1133	0.0293	0.0043	0.0012	0.0002	0.0000
	4	0.9972	0.9496	0.8854	0.7897	0.5328	0.2744	0.0994	0.0216	0.0076	0.0020	0.0000
	5	0.9997	0.9883	0.9657	0.9218	0.7535	0.5000	0.2465	0.0782	0.0343	0.0117	0.0003
	6	1.0000	0.9980	0.9924	0.9784	0.9006	0.7256	0.4672	0.2103	0.1146	0.0504	0.0028
	7	1.0000	0.9998	0.9988	0.9957	0.9707	0.8867	0.7037	0.4304	0.2867	0.1611	0.0185
	8	1.0000	1.0000	0.9999	0.9994	0.9941	0.9673	0.8811	0.6873	0.5448	0.3826	0.0896
	9	1.0000	1.0000	1.0000	1.0000	0.9993	0.9941	0.9698	0.8870	0.8029	0.6779	0.3026
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9964	0.9802	0.9578	0.9141	0.6862
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
12	0	0.2824	0.0687	0.0317	0.0138	0.0022	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.6590	0.2749	0.1584	0.0850	0.0196	0.0032	0.0003	0.0000	0.0000	0.0000	0.0000
	2	0.8891	0.5583	0.3907	0.2528	0.0834	0.0193	0.0028	0.0002	0.0000	0.0000	0.0000
	3	0.9744	0.7946	0.6488	0.4925	0.2253	0.0730	0.0153	0.0017	0.0004	0.0001	0.0000
	4	0.9957	0.9274	0.8424	0.7237	0.4382	0.1938	0.0573	0.0095	0.0028	0.0006	0.0000
	5	0.9995	0.9806	0.9456	0.8822	0.6652	0.3872	0.1582	0.0386	0.0143	0.0039	0.0001
	6	0.9999	0.9961	0.9857	0.9614	0.8418	0.6128	0.3348	0.1178	0.0544	0.0194	0.0005
	7	1.0000	0.9994	0.9972	0.9905	0.9427	0.8062	0.5618	0.2763	0.1576	0.0726	0.0043
	8	1.0000	0.9999	0.9996	0.9983	0.9847	0.9270	0.7747	0.5075	0.3512	0.2054	0.0256
	9	1.0000	1.0000	1.0000	0.9998	0.9972	0.9807	0.9166	0.7472	0.6093	0.4417	0.1109
	10	1.0000	1.0000	1.0000	1.0000	0.9997	0.9968	0.9804	0.9150	0.8416	0.7251	0.3410

Tabla I. Distribución binomial acumulada *(continuación)*

		P									
n	***t***	0.1	**0.2**	**0.25**	0.3	0.4	**0.5**	0.6	**0.7**	**0.75**	0.8	**0.9**
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9978	0.9862	0.9683	0.9313	0.7176
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	0	0.2542	0.0550	0.0238	0.0097	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.6213	0.2336	0.1267	0.0637	0.0126	0.0017	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.8661	0.5017	0.3326	0.2025	0.0579	0.0112	0.0013	0.0001	0.0000	0.0000	0.0000
	3	0.9658	0.7473	0.5843	0.4206	0.1686	0.0461	0.0078	0.0007	0.0001	0.0000	0.0000
	4	0.9935	0.9009	0.7940	0.6543	0.3530	0.1334	0.0321	0.0040	0.0010	0.0002	0.0000
	5	0.9991	0.9700	0.9198	0.8346	0.5744	0.2905	0.0977	0.0182	0.0056	0.0012	0.0000
	6	0.9999	0.9930	0.9757	0.9376	0.7712	0.5000	0.2288	0.0624	0.0243	0.0070	0.0001
	7	1.0000	0.9988	0.9944	0.9818	0.9023	0.7095	0.4256	0.1654	0.0802	0.0300	0.0009
	8	1.0000	0.9998	0.9990	0.9960	0.9679	0.8666	0.6470	0.3457	0.2060	0.0991	0.0065
	9	1.0000	1.0000	0.9999	0.9993	0.9922	0.9539	0.8314	0.5794	0.4157	0.2527	0.0342
	10	1.0000	1.0000	1.0000	0.9999	0.9987	0.9888	0.9421	0.7975	0.6674	0.4983	0.1339
	11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9983	0.9874	0.9363	0.8733	0.7664	0.3787
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9987	0.9903	0.9762	0.9450	0.7458
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0	0.2288	0.0440	0.0178	0.0068	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5846	0.1979	0.1010	0.0475	0.0081	0.0009	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.8416	0.4481	0.2811	0.1608	0.0398	0.0065	0.0006	0.0000	0.0000	0.0000	0.0000
	3	0.9559	0.6982	0.5213	0.3552	0.1243	0.0287	0.0039	0.0002	0.0000	0.0000	0.0000
	4	0.9908	0.8702	0.7415	0.5842	0.2793	0.0898	0.0175	0.0017	0.0003	0.0000	0.0000
	5	0.9985	0.9561	0.8883	0.7805	0.4859	0.2120	0.0583	0.0083	0.0022	0.0004	0.0000
	6	0.9998	0.9884	0.9617	0.9067	0.6925	0.3953	0.1501	0.0315	0.0103	0.0024	0.0000
	7	1.0000	0.9976	0.9897	0.9685	0.8499	0.6047	0.3075	0.0933	0.0383	0.0116	0.0002
	8	1.0000	0.9996	0.9978	0.9917	0.9417	0.7880	0.5141	0.2195	0.1117	0.0439	0.0015
	9	1.0000	1.0000	0.9997	0.9983	0.9825	0.9102	0.7207	0.4158	0.2585	0.1298	0.0092
	10	1.0000	1.0000	1.0000	0.9998	0.9961	0.9713	0.8757	0.6448	0.4787	0.3018	0.0441
	11	1.0000	1.0000	1.0000	1.0000	0.9994	0.9935	0.9602	0.8392	0.7189	0.5519	0.1584
	12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9991	0.9919	0.9525	0.8990	0.8021	0.4154
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9932	0.9822	0.9560	0.7712
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000	0.0000	0.0000	0.0000
	3	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000	0.0000	0.0000
	4	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0001	0.0000	0.0000
	5	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0008	0.0001	0.0000
	6	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0042	0.0008	0.0000
	7	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0173	0.0042	0.0000
	8	1.0000	0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0566	0.0181	0.0003
	9	1.0000	0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.1484	0.0611	0.0022
	10	1.0000	1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.3135	0.1642	0.0127
	11	1.0000	1.0000	1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.5387	0.3518	0.0556
	12	1.0000	1.0000	1.0000	1.0000	0.9997	0.9963	0.9729	0.8732	0.7639	0.6020	0.1841
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9948	0.9647	0.9198	0.8329	0.4510
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9866	0.9648	0.7941
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla 1. Distribución binomial acumulada *(continuación)*

		P										
n	***t***	0.1	**0.2**	**0.25**	0 3	0.4	**0.5**	0.6	**0.7**	**0.75**	0.8	0.9
16	0	0.1853	0.0281	0.0100	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.5147	0.1407	0.0635	0.0261	0.0033	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.7892	0.3518	0.1971	0.0994	0.0183	0.0021	0.0001	0.0000	0.0000	0.0000	0.0000
	3	0.9316	0.5981	0.4050	0.2459	0.0651	0.0106	0.0009	0.0000	0.0000	0.0000	0.0000
	4	0.9830	0.7982	0.6302	0.4499	0.1666	0.0384	0.0049	0.0003	0.0000	0.0000	0.0000
	5	0.9967	0.9183	0.8103	0.6598	0.3288	0.1051	0.0191	0.0016	0.0003	0.0000	0.0000
	6	0.9995	0.9733	0.9204	0.8247	0.5272	0.2272	0.0583	0.0071	0.0016	0.0002	0.0000
	7	0.9999	0.9930	0.9729	0.9256	0.7161	0.4018	0.1423	0.0257	0.0075	0.0015	0.0000
	8	1.0000	0.9985	0.9925	0.9743	0.8577	0.5982	0.2839	0.0744	0.0271	0.0070	0.0001
	9	1.0000	0.9998	0.9984	0.9929	0.9417	0.7728	0.4728	0.1753	0.0796	0.0267	0.0005
	10	1.0000	1.0000	0.9997	0.9984	0.9809	0.8949	0.6712	0.3402	0.1897	0.0817	0.0033
	11	1.0000	1.0000	1.0000	0.9997	0.9951	0.9616	0.8334	0.5501	0.3698	0.2018	0.0170
	12	1.0000	1.0000	1.0000	1.0000	0.9991	0.9894	0.9349	0.7541	0.5950	0.4019	0.0684
	13	1.0000	1.0000	1.0000	1.0000	0.9999	0.9979	0.9817	0.9006	0.8029	0.6482	0.2108
	14	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9967	0.9739	0.9365	0.8593	0.4853
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9967	0.9900	0.9719	0.8147
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17	0	0.1668	0.0225	0.0075	0.0023	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.4818	0.1182	0.0501	0.0193	0.0021	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.7618	0.3096	0.1637	0.0774	0.0123	0.0012	0.0001	0.0000	0.0000	0.0000	0.0000
	3	0.9174	0.5489	0.3530	0.2019	0.0464	0.0064	0.0005	0.0000	0.0000	0.0000	0.0000
	4	0.9779	0.7582	0.5739	0.3887	0.1260	0.0245	0.0025	0.0001	0.0000	0.0000	0.0000
	5	0.9953	0.8943	0.7653	0.5968	0.2639	0.0717	0.0106	0.0007	0.0001	0.0000	0.0000
	6	0.9992	0.9623	0.8929	0.7752	0.4478	0.1662	0.0348	0.0032	0.0006	0.0001	0.0000
	7 •	0.9999	0.9891	0.9598	0.8954	0.6405	0.3145	0.0919	0.0127	0.0031	0.0005	0.0000
	8	1.0000	0.9974	0.9876	0.9597	0.8011	0.5000	0.1989	0.0403	0.0124	0.0026	0.0000
	9	1.0000	0.9995	0.9969	0.9873	0.9081	0.6855	0.3595	0.1046	0.0402	0.0109	0.0001
	10	1.0000	0.9999	0.9994	0.9968	0.9652	0.8338	0.5522	0.2248	0.1071	0.0377	0.0008
	11	1.0000	1.0000	0.9999	0.9993	0.9894	0.9283	0.7361	0.4032	0.2347	0.1057	0.0047
	12	1.0000	1.0000	1.0000	0.9999	0.9975	0.9755	0.8740	0.6113	0.4261	0.2418	0.0221
	13	1.0000	1.0000	1.0000	1.0000	0.9995	0.9936	0.9536	0.7981	0.6470	0.4511	0.0826
	14	1.0000	1.0000	1.0000	1.0000	0.9999	0.9988	0.9877	0.9226	0.8363	0.6904	0.2382
	15	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9979	0.9807	0.9499	0.8818	0.5182
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9977	0.9925	0.9775	0.8332
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	0	0.1501	0.0180	0.0056	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.4503	0.0991	0.0395	0.0142	0.0013	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.7338	0.2713	0.1353	0.0600	0.0082	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.9018	0.5010	0.3057	0.1646	0.0328	0.0038	0.0002	0.0000	0.0000	0.0000	0.0000
	4	0.9718	0.7164	0.5187	0.3327	0.0942	0.0154	0.0013	0.0000	0.0000	0.0000	0.0000
	5	0.9936	0.8671	0.7175	0.5344	0.2088	0.0481	0.0058	0.0003	0.0000	0.0000	0.0000
	6	0.9988	0.9487	0.8610	0.7217	0.3743	0.1189	0.0203	0.0014	0.0002	0.0000	0.0000
	7	0.9998	0.9837	0.9431	0.8593	0.5634	0.2403	0.0576	0.0061	0.0012	0.0002	0.0000
	8	1.0000	0.9957	0.9807	0.9404	0.7368	0.4073	0.1347	0.0210	0.0054	0.0009	0.0000
	9	1.0000	0.9991	0.9946	0.9790	0.8653	0.5927	0.2632	0.0596	0.0193	0.0043	0.0000
	10	1.0000	0.9998	0.9988	0.9939	0.9424	0.7597	0.4366	0.1407	0.0569	0.0163	0.0002
	11	1.0000	1.0000	0.9998	0.9986	0.9797	0.8811	0.6257	0.2783	0.1390	0.0513	0.0012
	12	1.0000	1.0000	1.0000	0.9997	0.9942	0.9519	0.7912	0.4656	0.2825	0.1329	0.0064

Tabla I. Distribución binomial acumulada *(continuación)*

		P										
n	***t***	0.1	**0.2**	**0.25**	0 3	0.4	**0.5**	0.6	**0.7**	**0.75**	0.8	**0.9**
	13	1.0000	1.0000	1.0000	1.0000	0.9987	0.9846	0.9058	0.6673	0.4813	0.2836	0.0282
	14	1.0000	1.0000	1.0000	1.0000	0.9998	0.9962	0.9672	0.8354	0.6943	0.4990	0.0982
	15	1.0000	1.0000	1.0000	1.0000	1.0000	0.9993	0.9918	0.9400	0.8647	0.7287	0.2662
	16	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9987	0.9858	0.9605	0.9009	0.5497
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9984	0.9944	0.9820	0.8499
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
19	0	0.1351	0.0144	0.0042	0.0011	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.4203	0.0829	0.0310	0.0104	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.7054	0.2369	0.1113	0.0462	0.0055	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.8850	0.4551	0.2631	0.1332	0.0230	0.0022	0.0001	0.0000	0.0000	0.0000	0.0000
	4	0.9648	0.6733	0.4654	0.2822	0.0696	0.0096	0.0006	0.0000	0.0000	0.0000	0.0000
	5	0.9914	0.8369	0.6678	0.4739	0.1629	0.0318	0.0031	0.0001	0.0000	0.0000	0.0000
	6	0.9983	0.9324	0.8251	0.6655	0.3081	0.0835	0.0116	0.0006	0.0001	0.0000	0.0000
	7	0.9997	0.9767	0.9225	0.8180	0.4878	0.1796	0.0352	0.0028	0.0005	0.0000	0.0000
	8	1.0000	0.9933	0.9713	0.9161	0.6675	0.3238	0.0885	0.0105	0.0023	0.0003	0.0000
	9	1.0000	0.9984	0.9911	0.9674	0.8139	0.5000	0.0861	0.0326	0.0089	0.0016	0.0000
	10	1.0000	0.9997	0.9977	0.9895	0.9115	0.6762	0.3325	0.0839	0.0287	0.0067	0.0000
	11	1.0000	1.0000	0.9995	0.9972	0.9648	0.8204	0.5122	0.1820	0.0775	0.0233	0.0003
	12	1.0000	1.0000	0.9999	0.9994	0.9884	0.9165	0.6919	0.3345	0.1749	0.0676	0.0017
	13	1.0000	1.0000	1.0000	0.9999	0.9969	0.9682	0.8371	0.5261	0.3322	0.1631	0.0086
	14	1.0000	1.0000	1.0000	1.0000	0.9994	0.9904	0.9304	0.7178	0.5346	0.3267	0.0352
	15	1.0000	1.0000	1.0000	1.0000	0.9999	0.9978	0.9770	0.8668	0.7369	0.5449	0.1150
	16	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996	0.9945	0.9538	0.8887	0.7631	0.2946
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9896	0.9690	0.9171	0.5797
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9989	0.9958	0.9856	0.8649
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	0.1216	0.0115	0.0032	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003	0.0000	0.0000	0.0000	0.0000
	5	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000	0.0000	0.0000	0.0000
	6	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003	0.0000	0.0000	0.0000
	7	0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0002	0.0000	0.0000
	8	0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0009	0.0001	0.0000
	9	1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0039	0.0006	0.0000
	10	1.0000	0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0139	0.0026	0.0000
	11	1.0000	0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0409	0.0100	0.0001
	12	1.0000	1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.1018	0.0321	0.0004
	13	1.0000	1.0000	1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.2142	0.0867	0.0024
	14	1.0000	1.0000	1.0000	1.0000	0.9984	0.9793	0.8744	0.5836	0.3828	0.1958	0.0113
	15	1.0000	1.0000	1.0000	1.0000	0.9997	0.9941	0.9490	0.7625	0.5852	0.3704	0.0432
	16	1.0000	1.0000	1.0000	1.0000	1.0000	0.9987	0.9840	0.8929	0.7748	0.5886	0.1330
	17	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9964	0.9645	0.9087	0.7939	0.3231
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9924	0.9757	0.9308	0.6083
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9968	0.9885	0.8784
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

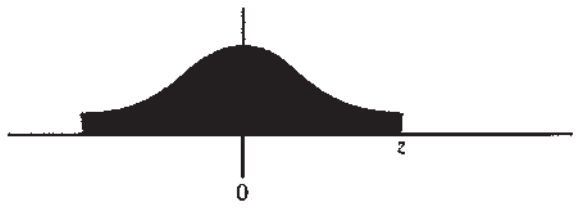
Tabla II. Función de distribución de Poisson

$$F_X(t) = P[X \leq t] = \sum_{x \leq t} e^{-\lambda s}(\lambda s)^x / x!$$

								λs								
t	*0.5*	**1.0**	2.0	**3.0**	**4.0**	**5.0**	6.0	7.0	8.0	9.0	**10.0**	**11.0**	**12.0**	**13.0**	**14.0**	**15.0**
0	0.607	0.368	0.135	0.050	0.018	0.007	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.910	0.736	0.406	0.199	0.092	0.040	0.017	0.007	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000
2	0.986	0.920	0.677	0.423	0.238	0.125	0.062	0.030	0.014	0.006	0.003	0.001	0.001	0.000	0.000	0.000
3	0.998	0.981	0.857	0.647	0.433	0.265	0.151	0.082	0.042	0.021	0.010	0.005	0.002	0.001	0.000	0.000
4	1.000	0.996	0.947	0.815	0.629	0.440	0.285	0.173	0.100	0.055	0.029	0.015	0.008	0.004	0.002	0.001
5	1.000	0.999	0.983	0.916	0.785	0.616	0.446	0.301	0.191	0.116	0.067	0.038	0.020	0.011	0.006	0.003
6	1.000	1.000	0.995	0.966	0.889	0.762	0.606	0.450	0.313	0.207	0.130	0.079	0.046	0.026	0.014	0.008
7	1.000	1.000	0.999	0.988	0.949	0.867	0.744	0.599	0.453	0.324	0.220	0.143	0.090	0.054	0.032	0.018
8	1.000	1.000	1.000	0.996	0.979	0.932	0.847	0.729	0.593	0.456	0.333	0.232	0.155	0.100	0.062	0.037
9	1.000	1.000	1.000	0.999	0.992	0.968	0.916	0.830	0.717	0.587	0.458	0.341	0.242	0.166	0.109	0.070
10	1.000	1.000	1.000	1.000	0.997	0.986	0.957	0.901	0.816	0.706	0.583	0.460	0.347	0.252	0.176	0.118
11	1.000	1.000	1.000	1.000	0.999	0.995	0.980	0.947	0.888	0.803	0.697	0.579	0.462	0.353	0.260	0.185
12	1.000	1.000	1.000	1.000	1.000	0.998	0.991	0.973	0.936	0.876	0.792	0.689	0.576	0.463	0.358	0.268
13	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.987	0.966	0.926	0.864	0.781	0.682	0.573	0.464	0.363
14	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.994	0.983	0.959	0.917	0.854	0.772	0.675	0.570	0.466
15	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.992	0.978	0.951	0.907	0.844	0.764	0.669	0.568
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.989	0.973	0.944	0.899	0.835	0.756	0.664
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.995	0.986	0.968	0.937	0.890	0.827	0.749
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.993	0.982	0.963	0.930	0.883	0.819
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.997	0.991	0.979	0.957	0.923	0.875
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.995	0.988	0.975	0.952	0.917
21											0.999	0.998	0.994	0.986	0.971	0.947
22											1.000	0.999	0.997	0.992	0.983	0.967
23											1.000	1.000	0.999	0.996	0.991	0.981
24											1.000	1.000	0.999	0.998	0.995	0.989
25											1.000	1.000	1.000	0.999	0.997	0.994
26											1.000	1.000	1.000	1.000	0.999	0.997
27											1.000	1.000	1.000	1.000	0.999	0.998
28											1.000	1.000	1.000	1.000	1.000	0.999
29											1.000	1.000	1.000	1.000	1.000	1.000

Tabla III. Distribución acumulada: normal estándar

$$F_Z(z) = P[Z \leq z]$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-.1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1921	0.1894	0.1867

-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Tabla IV. Números aleatorios

Une/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	43342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953
26	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
27	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
28	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253
29	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
31	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
32	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
33	69011	65797	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
34	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547
35	09763	83473	73577	12908	30883	18317	28290	35797	05998	41688	34952	37888	38917	88050
36	91567	42595	27958	30134	04024	86385	29880	99730	55536	84855	29080	09250	79656	73211
37	17955	56349	90999	49127	20044	59931	06115	20542	18059	02008	73708	83517	36103	42791
38	46503	18584	18845	49618	02304	51038	20655	58727	28168	15475	56942	53389	20562	87338
39	92157	89634	94824	78171	84610	82834	09922	25417	44137	48413	25555	21246	35509	20468
40	14577	62765	35605	81263	39667	47358	56873	56307	61607	49518	89656	20103	77490	18062
41	98427	07523	33362	64270	01638	92477	66969	98420	04880	45585	46565	04102	46880	45709
42	34914	63976	88720	82765	34476	17032	87589	40836	32427	70002	70663	88863	77775	69348
43	70060	28277	39475	46473	23219	53416	94970	25832	69975	94884	19661	72828	00102	66794
44	53976	54914	06990	67245	68350	82948	11398	42878	80287	88267	47363	46634	06541	97809
45	76072	29515	40980	07391	58745	25774	22987	80059	39911	96189	41151	14222	60697	59583
46	90725	52210	83974	29992	65831	38857	50490	83765	55657	14361	31720	57375	56228	41546
47	64364	67412	33339	31926	14883	24413	59744	92351	97473	89286	35931	04110	23726	51900
48	08962	00358	31662	25388	61642	34072	81249	35648	56891	69352	48373	45578	78547	81788
49	95012	68379	93526	70765	10593	04542	76463	54328	02349	17247	28865	14777	62730	92277
50	15664	10493	20492	38391	91132	21999	59516	81652	27195	48223	46751	22923	32261	85653

Tabla V. Diámetro medio a la altura del pecho de un conjunto de pinos <doblolly»

Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP
1		5.86	51	14	6.97	101	29	5.17
2		6.02	52	14	5.87	102	29	6.56
3		6.27	53	14	5.75	103	29	5.69
4		6.25	54	15	6.35	104	29	6.20
5		7.76	55	15	5.59	105	29	6.65
6	2	6.09	56	15	7.19	106	29	6.73
7	2	5.23	57	16	5.39	107	30	6.73
8	2	6.97	58	16	5.05	108	30	6.88
9	2	6.04	59	17	5.78	109	30	5.90
10	3	7.65	60	17	6.34	110	31	6.33
11	3	5.51	61	17	6.10	111	31	6.24
12	4	5.15	62	17	7.25	112	31	7.39
13	4	7.07	63	18	6.63	113	31	5.66
14	4	5.80	64	18	7.50	114	31	5.85
15	4	6.17	65	18	5.79	115	31	7.46
16	4	5.92	66	18	5.99	116	32	6.87
17	5	5.62	67	19	5.86	117	32	5.45
18	5	5.66	68	19	5.37	118	33	6.13
19	5	6.80	69	19	4.92	119	33	4.88
20	6	7.03	70	19	6.24	120	33	5.92
21	6	5.88	71	19	6.14	121	33	4.46
22	6	6.62	72	20	5.70	122	34	6.07
23	6	7.01	73	20	6.54	123	34	6.25
24	6	5.19	74	20	8.29	124	34	5.09
25	7	6.15	75	21	6.55	125	35	5.74
26	7	6.51	76	21	4.97	126	35	5.83
27	8	6.01	77	21	6.29	127	35	5.22
28	8	7.82	78	21	6.67	128	35	6.59
29	8	6.47	79	21	5.49	129	36	5.70
30	8	6.54	80	22	5.24	130	36	7.29
31	8	6.13	81	22	7.11	131	37	6.66
32	9	6.66	82	22	6.26	132	37	6.18
33	9	6.92	83	23	5.33	133	37	5.02
34	9	7.21	84	23	6.19	134	37	5.23
35	9	6.62	85	24	5.77	135	37	5.77
36	10	6.96	86	24	5.60	136	38	6.11
37	10	6.71	87	24	6.20	137	38	5.74
38	10	6.47	88,	24	5.95	138	39	7.86
39	11	6.29	89	24	6.87	139	39	5.50
40	11	5.88	90	24	6.56	140	40	6.71
41	12	6.89	91	25	5.74	141	40	5.89
42	12	6.53	92	25	6.27	142	40	6.56
43	12	6.32	93	25	6.16	143	40	5.28
44	13	5.42	94	25	6.63	144	40	7.00
45	13	6.70	95	26	7.07	145	41	6.24
46	13	7.19	**%**	26	6.92	146	41	6.37
47	13	7.08	97	26	6.47	147	41	6.42
48	13	6.23	98	27	6.63	148	42	5.53
49	14	5.41	99	27	6.34	149	42	7.07
50	14	7.54	100	28	7.11	150	42	6.74

Tabla V Diámetro medio a la altura del pecho de un conjunto de pinos «loblolly» (conf.)

Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP
151	42	6.74	201	55	6.57	251	71	6.17
152	42	4.72	202	56	6.83	252	71	6.60
153	42	5.17	203	56	7.02	253	71	7.67
154	43	7.37	204	57	6.03	254	71	7.13
155	43	6.10	205	57	6.54	255	72	6.95
156	44	5.93	206	57	6.11	256	72	5.14
157	44	6.96	207	58	6.33	257	73	6.17
158	44	5.52	208	58	5.88	258	73	6.13
159	45	7.06	209	58	6.37	259	74	7.16
160	45	7.25	210	59	5.46	260	74	7.49
161	45	6.53	211	59	6.57	261	74	6.52
162	46	6.51	212	59	6.25	262	75	6.24
163	46	6.03	213	60	7.23	263	75	7.29
164	46	6.10	214	60	5.21	264	75	6.64
165	46	6.52	215	60	5.04	265	76	6.35
166	47	7.20	216	60	6.16	266	76	6.62
167	47	5.91	217	60	5.70	267	77	6.00
168	47	6.51	218	61	5.80	268	77	6.49
169	48	7.48	219	61	5.15	269	77	6.52
170	48	6.73	220	61	6.79	270	77	6.16
171	49	6.16	221	62	5.80	271	77	6.34
172	49	4.76	222	62	7.60	272	78	5.61
173	49	6.70	223	62	6.35	273	78	6.80
174	49	5.83	224	62	6.01	274	78	6.64
175	49	6.60	225	63	7.16	275	79	6.54
176	50	8.07	226	63	6.55	276	79	6.46
177	50	5.66	227	63	5.69	277	79	6.54
178	50	6.12	228	63	5.46	278	79	5.86
179	50	6.27	229	63	6.01	279	80	6.12
180	51	6.66	230	64	6.47	280	80	5.85
181	51	5.99	231	64	5.56	281	80	5.84
182	51	5.51	232	64	5.37	282	81	4.86
183	51	6.98	233	64	6.26	283	84	6.41
184	52	6.13	234	65	6.51	284	84	6.52
185	52	7.11	235	65	6.40	285	81	6.30
186	52	5.47	236	65	6.97	286	81	6.85
187	52	5.18	237	66	5.56	287	82	6.36
188	53	6.30	238	66	6.20	288	82	6.95
189	53	5.11	239	66	6.26	289	83	7.00
190	53	7.32	240	67	6.13	290	83	5.91
191	53	7.20	241	68	7.00	291	83	6.88
192	54	5.41	242	68	5.70	292	84	7.03
193	54	6.43	243	68	6.56	293	84	6.06
194	54	5.79	244	69	7.49	294	84	5.20
195	54	6.24	245	69	6.91	295	84	5.54
196	55	6.45	246	69	5.97	296	85	6.78
197	55	4.95	247	69	5.23	297	85	5.98
198	55	6.88	248	70	6.46	298	85	5.49
199	55	6.22	249	70	5.11	299	86	5.48
200	55	4.61	250	70	6.47	300	86	6.73

Tabla V. Diámetro medio a la altura del pecho de un conjunto de pinos «loblolly» *(cont.)*

Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP
301	87	5.33	351	99	6.15	401	120	6.40
302	87	6.26	352	100	7.16	402	120	5.76
303	87	4.25	353	100	5.37	403	120	5.15
304	87	6.36	354	101	7.84	404	121	7.34
305	87	5.54	355	101	5.94	405	121	5.22
306	88	6.60	356	101	6.59	406	121	6.92
307	88	5.91	357	102	7.04	407	122	6.77
308	88	6.44	358	102	7.05	408	122	6.53
309	88	6.41	359	103	5.61	409	123	6.15
310	89	6.42	360	103	5.82	410	123	5.98
311	89	6.03	361	103	5.52	411	123	7.38
312	89	6.17	362	104	6.62	412	124	4.99
313	90	6.85	363	104	5.70	413	124	5.11
314	90	5.55	364	105	6.84	414	125	6.60
315	90	5.73	365	105	7.10	415	125	6.15
316	91	6.07	366	105	5.82	416	125	6.44
317	91	4.57	367	106	6.35	417	126	6.56
318	91	5.45	368	106	6.64	418	126	5.60
319	91	5.77	369	107	6.13	419	127	6.82
320	91	5.42	370	107	7.07	420	128	6.35
321	92	7.24	371	108	5.79	421	128	6.44
322	92	5.36	372	108	6.99	422	128	5.86
323	93	6.99	373	108	5.45	423	129	6.87
324	93	6.65	374	109	6.22	424	129	8.32
325	93	6.22	375	109	7.61	425	130	5.99
326	93	7.57	376	109	6.95	426	130	6.28
327	94	5.61	377	109	6.29	427	131	7.61
328	94	6.07	378	109	6.10	428	132	6.32
329	94	7.29	379	110	6.10	429	132	6.56
330	95	4.98	380	110	7.91	430	132	8.37
331	95	5.96	381	110	6.95	431	132	7.63
332	95	6.88	382	111	7.44	432	133	5.53
333	95	6.49	383	112	5.66	433	133	6.97
334	95	6.57	384	112	6.70	434	134	6.80
335	95	6.12	385	113	6.84	435	134	6.07
336	96	5.52	386	113	6.52	436	134	5.76
337	96	6.05	387	113	6.19	437	134	6.60
338	96	5.62	388	114	5.63	438	135	6.52
339	96	5.78	389	114	4.37	439	135	6.57
340	97	6.11	390	115	6.27	440	136	6.16
341	97	5.80	391	115	8.48	441	136	6.57
342	97	6.27	392	115	5.52	442	136	7.11
343	97	6.09	393	116	6.68	443	137	5.88
344	98	5.38	394	117	6.68	444	137	5.26
345	98	6.03	395	117	5.70	445	138	6.36
346	98	6.27	396	118	6.21	446	138	5.90
347	99	6.07	397	118	6.48	447	138	5.91
348	99	6.79	398	118	6.05	448	139	6.89
349	99	7.12	399	119	6.16	449	139	5.03
350	99	6.70	400	119	5.66	450	140	6.61

Tabla V. Diámetro medio a la altura del pecho de un conjunto de pinos «loblolly» *(cont.)*

Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP	Árbol	Cuadrado	DMAP
451	140	6.79	503	160	6.50	555	179	6.30
452	140	5.88	504	160	6.40	556	180	6.42
453	141	6.89	505	160	7.29	557	180	6.70
454	141	6.09	506	160	5.89	558	181	5.42
455	142	6.12	507	160	7.08	559	181	6.90
456	142	7.06	508	161	6.59	560	181	7.20
457	142	7.19	509	161	5.92	561	182	6.97
458	143	5.78	510	161	7.48	562	182	6.24
459	143	6.50	511	162	6.70	563	183	7.05
460	143	6.55	512	162	5.95	564	183	6.26
461	143	5.71	513	163	6.38	565	184	5.91
462	144	5.55	514	163	5.65	566	184	6.27
463	144	5.97	515	164	6.15	567	184	5.60
464	145	6.23	516	164	6.83	568	185	5.95
465	146	5.72	517	164	5.30	569	185	5.21
466	146	5.90	518	165	7.89	570	186	6.38
467	146	6.26	519	166	5.41	571	186	5.42
468	147	6.28	520	166	5.73	572	186	5.86
469	147	4.46	521	166	5.94	573	186	6.11
470	148	5.81	522	167	6.45	574	187	6.45
471	148	6.22	523	167	4.99	575	187	5.89
472	149	6.25	524	168	5.80	576	188	6.51
473	149	5.58	525	168	7.39	577	188	7.31
474	149	5.69	526	169	5.92	578	189	6.46
475	150	5.17	527	169	5.89	579	189	5.21
476	150	6.11	528	170	5.49	580	190	6.64
477	150	6.16	529	170	7.11	581	191	5.35
478	151	6.14	530	171	6.47	582	191	6.35
479	151	5.60	531	171	6.26	583	191	6.05
480	152	5.34	532	171	6.76	584	192	6.30
481	152	6.19	533	172	6.62	585	193	7.29
482	152	6.53	534	172	5.64	586	194	6.53
483	152	6.10	535	172	5.67	587	194	6.29
484	153	4.81	536	172	5.89	588	195	6.21
485	153	6.28	537	173	7.26	589	195	6.21
486	154	6.56	538	173	5.36	590	195	6.49
487	154	6.81	539	174	5.49	591	196	5.68
488	154	5.73	540	174	5.40	592	196	7.53
489	154	5.62	541	174	4.90	593	197	6.05
490	155	5.62	542	175	5.96	594	198	5.47
491	155	6.60	543	175	6.69	595	198	6.78
492	155	6.39	544	175	6.94	596	199	6.61
493	156	6.82	545	175	5.75	597	199	4.73
494	156	5.99	546	176	5.63	598	200	5.41
495	157	5.80	547	176	8.02	599	200	6.21
496	157	6.12	548	177	7.83	600	200	7.01
497	157	7.18	549	177	7.60			
498	158	6.30	550	177	5.82			
499	158	5.37	551	178	4.98			
500	158	5.48	552	178	6.01			
501	159	5.92	553	178	7.08			
502	159	6.05	554	179	6.80			

DMAP	
Media	Desviación típica
6.2536	0.6949

Tabla VI. Distribución acumulada de *T*

$$F(t) = P[T \leq t]$$

	0.40	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005 (Área a la dcha.)**
v	**0.60**	**0.75**	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**	**0.999**	**0.9995 (Área a la izqda.)**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.317	636.607
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.611	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
31	0.256	0.682	1.309	1.696	2.040	2.453	2.744	3.375	3.633
32	0.255	0.682	1.309	1.694	2.037	2.449	2.738	3.365	3.622
33	0.255	0.682	1.308	1.692	2.035	2.445	2.733	3.356	3.611
34	0.255	0.682	1.307	1.691	2.032	2.441	2.728	3.348	3.601
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591

Tabla VI. Distribución acumulada de *T(cont)*

$$F(t) = P[T \leq t]$$

	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005 (Área a la dcha.)
v	0.60	0.75	0.90	0.95	0.975	0.99	0.995	0.999	0.9995 (Área a la izqda.)
36	0.255	0.681	1.306	1.688	2.028	2.434	2.719	3.333	3.582
37	0.255	0.681	1.305	1.687	2.026	2.431	2.715	3.326	3.574
38	0.255	0.681	1.304	1.686	2.024	2.429	2.712	3.319	3.566
39	0.255	0.681	1.304	1.685	2.023	2.426	2.708	3.313	3.558
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
41	0.255	0.681	1.303	1.683	2.020	2.421	2.701	3.301	3.544
42	0.255	0.680	1.302	1.682	2.018	2.418	2.698	3.296	3.538
43	0.255	0.680	1.302	1.681	2.017	2.416	2.695	3.291	3.532
44	0.255	0.680	1.301	1.680	2.015	2.414	2.692	3.286	3.526
45	0.255	0.680	1.301	1.679	2.014	2.412	2.690	3.281	3.520
46	0.255	0.680	1.300	1.679	2.013	2.410	2.687	3.277	3.515
47	0.255	0.680	1.300	1.678	2.012	2.408	2.685	3.273	3.510
48	0.255	0.680	1.299	1.677	2.011	2.407	2.682	3.269	3.505
49	0.255	0.680	1.299	1.677	2.010	2.405	2.680	3.265	3.500
50	0.255	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
51	0.255	0.679	1.298	1.675	2.008	2.402	2.676	3.258	3.492
52	0.255	0.679	1.298	1.675	2.007	2.400	2.674	3.255	3.488
53	0.255	0.679	1.298	1.674	2.006	2.399	2.672	3.251	3.484
54	0.255	0.679	1.297	1.674	2.005	2.397	2.670	3.248	3.480
55	0.255	0.679	1.297	1.673	2.004	2.396	2.668	3.245	3.476
56	0.255	0.679	1.297	1.673	2.003	2.395	2.667	3.242	3.473
57	0.255	0.679	1.297	1.672	2.002	2.394	2.665	3.239	3.470
58	0.255	0.679	1.296	.672	2.002	2.392	2.663	3.237	3.466
59	0.254	0.679	1.296	.671	2.001	2.391	2.662	3.234	3.463
60	0.254	0.679	1.296	.671	2.000	2.390	2.660	3.232	3.460
61	0.254	0.679	1.296	.670	2.000	2.389	2.659	3.229	3.457
62	0.254	0.678	1.295	.670	1.999	2.388	2.658	3.227	3.455
63	0.254	0.678	1.295	.669	1.998	2.387	2.656	3.225	3.452
64	0.254	0.678	1.295	.669	1.998	2.386	2.655	3.223	3.449
65	0.254	0.678	1.295	.669	1.997	2.385	2.654	3.221	3.447

Tabla VI. Distribución acumulada de *T(cont.)*

	0.40	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005 (Área a la dcha.)**
v	**0.60**	**0.75**	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**	**0.999**	**0.9995 (Área a la izqda.)**
66	0.254	0.678	1.295	1.668	1.997	2.384	2.652	3.218	3.444
67	0.254	0.678	1.294	1.668	1.996	2.383	2.651	3.217	3.442
68	0.254	0.678	1.294	1.668	1.995	2.382	2.650	3.215	3.440
69	0.254	0.678	1.294	1.667	1.995	2.382	2.649	3.213	3.437
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	3.211	3.435
71	0.254	0.678	1.294	1.667	1.994	2.380	2.647	3.209	3.433
72	0.254	0.678	1.293	1.666	1.993	2.379	2.646	3.207	3.431
73	0.254	0.678	1.293	1.666	1.993	2.379	2.645	3.206	3.429
74	0.254	0.678	1.293	1.666	1.993	2.378	2.644	3.204	3.427
75	0.254	0.678	1.293	1.665	1.992	2.377	2.643	3.203	3.425
76	0.254	0.678	1.293	1.665	1.992	2.376	2.642	3.201	3.423
77	0.254	0.678	1.293	1.665	1.991	2.376	2.641	3.200	3.422
78	0.254	0.678	1.292	1.665	1.991	2.375	2.640	3.198	3.420
79	0.254	0.678	1.292	1.664	1.990	2.375	2.640	3.197	3.418
80	0.254	0.678	1.292	1.664	1.990	2.374	2.639	3.195	3.416
81	0.254	0.678	1.292	1.664	1.990	2.373	2.638	3.194	3.415
82	0.254	0.677	1.292	1.664	1.989	2.373	2.637	3.193	3.413
83	0.254	0.677	1.292	1.663	1.989	2.372	2.636	3.191	3.412
84	0.254	0.677	1.292	1.663	1.989	2.372	2.636	3.190	3.410
85	0.254	0.677	1.292	1.663	1.988	2.371	2.635	3.189	3.409
86	0.254	0.677	1.291	1.663	1.988	2.371	2.634	3.188	3.407
87	0.254	0.677	1.291	1.663	1.988	2.370	2.634	3.187	3.406
88	0.254	0.677	1.291	1.662	1.987	2.369	2.633	3.186	3.405
89	0.254	0.677	1.291	1.662	1.987	2.369	2.632	3.184	3.403
90	0.254	0.677	1.291	1.662	1.987	2.369	2.632	3.183	3.402
91	0.254	0.677	1.291	1.662	1.986	2.368	2.631	3.182	3.401
92	0.254	0.677	1.291	1.662	1.986	2.368	2.630	3.181	3.400
93	0.254	0.677	1.291	1.661	1.986	2.367	2.630	3.180	3.398
94	0.254	0.677	1.291	1.661	1.986	2.367	2.629	3.179	3.397
95	0.254	0.677	1.291	1.661	1.985	2.366	2.629	3.178	3.396
96	0.254	0.677	1.290	1.661	1.985	2.366	2.628	3.177	3.395
97	0.254	0.677	1.290	1.661	1.985	2.365	2.627	3.176	3.394
98	0.254	0.677	1.290	1.661	1.984	2.365	2.627	3.176	3.393
99	0.254	0.677	1.290	1.660	1.984	2.365	2.626	3.175	3.392
100	0.254	0.677	1.290	1.660	1.984	2.364	2.626	3.174	3.391
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

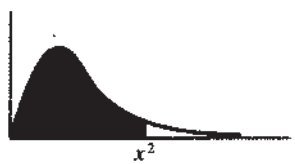
Tabla II. Tamaño muestral para el contraste de la media

Contraste de una cola / Contraste dos colas		Nivel del contraste t																				
		$\alpha = 0.005$ / $\alpha = 0.01$					$\alpha = 0.01$ / $\alpha = 0.02$					$\alpha = 0.025$ / $\alpha = 0.05$					$\alpha = 0.05$ / $\alpha = 0.1$					
$\beta =$		0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	
	0.05																					0.05
	0.10																					0.10
	0.15																				122	0.15
	0.20										139					99					70	0.20
	0.25					110					90				128	64			139	101	45	0.25
	0.30				134	78				115	63			119	90	45		122	97	71	32	0.30
	0.35			125	99	58			109	85	47		109	88	67	34		90	72	52	24	0.35
	0.40		115	97	77	45		101	85	66	37	117	84	68	51	26	101	70	55	40	19	0.40
	0.45		92	77	62	37	110	81	68	53	30	93	67	54	41	21	80	55	44	33	15	0.45
Valor de $\Delta = \frac{\mu - \mu_0}{\sigma}$	0.50	100	75	63	51	30	90	66	55	43	25	76	54	44	34	18	65	45	36	27	13	0.50
	0.55	83	63	53	42	26	75	55	46	36	21	63	45	37	28	15	54	38	30	22	11	0.55
	0.60	71	53	45	36	22	63	47	39	31	18	53	38	32	24	13	46	32	26	19	9	0.60
	0.65	61	46	39	31	20	55	41	34	27	16	46	33	27	21	12	39	28	22	17	8	0.65
	0.70	53	40	34	28	17	47	35	30	24	14	40	29	24	19	10	34	24	19	15	8	0.70
	0.75	47	36	30	25	16	42	31	27	21	13	35	26	21	16	9	30	21	17	13	7	0.75
	0.80	41	32	27	22	14	37	28	24	19	12	31	22	19	15	9	27	19	15	12	6	0.80
	0.85	37	29	24	20	13	33	25	21	17	11	28	21	17	13	8	24	17	14	11	6	0.85
	0.90	34	26	22	18	12	29	23	19	16	10	25	19	16	12	7	21	15	13	10	5	0.90
	0.95	31	24	20	17	11	27	21	18	14	9	23	17	14	11	7	19	14	11	9	5	0.95
	1.00	28	22	19	16	10	25	19	16	13	9	21	16	13	10	6	18	13	11	8	5	1.00

Valor de $\Delta = \frac{\mu - \mu_0}{\sigma}$																					
	1.1	24	19	16	14	9	21	16	14	12	8	18	13	11	9	6	15	11	9	7	1.1
	1.2	21	16	14	12	8	18	14	12	10	7	15	12	10	8	5	13	10	8	6	1.2
	1.3	18	15	13	11	8	16	13	11	9	6	14	10	9	7		11	8	7	6	1.3
	1.4	16	13	12	10	7	14	11	10	9	6	12	9	8	7		10	8	7	5	1.4
	1.5	15	12	11	9	7	13	10	9	8	6	11	8	7	6		9	7	6		1.5
	1.6	13	11	10	8	6	12	10	9	7	5	10	8	7	6		8	6	6		1.6
	1.7	12	10	9	8	6	11	9	8	7		9	7	6	5		8	6	5		1.7
	1.8	12	10	9	8	6	10	8	7	7		8	7	6			7	6			1.8
	1.9	11	9	8	7	6	10	8	7	6		8	6	6			7	5			1.9
	2.0	10	8	8	7	5	9	7	7	6		7	6	5			6				2.0
	2.1	10	8	7	7		8	7	6	6		7	6				6				2.1
	2.2	9	8	7	6		8	7	6	5		7	6				6				2.2
	2.3	9	7	7	6		8	6	6			6	5				5				2.3
	2.4	8	7	7	6		7	6	6			6									2.4
	2.5	8	7	6	6		7	6	6			6									2.5
	3.0	7	6	6	5		6	5	5			5									3.0
	3.5	6	5	5			5														3.5
	4.0	6																			4.0

Reproducido con autorización de W. H. Beyer (ed.), *CRC Handbook of Tables for Probability and Statistics,* 2ª ed., 1968, p. 282. Copyright CRC Press, Boca Ratón, Florida

Tabla VIII. Distribución acumulada de ji-cuadrado

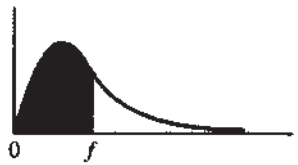


$$F(\chi^2) = P[X^2 \leq \chi^2]$$

γ \ F	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995
1	0.0000393	0.000157	0.000982	0.00393	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8

16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

Tabla IX. Distribución acumulada de *F*

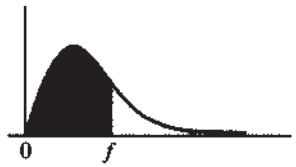


$$P[F_{\nu_1,\nu_2} \leq f] = 0.90$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	**10**	12	15	20	24	30	40	60	**120**	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.%	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	**1.66**	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	130	1.24	1.17	1.00

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90.	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Tabla IX. Distribución acumulada de *F(cont.)*



$P[F_{\gamma_1,\gamma_2} \leq f] = 0.975$

γ_2 \ γ_1	1	2	3	4	5	6	7	8	9	**10**	**12**	15	**20**	24	30	40	60	**120**	**∞**
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.%	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.%	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

$$P[F_{\nu_1,\nu_2} \leq f] = 0.99$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Tabla X. Tablas de Duncan

Rang;o mínimo significativo «estudentizado» r_p $a = 0.05$ P						Rango mínimo significativo «estudentizado» r $a = \mathbf{0.01}$ P					
r	2	3	4	5	6	*r*	*2*	3	4	5	6
1	17.97	17.97	17.97	17.97	17.97	1	90.03	90.03	90.03	90.03	9O.03
2	6.085	6.085	6.085	6.085	6.085	2	14.04	14.04	14.04	14.04	14.04
3	4.501	4.516	4.516	4.516	4.516	3	8.261	8.321	8.321	8.321	8.321
4	3.927	4.013	4.033	4.033	4.033	4	6.512	6.677	6.740	6.756	6.756
5	3.635	3.749	3.797	3.814	3.814	5	5.702	5.893	5.898	6.040	6C65
6	3.461	3.587	3.649	3.680	3.694	6	5.243	5.439	5.549	5.614	5.655
7	3.344	3.477	3.548	3.588	3.611	7	4.949	5.145	5.260	5.334	5.383
8	3.261	3.399	3.475	3.521	3.549	8	4.746	4.939	5.057	5.135	5.189
9	3.199	3.339	3.420	3.470	3.502	9	4.596	4.787	4.906	4.986	5.Ó43
10	3.151	3.293	3.376	3.430	3.465	10	4.482	4.671	4.790	4.871	4.931
11	3.113	3.256	3.342	3.397	3.435	11	4.392	4.579	4.697	4.780	4.841
12	3.082	3.225	3.313	3.370	3.410	12	4.320	4.504	4.622	4.706	4.767
13	3.055	3.200	3.289	3.348	3.389	13	4.260	4.442	4.560	4.644	4.706
14	3.033	3.178	3.268	3.329	3.372	14	4.210	4.391	4.508	4.591	4.654
15	3.014	3.160	3.250	3.312	3.356	15	4.168	4.347	4.463	4.547	4.610
16	2.998	3.144	3.235	3.298	3.343	16	4.131	4.309	4.425	4.509	4.572
17	2.984	3.130	3.222	3.285	3.331	17	4.099	4.275	4.391	4.475	4.539
18	2.971	3.118	3.210	3.274	3.321	18	4.071	4.246	4.362	4.445	4.509
19	2.960	3.107	3.199	3.264	3.311	19	4.046	4.220	4.335	4.419	4.483
20	2.950	3.097	3.190	3.255	3.303	20	4.024	4.197	4.312	4.395	4.459
24	2.919	3.066	3.160	3.226	3.276	24	3.956	4.126	4.239	4.322	4.386
30	2.888	3.035	3.131	3.199	3.250	30	3.889	4.506	4.168	4.250	4.314
40	2.858	3.006	3.102	3.171	3.224	40	3.825	3.988	4.098	4.180	4.244
60	2.829	2.976	3.073	3.143	3.198	60	3.762	3.922	4.031	4.111	4.174
120	2.800	2.947	3.045	3.116	3.172	120	3.702	3.858	3.965	4.044	4.137
00	2.772	2.918	3.017	3.089	3.146	00	3.643	3.796	3.900	3.978	4.040

Compendio de H L. Harter's «Critical Valúes for Duncan's New Múltiple Range Test,» *Biometrics,* vol. 16, n.°4(1960).Cc n autorización de Biometric Society.

Tabla XI. Contraste de rangos de signos de Wilcoxon

Una cola	Dos colas	« = 5	« = 6	« = 7	**B = 8**	n = 9	**71= 10**
P = 0.05	P = 0.10	1	2	4	6	8	11
P = 0.025	P = 0.05		1	2	4	6	8
P = 0.01	P = 0.02			0	2	3	5
P= 0.005	P = 0.01				0	2	3
Una cola	Dos colas	«=1 1	**71= 12**	«=1 3	«=1 4	«=1 5	«=1 6
p = 0.05	P = 0.10	14	17	21	26	30	36
P = 0.025	P = 0.05	11	14	17	21	25	30
P = 0.01	P = 0.02	7	10	13	16	20	24
P = 0.005	P = 0.01	5	7	10	13	16	19
Una cola	Dos colas	«=1 7	«=1 8	«=1 9	« = 20	« = 21	« = 22
P = 0.05	P = 0.10	41	47	54	60	68	75
P = 0.025	P = 0.05	35	40	46	52	59	66
P = 0.01	P = 0.02	28	33	38	43	49	56
P = 0.005	P = 0.01	23	28	32	37	43	49
Una cola	Dos colas	« = 23	« = 24	« = 25	« = 26	« = 27	« = 28
P = 0.05	P = 0.10	83	92	101	110	120	130
p = 0.025	P = 0.05	73	81	90	98	107	117
P = 0.01	P = 0.02	62	69	77	85	93	102
P = 0.005	P = 0.01	55	61	68	76	84	92
'Una cola	Dos colas	**7i = 29**	« = 30	**71=31**	« = 32	« = 33	« = 34
P = 0.05	P = 0.10	141	152	163	175	188	201
p = 0.025	P = 0.05	127	137	148	159	171	183
P = 0.01	P = 0.02	111	120	130	141	151	162
P = 0.005	P = 0.01	100	109	118	128	138	149
'Una cola	Dos colas	« = 35	« = 36	« = 37	« = 38	« = 39	
p = 0.05	P = 0.10	214	228	242	256	271	
p = 0.025	P = 0.05	195	208	222	235	250	
p = 0.01	P = 0.02	174	186	198	211	224	
P = 0.005	P = 0.01	160	171	183	195	208	
'Una cola	Dos colas	« = 40	« = 41	« = 42	« = 43	n = 44	n = 45
/> = 0.05	*P*= 0.10	287	303	319	336	353	371
P = 0.025	/>= 0.05	264	279	295	311	327	344
P = 0.01	P = 0.02	238	252	267	281	297	313
P = 0.005	P = 0.01	221	234	248	262	277	292
Una cola	Dos colas	n = 46	« = 47	« = 48	« = 49	« = 50	
0.05	P = 0.10	389	408	427	446	466	
P = 0.025	P = 0.05	361	379	397	415	434	
P = 0.01	P = 0.02	329	345	362	380	398	
P = 0.005	P = 0.01	307	323	339	356	373	

Tabla XII. Contraste de la suma de rangos de Wilcoxon

$m = 3(1)25$ y $n = m(1)m + 25$
$P = 0.05$ una cola; $P = 0.10$ dos colas

n	m = 3	m = 4	m = 5	m = 6	m = 7	m = 8	m = 9	m= 10	m= 11	m= 12	m= 13	m = 14
$n = m$	6,15	12,24	19,36	28,50	39,66	52,84	66,105	83,127	101,152	121,179	143,208	167,239
$n = m + 1$	7,17	13,27	20,40	30,54	41,71	54,90	69,111	86,134	105,159	125,187	148,216	172,248
$n = m + 2$	7,20	14,30	22,43	32,58	43,76	57,95	72,117	89,141	109,166	129,195	152,225	177,257
$n = m + 3$	8,22	15,33	24,46	33,63	46,80	60,100	75,123	93,147	112,174	134,202	157,233	182,266
n = m + 4	9,24	16,36	25,50	35,67	48,85	62,106	78,129	96,154	116,181	138,210	162,241	187,275
n = m + 5	9,27	17,39	26,54	37,71	50,90	65,111	81,135	100,160	120,188	142,218	166,250	192,284
n = m + 6	10,29	18,42	27,58	39,75	52,95	67,117	84,141	103,167	124,195	147,225	171,258	197,293
n = m + 7	11,31	19,45	29,61	41,79	54,100	70,122	87,147	107,173	128,202	151,233	176,266	203,301
n = m + 8	11,34	20,48	30,65	42,84	57,104	73,127	90,153	110,180	132,209	155,241	181,274	208,310
n = m + 9	12,36	21,51	32,68	44,88	59,109	75,133	93,159	114,186	136,216	159,249	185,283	213,319
$n = m$ + 10	13,38	22,54	33,72	46,92	61,114	78,138	96,165	117,193	139,224	164,256	190,291	218,328
n = m + 11	13,41	23,57	34,76	48,96	63,119	80,144	100,170	120,200	143,231	168,264	195,299	223,337
n = m + 12	14,43	24,60	36,79	50,100	65,124	83,149	103,176	124,206	147,238	172,272	199,308	228,346
n = n + 13	15,45	25,63	37,83	52,104	68,128	86,154	106,182	127,213	151,245	177,279	204,316	234,354
n = m + 14	15,48	26,66	39,86	53,109	70,133	88,160	109,188	131,219	155,252	181,287	209,324	239,363
n = m + 15	16,50	27,69	40,90	55,113	72,138	91,165	112,194	134,226	159,259	185,295	214,332	244,372
n = m + 16	17,52	28,72	42,93	57,117	74,143	94,170	115,200	138,232	163,266	190,302	218,341	249,381
n = m + 17	17,55	29,75	43,97	59,121	77,147	96,176	118,206	141,239	167,273	194,310	223,349	254,390
n = m+ 18	18,57	30,78	44,101	61,125	79,152	99,181	121,212	145,245	171,280	198,318	228,357	260,398
n = m+ 19	19,59	31,81	46,104	62,130	81,157	102,186	124,218	148,252	175,287	203,325	233,365	265,407
n = m + 20	19,62	32,84	47,108	64,134	83,162	104,192	127,224	152,258	178,295	207,333	237,374	270,416
n = m + 21	20,64	33,87	49,111	66,138	86,166	107,197	130,230	155,265	182,302	211,341	242,382	275,425
n = m + 22	21,66	34,90	50,115	68,142	88,171	109,203	133,236	159,271	186,309	216,348	247,390	280,434
n = m + 23	21,69	35,93	52,118	70,146	90,176	112,208	136,242	162,278	190,316	220,356	252,398	285,443
n = m + 24	22,71	37,95	53,122	72,150	92,181	115,213	139,248	166,284	194,323	224,364	257,406	291,451
n = m + 25	23,73	38,98	54,126	73,155	94,186	117,219	142,254	169,291	198,330	229,371	261,415	296.460

m= 3(1)25 y n=m(1)m +25
P = 0.05 una cola; P = 0.10 dos colas

n	m = 15	m = 16	m = 17	m = 18	m = 19	m = 20	m = 21	m = 22	m = 23	m = 24	m = 25
n = m	192,273	220,308	249,346	280,386	314,427	349,471	386,517	424,566	465,616	508,668	552,723
n = m + 1	198,282	226,318	256,356	287,397	321,439	356,484	394,530	433,579	474,630	517,683	562,738
n = m + 2	203,292	232,328	262,367	294,408	328,451	364,496	402,543	442,592	483,644	527,697	572,753
n = m + 3	209,301	238,338	268,378	301,419	336,462	372,508	410,556	450,606	492,658	536,712	582,768
n = m + 4	215,310	244,348	275,388	308,430	343,474	380,520	418,569	459,619	501,672	546,726	592,783
n = m + 5	220,320	250,358	281,399	315,441	350,486	387,533	427,581	468,632	511,685	555,741	602,798
n = m + 6	226,329	256,368	288,409	322,452	358,497	395,545	435,594	476,646	520,699	565,755	612,813
n = m + 7	231,339	262,378	294,420	329,463	365,509	403,557	443,607	485,659	529,713	574,770	622,828
n = m + 8	237,348	268,388	301,430	336,474	372,521	411,569	451,620	494,672	538,727	584,784	632,843
n = m + 9	242,358	274,398	307,441	342,486	380,532	419,581	459,633	502,686	547,741	594,798	642,858
n = m + 10	248,367	280,408	314,451	349,497	387,544	426,594	468,645	511,699	556,755	603,813	652,873
n = m + 11	254,376	286,418	320,462	356,508	394,556	434,606	476,658	520,712	565,769	613,827	662,888
n = m + 12	259,386	292,428	327,472	363,519	402,567	442,618	484,671	528,726	574,783	622,842	672,903
n = m + 13	265,395	298,438	333,483	370,530	409,579	450,630	492,684	537,739	584,796	632,856	682,918
n = m + 14	270,405	304,448	340,493	377,541	416,591	458,642	501,696	546,752	593,810	642,870	692,933
n = m + 15	276,414	310,458	346,504	384,552	424,602	465,655	509,709	554,766	602,824	651,885	702,948
n = m + 16	282,423	316,468	353,514	391,563	431,614	473,667	517,722	563,779	611,838	661,899	712,963
n = m + 17	287,433	322,478	359,525	398,574	438,626	481,679	526,734	572,792	620,852	670,914	723,977
n = m + 18	293,442	328,488	366,535	405,585	446,637	489,691	534,747	581,805	629,866	680,928	733,992
n = m + 19	299,451	334,498	372,546	412,596	453,649	487,703	542,760	589,819	639,879	690,942	743,1007
n = m + 20	304,461	340,508	379,556	419,607	461,660	505,715	550,773	598,832	648,893	699,957	753,1022
n = m + 21	310,470	347,517	385,568	426,618	468,672	512,728	559,785	607,845	657,907	709,971	763,1037
n = m + 22	315,480	353,527	392,577	433,629	475,684	520,740	567,798	615,859	666,921	718,986	773,1052
n = m + 23	321,489	359,537	398,588	439,641	483,695	528,752	575,811	624,872	675,935	728,1000	783,1067
n = m + 24	327,498	365,547	405,598	446,652	490,707	536,764	583,824	633,885	684,949	738,1014	793,1082
n = m + 25	332,508	371,557	411,609	453,663	498,718	544,776	592,836	642,898	694,962	747,1029	803,1097

Tabla XII. Contraste de la suma de rangos de Wilcoxon (cont.)

m = 3(1)25 y n = m(1)m + 25
P = 0.025 una cola; P = 0.05 dos colas

n	m = 3	m = 4	m — 5	m = 6	m = l	m = 8	m = 9	m = 10	m = 11	m = 12	m = 13	m = 14
n = m	5,16	11,25	18,37	26,52	37,68	49,87	63,108	79,131	96,157	116,184	137,214	160,246
n = m + 1	6,18	12,28	19,41	28,56	39,73	51,93	66,114	82,138	100,164	120,192	141,223	165,255
n = m + 2	6,21	12,32	20,45	29,61	41,78	54,98	68,121	85,145	103,172	124,200	146,231	170,264
n = m + 3	7,23	13,35	21,49	31,65	43,83	56,104	71,127	88,152	107,179	128,208	150,240	174,274
n = m + 4	7,26	14,38	22,53	32,70	45,88	58,110	74,133	91,159	110,187	131,217	154,249	179,283
n = m + 5	8,28	15,41	24,56	34,74	46,94	61,115	77,139	94,166	114,194	135,225	159,257	184,292
n = m + 6	8,31	16,44	25,60	36,78	48,99	63,121	79,146	97,173	118,201	139,233	163,266	189,301
n = m + 7	9,33	17,47	26,64	37,83	50,104	65,127	82,152	101,179	121,209	143,241	168,274	194,310
n = m + 8	10,35	17,51	27,68	39,87	52,109	68,132	85,158	104,186	125,216	147,249	172,283	198,320
n = m + 9	10,38	18,54	29,71	41,91	54,114	70,138	88,164	107,193	128,224	151,257	176,292	203,329
n = m + 10	11,40	19,57	30,75	42,96	56,119	72,144	90,171	110,200	132,231	155,265	181,300	208,338
n = m + 11	11,43	20,60	31,79	44,100	58,124	75,149	93,177	113,207	135,239	159,273	185,309	213,347
n = m + 12	12,45	21,63	32,83	45,105	60,129	77,155	96,183	117,213	139,246	163,281	190,317	218,356
n = m + 13	12,48	22,66	33,87	47,109	62,134	80,160	99,189	120,220	143,253	167,289	194,326	222,366
n = m + H	13,50	23,69	35,90	49,113	64,139	82,166	101,196	123,227	146,261	171,297	198,335	227,375
n = m + 15	13,53	24,72	36,94	50,118	66,144	84,172	104,202	126,234	150,268	175,305	203,343	232,384
n = m + 16	14,55	24,76	37,98	52,122	68,149	87,177	107,208	129,241	153,276	179,313	207,352	237,393
n = m + 17	14,58	25,79	38,102	53,127	70,154	89,183	110,214	132,248	157,283	183,321	212,360	242,402
n = m + 18	15,60	26,82	40,105	55,131	72,159	92,188	113,220	136,254	161,290	187,329	216,369	247,411
n = m + 19	15,63	27,85	41,109	57,135	74,164	94,194	115,227	139,261	164,298	191,337	221,377	252,420
n = m + 20	16,65	28,88	42,113	58,140	76,169	96,200	118,233	142,268	168,305	195,345	225,386	256,430
n = m + 21	16,68	29,91	43,117	60,144	78,174	99,205	121,239	145,275	171,313	199,353	229,395	261,439
n = m + 22	17,70	30,94	45,120	61,149	80,179	101,211	124,245	148,282	175,320	203,361	234,403	266,448
n = m + 23	17,73	31,97	46,124	63,153	82,184	103,217	127,251	152,288	179,327	207,369	238,412	271,457
n = m + 24	18,75	31,101	47,128	65,157	84,189	106,222	129,258	155,295	182,335	211,377	243,420	276,466
n = m + 25	18,78	32,104	48,132	66,162	86,194	108,228	132,264	158,302	186.342	216.384	247,429	281,475

$m = 3(1)25$ y $n = m(1)m + 25$
$P = 0.025$ una cola; $P = 0.05$ dos colas

n	$m = 15$	$m = 16$	$m = 17$	$m = 18$	$m = 19$	$m = 20$	$m = 21$	$m = 22$	$m = 23$	$m = 24$	$m = 25$
$n = m$	185,280	212,316	240,355	271,395	303,438	337,483	373,530	411,579	451,630	493,683	536,739
$n = m + 1$	190,290	217,327	246,366	277,407	310,450	345,495	381,543	419,593	460,644	502,698	546,754
$n = m + 2$	195,300	223,337	252,377	284,418	317,462	352,508	389,556	428,606	468,659	511,713	555,770
$n = m + 3$	201,309	229,347	258,388	290,430	324,474	359,521	397,569	436,620	477,673	520,728	565,785
$n = m + 4$	206,319	234,358	264,399	297,441	331,486	367,533	404,583	444,634	486,687	529,743	574,801
$n = m + 5$	211,329	240,368	271,409	303,453	338,498	374,546	412,596	452,648	494,702	538,758	584,816
$n = m + 6$	216,339	245,379	277,420	310,464	345,510	381,559	420,609	460,662	503,716	547,773	593,832
$n = m + 7$	221,349	251,389	283,431	316,476	351,523	389,571	428,622	469,675	512,730	556,788	603,847
$n = m + 8$	227,358	257,399	289,442	323,487	358,535	396,584	436,635	477,689	520,745	565,803	612,863
$n = m + 9$	232,368	262,410	295,453	329,499	365,547	403,597	443,649	485,703	529,759	575,817	622,878
$n = m + 10$	237,378	268,420	301,464	336,510	372,559	411,609	451,662	493,717	538,773	584,832	632,893
$n = m + 11$	242,388	274,430	307,475	342,522	379,571	418,622	459,675	502,730	546,788	593,847	641,909
$n = m + 12$	248,397	279,441	313,486	349,533	386,583	426,634	467,688	510,744	555,802	602,862	651,924
$n = m + 13$	253,407	285,451	319,497	355,545	393,595	433,647	475,701	518,758	564,816	611,877	660,940
$n = m + 14$	258,417	291,461	325,508	362,556	400,607	440,660	482,715	526,772	572,831	620,892	670,955
$n = m + 15$	263,427	296,472	331,519	368,568	407,619	448,672	490,728	535,785	581,845	629,907	679,971
$n = m + 16$	269,436	302,482	338,529	375,579	414,631	455,685	498,741	543,799	590,859	638,922	689,986
$n = m + 17$	274,446	308,492	344,540	381,591	421,643	463,697	506,754	551,813	599,873	648,936	699,1001
$n = m + 18$	279,456	314,502	350,551	388,602	428,655	470,710	514,767	560,826	607,888	657,951	708,1017
$n = m + 19$	284,466	319,513	356,562	395,613	435,667	477,723	522,780	568,840	616,902	666,966	718,1032
$n = m + 20$	290,475	325,523	362,573	401,625	442,679	485,735	530,793	576,854	625,916	675,981	727,1048
$n = m + 21$	295,485	331,533	368,584	408,636	449,691	492,748	537,807	584,868	633,931	684,996	737,1063
$n = m + 22$	300,495	336,544	374,595	414,648	456,703	500,760	545,820	593,881	642,945	693,1011	747,1078
$n = m + 23$	306,504	342,554	380,606	421,659	463,715	507,773	553,833	601,895	651,959	703,1025	756,1094
$n = m + 24$	311,514	348,564	387,616	427,671	470,727	515,785	561,846	609,909	660,973	712,1040	766,1109
$n = m + 25$	316,524	353,575	393,627	434,682	477,739	522,798	569,859	618,922	668,988	721,1055	775,1125

Reproducido con autorización de W. H. Beyer (ed.), *CRC Handbook ofTablesfor Probability and Statistics,* 2.ª ed., 1968, pp. 410-411. Copyright CRC Press, Boca Ratón, Florida

Tabla XIII. Datos de tensión arterial

Paciente	Sexo	Tensión sistólica	Tensión diastólica	Paciente	Sexo	Tensión sistólica	Tensión diastólica
1	M	112	80	61	F	142	80
2	F	112	80	62	F	120	80
3	M	130	76	63	M	136	86
4	M	157	86	64	M	142	80
5	F	152	80	65	F	95	80
6	F	130	80	66	M	110	80
7	M	140	90	67	F	118	80
8	M	120	74	68	M	97	75
9	F	100	82	69	M	120	84
10	M	118	100	70	F	110	60
11	F	120	90	71	M	110	80
12	M	122	72	72	F	118	88
13	F	130	80	73	M	117	76
14	M	96	65	74	M	100	78
15	F	90	56	75	F	152	88
16	M	102	70	76	M	112	80
17	F	115	80	77	F	160	88
18	M	130	80	78	M	110	75
19	M	136	94	79	M	128	85
20	F	130	80	80	F	118	90
21	M	122	80	81	M	140	100
22	M	118	84	82	F	90	58
23	M	108	76	83	M	110	70
24	M	110	86	84	F	102	60
25	F	140	88	85	M	140	90
26	M	130	80	86	F	110	80
27	F	120	80	87	M	116	74
28	M	120	76	88	M	118	81
29	M	125	80	89	F	140	85
30	F	130	80	90	M	106	70
31	M	120	80	91	F	110	72
32	M	110	84	92	M	120	75
33	M	115	70	93	F	110	80
34	F	120	100	94	M	120	80
35	M	142	94	95	M	116	80
36	F	120	85	96	F	122	95
37	M	102	64	97	M	130	90
38	M	110	78	98	M	110	70
39	F	118	80	99	F	150	80
40	M	120	80	100	M	122	80
41	F	120	78	101	M	108	58
42	M	126	76	102	F	132	88
43	M	144	100	103	M	110	70
44	F	140	95	104	M	122	82
45	M	116	80	105	F	140	100
46	M	108	80	106	M	138	96
47	F	136	98	107	M	127	90
48	M	135	85	108	F	130	80
49	M	114	76	109	M	130	80
50	F	150	90	110	M	110	70
51	F	110	50	111	F	130	90
52	M	120	70	112	M	130	82
53	M	120	80	113	M	160	105
54	F	122	80	114	F	132	80
55	M	120	80	115	M	120	80
56	M	165	70	116	M	123	73
57	M	120	80	117	F	110	80
58	F	136	90	118	M	112	80
59	M	140	70	119	F	118	72
60	M	110	66	120	M	140	78

APÉNDICE 

Métodos estadísticos STATGRAPHICS Plus

Agustín Turrero y Pilar Zuluaga

En este Apéndice explicaremos el manejo del paquete estadístico STATGRAPHICS Plus, mostrando los procedimientos adecuados para realizar los análisis estadísticos presentados a lo largo del libro. Se utilizarán ejemplos biológicos para ilustrar dichos procedimientos y facilitar la interpretación de resultados. En lo que se refiere al aspecto más formal de los métodos estadísticos utilizados, se incluyen algunos comentarios que pretenden extender lo aprendido en los capítulos precedentes, con el único objetivo de facilitar la comprensión de los resultados obtenidos así como permitir la mejora y/o ampliación de dichos métodos estadísticos. En particular, en el Apéndice C2, dedicado a las técnicas de Estadística descriptiva, dichos comentarios serán más exhaustivos por dos razones: a) el análisis inferencial de los datos recogidos en una muestra comienza siempre con el análisis descriptivo de los mismos, de la calidad de este análisis dependerá el alcance de los resultados obtenidos, y b) al ejecutar diversos procedimientos de inferencia estadística con el STATGRAPHICS Plus aparecen, por defecto, resúmenes estadísticos y gráficos que se comentan en el Apéndice C2.

Introducción al STATGRAPHICS Plus

Este Apéndice no pretende ser un manual exhaustivo de STATGRAPHICS Plus, sino solamente una breve guía que permita ejecutar los procedimientos estadísticos vistos en el texto del libro y que; se desarrollarán a continuación con STATGRAPHICS Plus.

La primera pantalla ejecutable que aparece en STATGRAPHICS Plus, después de la carátula), se representa en la Figura C1.l

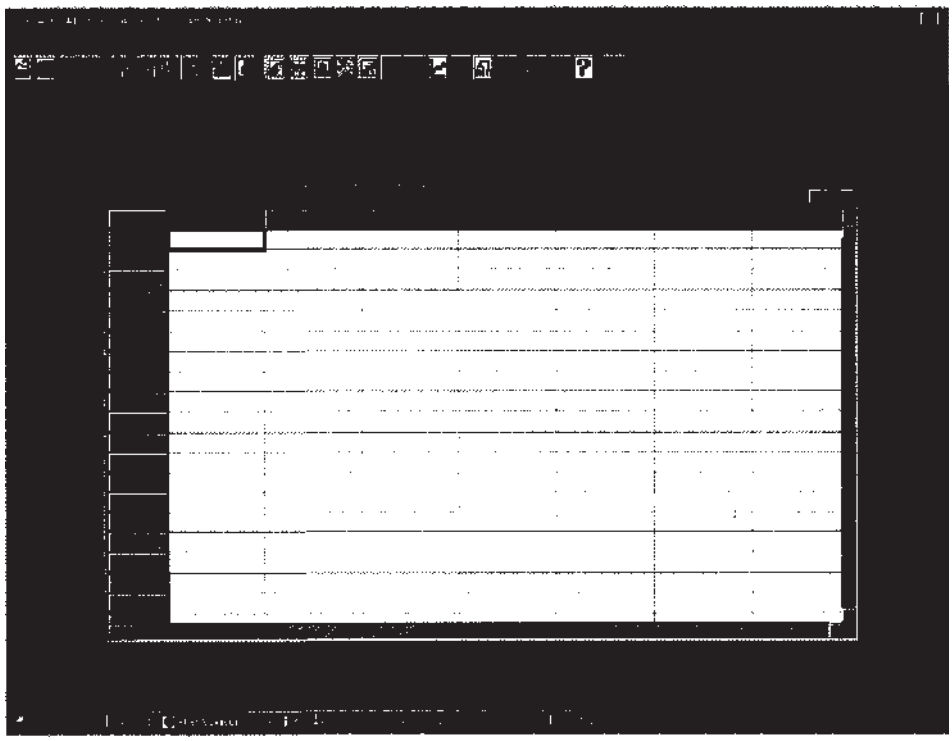


Figura C1.1. Pantalla inicial.

Distinguimos en ella cuatro elementos que nos permitirán comunicamos con STATGRAPHICS Plus para realizar nuestros análisis: la barra de menú, la barra de herramientas, la pantalla de datos y la barra de tareas. Describimos a continuación cada uno de estos cuatro elementos. Además, existe la barra de ejecución, que emerge cuando ya se ha elegido un procedimiento estadístico a realizar; dicha barra se comentará posteriormente:

C1.1 BARRA DE MENÚ

Aparece siempre; está formada por 11 procedimientos (véase Fig. C1.2).

Figura C1.2. Barra de menú.

Según nos posicionamos en cada procedimiento y hacemos un clic, aparece un menú emergente con las distintas posibilidades, algunas de ellas además acaban en una punta de flecha, que indicará la existencia de un submenú, el cual vuelve a emerger al posicionamos encima con el ratón (véase Fig. C 1.3).

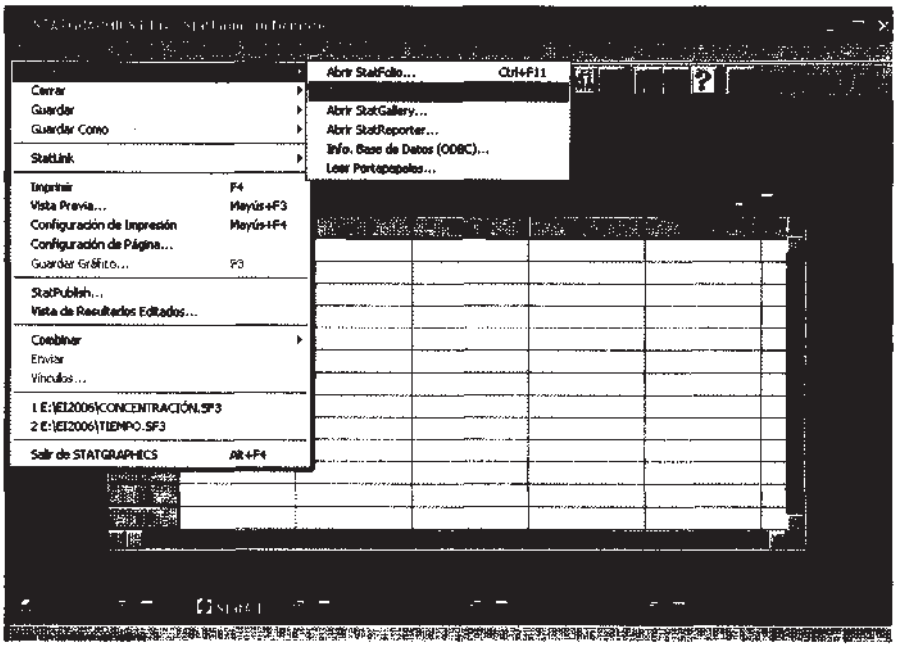


Figura C1.3. Menús emergentes de la barra de menú.

Comentaremos brevemente los 11 procedimientos:

- **Archivo:** Permite realizar tareas generales con los ficheros de datos como **Abrir, Cerrar, Guardar, Imprimir, Salir...,** algunos tienen submenús.
- **Edición:** Hace las mismas funciones que los procedimientos del mismo nombre en cualquier otra aplicación del entorno Windows, como **Copiar, Corta, Pegar,** pero además permite **Modificar** y **Generar** datos.

- **Gráficos, Descripción, Comparación, Dependencia, Avanzado:** Permiten realizar distintos tratamientos estadísticos que se explicarán con cada técnica estadística.
- **SnapStats:** Como su nombre indica (Estadística rápida) realiza de una vez varias técnicas estadísticas, tanto gráficas como numéricas, las cuales son consideradas por el programa como más habituales en ese tipo de problemas.
- **Ver, Ventana y Ayuda** hacen funciones similares que los procedimientos del mismo nombre en cualquier otra aplicación del entorno Windows.

C1.2. BARRA DE HERRAMIENTAS

Consta de distintos iconos, que realizan los procedimientos más habituales (numéricos o gráficas) que se pueden encontrar en la barra de menú y/o en la barra de ejecución. Para ejecutarlos bastará con situarnos en el icono correspondiente, hacer «clic» y se pondrá operativo (véase Fig. C1.4).

Figura C1.4. Barra de herramientas.

C1.3. BARRA DE TAREAS

Aparece en la parte baja de la pantalla, consta de 4 procedimientos (véase Fig. C1.5), los cuales se describen a continuación.

Figura C1.5. Barra de tareas.

• **StatAdvisor**
Es un asesor estadístico pues da una explicación breve sobre el interés y significado de la técnica estadística utilizada. Los comentarios del StatAdvisor aparecen, por defecto, en pantalla al ejecutar una técnica numérica pero habrá que «maximizarlo» si estamos en una técnica gráfica.

• **StatGallery**
Permite ir almacenando resultados de los análisis sucesivos que se hacen con el STATGRAPHICS Plus, de ahí el nombre de «galería de resultados».

• **StatReporter**
Es similar al StatGallery, pero tiene la ventaja de permitir incorporar texto y realizar un documento con resultado de STATGRAPHICS Plus y nuestras propias anotaciones.

• **Sin Nombre**
Cuando se comienza un nuevo estudio el STATGRAPHICS Plus engloba todo lo que realicemos en dicha sesión en lo que denomina StatFolio (datos, técnicas, StatGallery, comentarios...), por si tenemos interés de guardarlo, por defecto lo denomina **Sin nombre.**

C1.4. PANTALLA DE DATOS

La pantalla de datos es muy parecida a la hoja de datos de otros programas (véase Fig. C1.6); lo más habitual es que cada fila corresponda a los datos de un mismo individuo referentes a distintas

características (variables) que aparecen en las columnas. Por lo tanto un archivo de datos se puede considerar como una matriz donde si nos fijamos en una columna tenemos los valores de una variable para todos los individuos, y si nos fijamos en una fila tenemos todas las variables para un individuo.

Los datos pueden leerse de ficheros ya existentes mediante el procedimiento **Archivo,** de la barra de menú, menú **Abrir** y submenú **Abrir Datos,** o bien sobre el tercer icono de la barra de herramientas. También pueden crearse directamente sobre la plantilla que aparece en pantalla. Los ficheros que crea el STATGRAPHICS Plus tienen la extensión sf3.

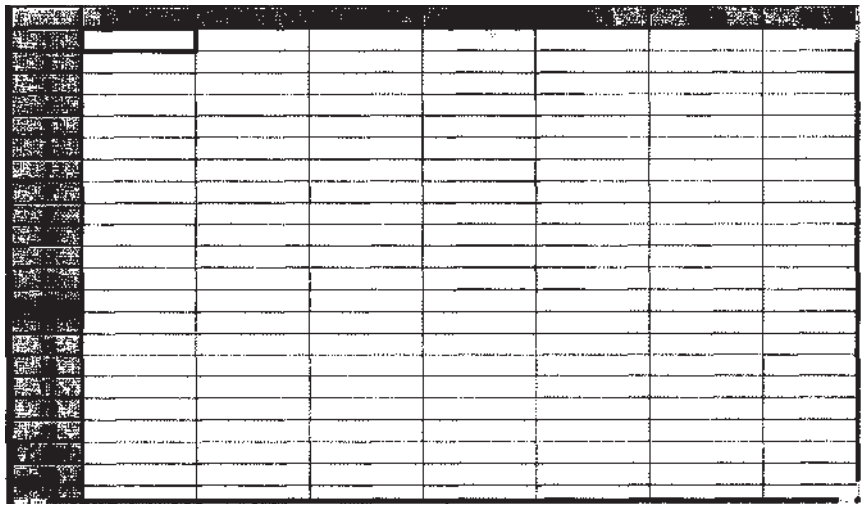
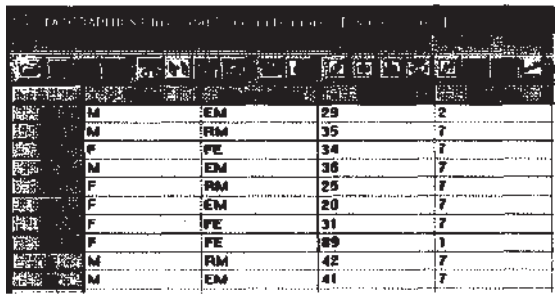


Figura C1.6. Pantalla de datos.

La Figura C1.7 muestra la pantalla de datos para los valores de las 4 variables correspondientes a los primeros 10 pacientes del Ejemplo 1.1.1 del Capítulo 1.



M	EM	29	2
M	RM	35	7
F	FE	34	7
M	EM	36	7
F	RM	25	7
F	EM	20	7
F	FE	31	7
F	FE	39	1
M	RM	42	7
M	EM	41	7

Figura C1.7. Fichero de datos.

Para modificar las características de una columna, ya sea nombre, tipo de variable..., basta con colocarse sobre la columna, hacer clic con el botón izquierdo y la columna entera se pondrá en negro, si ahora se hace che en **Edición** aparece un submenú que permite hacer diversas tareas, una de ellas es **Modificar columna.** Podemos destacar que el **Nombre** puede tener hasta 32 caracteres, comenzando por una letra (véase Fig. C1.8).

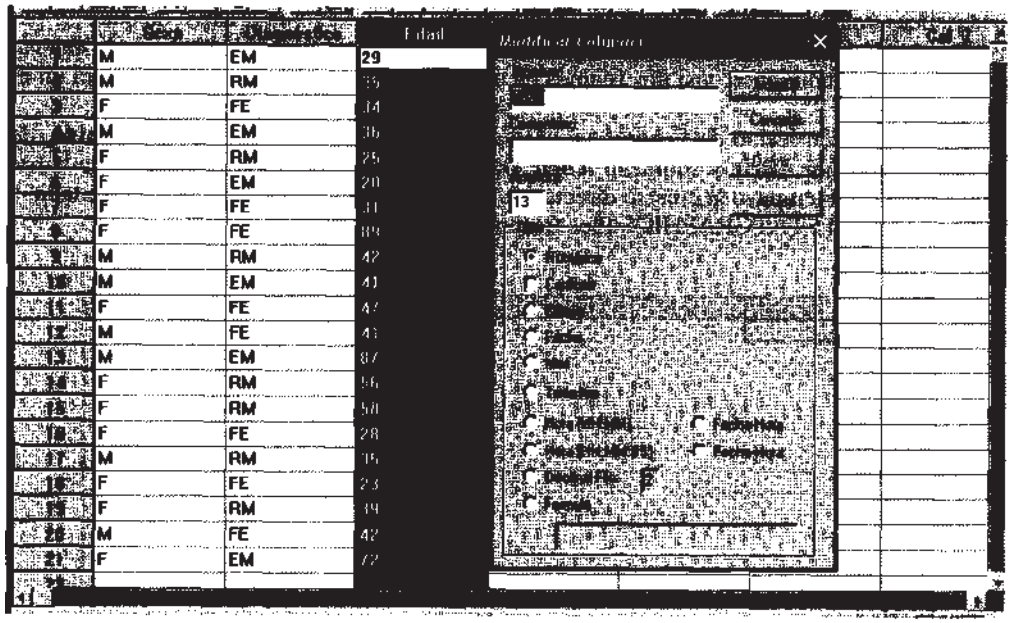


Figura C1.8. Modificación de una columna.

Una vez que se tiene el fichero de datos se pueden ejecutar los procedimientos del STATGRA PHICS Plus que están en la Barra de menú: **Gráficos, Descripción, Comparación, Dependencia Avanzado y SnapStats,** pero para ello hay que indicar previamente qué variable va a ser objeto de estudio, por ejemplo, nos posicionamos sobre edad, se pondrá en negro y al posicionarnos sobre la flecha pasará a Datos, según se ve en la Figura C1.9. Si ahora se pulsa **Aceptar** se ejecutará el procedimiento elegido.

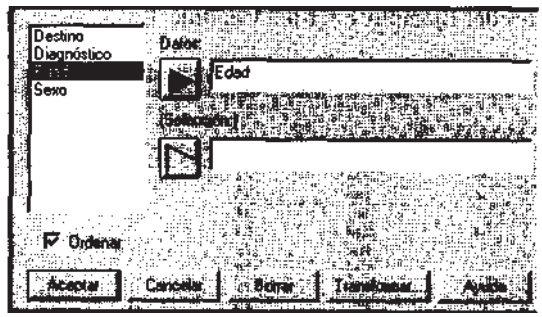


Figura C1.9. Selección de variable a estudiar.

Una vez ejecutado el procedimiento, aparece una nueva barra, la barra de ejecución.

C1.5. BARRA DE EJECUCIÓN

Cuando en la barra de menú elegimos un procedimiento estadístico y lo ejecutamos, se producen por defecto una serie de resultados que podrán ser ampliados en la barra de ejecución. Aunque la barra de ejecución consta de más iconos sólo describiremos los 4 principales:

- El icono  permite cambiar de variable con la que estamos trabajando.
- El icono  , **Opciones Tabulares,** permite ampliar o reducir resultados numéricos.
- El icono  , **Opciones Gráficas,** permite ampliar o reducir resultados gráficos.
- El icono  , **Opciones Guardar Resultados,** permite grabar algunos resultados numéricos.

Estadística descriptiva

Ejemplo C2.1

La tabla siguiente muestra la información relativa a 88 pacientes diagnosticados con carcinoma epidermoide de la cavidad oral. Se describen 9 variables de diferente naturaleza. De tipo discreto: RL, recidiva local (si, no), **Sexo** (hombre, mujer), **Localización,** del tumor principal (mucosa, lengua, labio, suelo de boca, orofaringe), **Infil.,** infiltración ganglionar (si, no), todas ellas de naturaleza cualitativa nominal, **Categoría,** del tumor (< 20mm, 20-40 mm, > 40 mm, invasivo) de naturaleza cualitativa ordinal, y N° **gang.,** número de ganglios infiltrados (0,1, 2, 3,4,....) de naturaleza cuantitativa. De tipo continuo: SG, supervivencia global (meses), **Edad** (años) y **Diám.,** diámetro máximo del tumor (mm).

N.°	RL	SG	Edad	Sexo	Localización	Categoría	Diám.	Infil.	N.° gak
1	sí	9	58	hombre	orofaringe	invasivo	50	si	1
2	sí	4	57	hombre	lengua	>40 mm	42	sí	2
3	no	19	84	mujer	lengua	20-40 mm	25	no	0
4	no	116	61	hombre	labio	20-40 mm	22	no	0
5	no	117	65	hombre	labio	20-40 mm	22	no	0
6	no	23	67	hombre	lengua	<20mm	15	sí	6
7	no	71	68	hombre	suelo boca	20-40 mm	24	no	0
8	no	107	62	hombre	labio	20-40 mm	30	no	0
9	sí	113	55	hombre	lengua	20-40 mm	30	no	0
10	sí	7	70	hombre	mucosa	20-40 mm	30	sí	9
11	no	100	71	hombre	labio	<20mm	20	no	0
12	sí	12	58	hombre	mucosa	20-40 mm	30	sí	2
13	no	24	55	hombre	lengua	20-40 mm	28	sí	1
14	sí	8	53	hombre	suelo boca	invasivo	34	no	0
15	sí	81	68	mujer	lengua	>40 mm	50	no	0
16	no	94	57	hombre	suelo boca	20-40 mm	40	no	0
17	no	94	70	mujer	lengua	<20mm	18	sí	1
18	no	15	47	hombre	orofaringe	>40 mm	60	no	0
19	no	25	53	hombre	lengua	>40 mm	45	sí	9
20	no	98	72	hombre	labio	<20mm	20	no	0
21	no	48	69	mujer	mucosa	<20mm	15	sí	3

N.°	RL	SG	Edad	Sexo	Localización	Categoría	Diám.	Infil.	N.° gan.
22	no	**93**	61	hombre	lengua	20-40 mm	**30**	no	**0**
23	**sí**	**14**	60	mujer	lengua	<20mm	**10**	no	**0**
24	**sí**	**12**	59	hombre	lengua	20-40 mm	**25**	no	**0**
25	no	**87**	67	hombre	lengua	<20mm	**18**	no	**0**
26	**sí**	**37**	44	hombre	lengua	20-40 mm	40	**sí**	**1**
27	no	**85**	68	hombre	labio	20-40 mm	**25**	no	**0**
28	**sí**	**6**	65	hombre	orofaringe	invasivo	**45**	**sí**	**6**
29	**sí**	**10**	65	hombre	suelo boca	invasivo	**35**	no	**0**
30	**sí**	**9**	74	mujer	mucosa	invasivo	**45**	**sí**	**1**
31	no	**90**	67	hombre	suelo boca	<20mm	**15**	no	**0**
32	no	**88**	53	mujer	lengua	20-40 mm	**30**	no	**0**
33	no	**34**	44	hombre	suelo boca	20-40 mm	**32**	no	**0**
34	no	**81**	73	hombre	suelo boca	20-40 mm	24	no	**0**
35	**sí**	**87**	54	hombre	suelo boca	>40mm	**35**	no	**0**
36	**sí**	**8**	45	hombre	mucosa	20-40 mm	**30**	**sí**	**2**
37	**sí**	124	33	mujer	lengua	<20mm	20	no	**0**
38	**sí**	**31**	56	hombre	suelo boca	invasivo	**35**	**sí**	**5**
39	**sí**	**75**	47	hombre	labio	<20mm	**8**	no	**0**
40	no	**77**	44	hombre	lengua	<20mm	**15**	no	**0**
41	no	**65**	68	hombre	lengua	20-40 mm	**25**	no	**0**
42	no	**66**	66	hombre	lengua	<20mm	20	no	**0**
43	no	**8**	66	hombre	labio	<20mm	20	no	**0**
44	**sí**	**11**	63	mujer	lengua	invasivo	**60**	no	**0**
45	no	**25**	63	hombre	suelo boca	<20mm	20	**sí**	**4**
46	no	**64**	66	mujer	lengua	<20mm	**10**	no	**0**
47	no	**17**	80	hombre	labio	<20mm	**7**	**sí**	**4**
48	no	**70**	52	hombre	lengua	<20mm	**15**	no	**0**
49	**sí**	**49**	64	hombre	labio	<20mm	**10**	no	**0**
50	**sí**	**26**	62	hombre	suelo boca	<20mm	20	no	**0**
51	**sí**	**10**	40	hombre	suelo boca	<20mm	**15**	**sí**	**1**
52	no	**58**	64	mujer	lengua	20-40 mm	**30**	no	**0**
53	**sí**	**75**	41	hombre	lengua	invasivo	40	**sí**	**1**
54	**sí**	**27**	81	hombre	orofaringe	20-40 mm	40	**sí**	**4**
55	**sí**	**12**	52	hombre	suelo boca	20-40 mm	**30**	**sí**	**3**
56	**sí**	**12**	50	mujer	lengua	20-40 mm	**35**	no	**0**
57	no	**62**	57	hombre	lengua	<20mm	20	no	**0**
58	**sí**	**9**	45	hombre	suelo boca	>40mm	**50**	**sí**	**1**
59	no	**60**	57	hombre	lengua	20-40 mm	**30**	no	**0**
60	no	**58**	**64**	mujer	labio	<20mm	**20**	no	**0**
61	**sí**	**34**	59	hombre	suelo boca	invasivo	**25**	no	**0**
62	no	**66**	60	hombre	orofaringe	20-40 mm	**30**	no	**0**
63	no	**65**	57	hombre	lengua	<20mm	20	no	**0**
64	no	**63**	61	hombre	suelo boca	20-40 mm	22	**sí**	**2**
65	**sí**	**17**	55	hombre	suelo boca	invasivo	**30**	no	**0**
66	**sí**	**10**	53	hombre	suelo boca	invasivo	**60**	no	**0**
67	no	**3**	51	hombre	suelo boca	<20mm	20	**sí**	**3**
68	no	**50**	53	hombre	suelo boca	<20mm	20	no	**0**
69	**sí**	183	44	hombre	lengua	<20mm	20	no	**0**
70	no	**16**	77	hombre	orofaringe	20-40 mm	40	no	**0**
71	no	**58**	56	mujer	lengua	<20mm	20	no	**0**
72	no	120	59	hombre	lengua	20-40 mm	**30**	no	**0**
73	**sí**	22	52	hombre	mucosa	20-40 mm	40	no	**0**
74	no	**94**	58	hombre	lengua	<20mm	**19**	no	**0**
75	no	**12**	66	hombre	suelo boca	20-40 mm	**26**	no	**0**
76	**sí**	24	51	hombre	lengua	<20rnm	20	no	**0**
77	no	120	55	mujer	lengua	<20mm	**12**	**sí**	**2**

N.°	RL	SG	Edad	Sexo	Localización	Categoría	Diám.	Infil.	N.° ga n
78	no	**99**	61	hombre	lengua	20-40 mm	**29**	no	**0**
79	**sí**	**52**	68	hombre	lengua	20-40 mm	**35**	no	**0**
80	no	159	45	hombre	suelo boca	20-40 mm	**25**	no	**0**
81	no	**33**	55	hombre	suelo boca	20-40 mm	**25**	no	**0**
82	**sí**	**11**	57	hombre	suelo boca	invasivo	**26**	no	**0**
83	**sí**	**8**	61	hombre	orofaringe	>40mm	42	**sí**	**4**
84	no	**8**	65	mujer	mucosa	20-40 mm	**25**	**sí**	**3**
85	**sí**	**11**	50	hombre	mucosa	invasivo	40	**sí**	**1**
86	no	**2**	70	hombre	labio	<20mm	**18**	**sí**	**1**
87	no	**19**	51	hombre	lengua	20-40 mm	22	no	**0**
88	no	129	64	hombre	orofaringe	20-40 mm	**35**	no	**0**

La forma más útil de introducir los datos, para su posterior análisis con el paquete STATGRÁPHICS Plus se muestra en el fichero de datos, **ejemplo 1.sf3,** creado en una hoja de datos STATGRAPHICS Plus, que aparece en la Figura C2.1

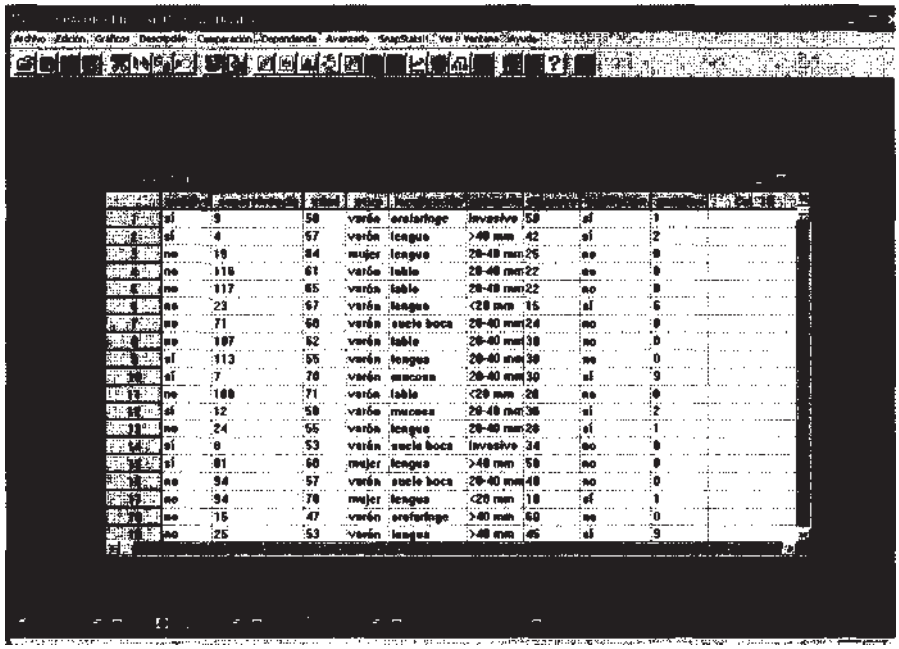


Figura C2.1. Fichero de datos correspondiente al ejemplo C2.1.

En la hoja de datos se muestran los valores relativos a los primeros 19 pacientes. Las variables de naturaleza discreta cualitativa pueden grabarse con sus modalidades o etiquetas no numéricas, por ejemplo el sexo como hombre o mujer; de esta forma los resultados del programa nos mostrarán dichas modalidades haciendo más fácil su interpretación. En caso de codificar numéricamente dichas modalidades, por ejemplo el sexo como 1 y 2, deberemos etiquetar sobre los resultados si queremos que se muestren los nombres de las modalidades. Las variables de naturaleza cuantitativa, discreta o continua, deben grabarse con sus valores respectivos. Conviene recordar aquí que ¡la descripción numérica de una variable cualitativa concluye con las tablas de frecuencias y los gráficos correspondientes, mientras que la descripción de una variable cuantitativa! es mucho más ruja, al permitir el cálculo de medidas centrales, de dispersión y de forma.

C2.1. DESCRIPCIÓN DE UNA VARIABLE CUALITATIVA

Para obtener la tabla de frecuencias de la localización del tumor elegimos el procedimiento **Descripción** de la barra de menú, a continuación **Datos cualitativos,** y finalmente **Tabulación.** Una vez desplegada la ventana del procedimiento, seleccionamos, sobre el listado de variables, **localización** y arrastramos dicha variable al campo *Datos*. Pulsando **Aceptar** se obtienen los resultados de la Figura C2.2.

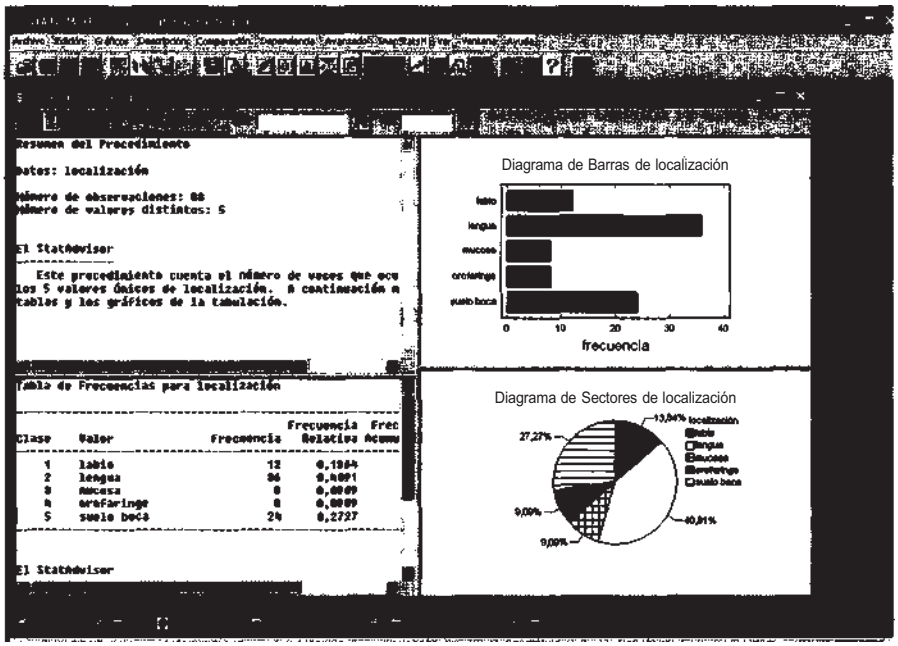


Figura C2.2. Pantalla de resultados correspondiente a la localización del tumor.

Las dos ventanas de la izquierda muestran los resultados numéricos de este procedimiento (Opciones tabulares), las dos de la derecha los resultados gráficos (Opciones gráficas). La ampliación o reducción de estas opciones se lleva a cabo, respectivamente, a través de los iconos y , segundo y tercero, de la barra de ejecución. Para este procedimiento no existen más opciones, ni numéricas ni gráficas. Haciendo doble clic sobre cualquiera de las ventanas se maximiza el contenido de la ventana elegida. La tabla de frecuencias completa, tras el doble clic sobre dicha ventana, se muestra en la Figura C2.3. Conviene puntualizar que las dos últimas columnas de resultados, las de frecuencias absolutas y relativas acumuladas, carecen de interés debido a la naturaleza cualitativa nominal de la variable localización. Las tablas de frecuencias acumuladas tienen sentido cuando al menos existe una ordenación entre las modalidades de la variable, por ello sólo resultan recomendables para variables cualitativas ordinales, como la categoría del tumor, y variables cuantitativas, discretas o continuas.

Las gráficas adecuadas para representar las distribuciones de frecuencias de una variable cualitativa son el diagrama de barras y el diagrama de sectores. Un clic con el botón derecho del ratón sobre cualquiera de los gráficos (no es necesario maximizarlos) abre una ventana de **Opciones** de dos tipos, **de ventana y gráficas.** Con las primeras podemos cambiar la forma del gráfico, con las segundas podemos editar dicho gráfico. Por ejemplo, para el diagrama de barras, si desplegamos las **Opciones de ventana** podemos cambiar la *Escala* del gráfico, porcentajes o frecuencias, la *Dirección* del gráfico, horizontal o vertical, el *Tipo de diagrama,* adosado o apilado, y la *Línea base,* cortando el gráfico por un valor fijado. En el gráfico de barras mostrado anteriormente las

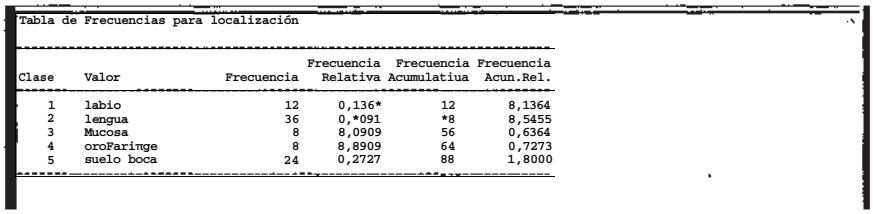


Tabla de Frecuencias para localización

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulatiua	Frecuencia Acun.Rel.
1	labio	12	0,136*	12	8,1364
2	lengua	36	0,*091	*8	8,5455
3	Mucosa	8	8,0909	56	0,6364
4	oroFaringe	8	8,8909	64	0,7273
5	suelo boca	24	0,2727	88	1,8000

Figura C2.3. Distribución de frecuencias de la localización del tumor.

Opciones de ventana elegidas fueron respectivamente, frecuencias, horizontal, adosado y línea base a partir de 0. Si para el mismo diagrama desplegamos las **Opciones gráficas** podemos editar el título, los efectos, colores, etc....del gráfico. Por ejemplo, si hubiésemos codificado la localización como 1, 2, 3, 4 y 5 en la hoja de datos, en dichas **Opciones** gráficas podríamos dar nombre a cada una de estas modalidades, bastaría seleccionar *eje Y,* y escribir los nombres de las modalidades en *Etiquetas.*

C2.2. DESCRIPCIÓN DE UNA VARIABLE CUANTITATIVA DISCRETA

Para obtener la tabla de frecuencias del número de ganglios infiltrados elegimos el procedimiento **Descripción** de la barra de menú, a continuación **Datos cualitativos,** y finalmente **Tabulación.** Seleccionamos la variable **ganglios** y la arrastramos al campo *Datos.* La elección de la opción **Datos cualitativos** puede llevar a confusión debido a la naturaleza numérica de la variable, sin embargo, para este programa estadístico es la opción adecuada para la obtención de tablas de frecuencias y algún gráfico. Por consiguiente, ésta será la secuencia de procedimientos para variables de tipo discreto, tanto cualitativas como cuantitativas. La tabla de frecuencias y el diagrama de barras se muestran en las Figuras C2.4 y C2.5.

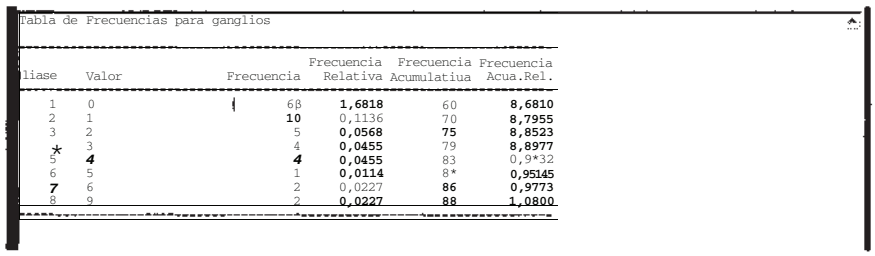


Tabla de Frecuencias para ganglios

liase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulatiua	Frecuencia Acua.Rel.
1	0	6β	1,6818	60	8,6810
2	1	10	0,1136	70	8,7955
3	2	5	0,0568	75	8,8523
*	3	4	0,0455	79	8,8977
5	4	4	0,0455	83	0,9*32
6	5	1	0,0114	8*	0,95145
7	6	2	0,0227	86	0,9773
8	9	2	0,0227	88	1,0800

Figura C2.4. Distribución de frecuencias del número de ganglios infiltrados.

Se ha suprimido el gráfico de sectores para destacar la naturaleza cuantitativa de la variable descrita, número de ganglios infiltrados, dado que dicho gráfico está indicado para variables cualitativas. Un déficit de este procedimiento para este tipo de variables (cuantitativas discretas) es que no permite escalar los valores de la variable, esto se puede ver observando la posición de la barra correspondiente al valor 9 que debería estar separada 3 unidades de la relativa al valor 6. Para continuar el análisis descriptivo de esta variable debemos volver a la barra de menú y elegir el procedimiento **Descripción,** a continuación **Datos numéricos,** y finalmente **Análisis unidimensional.** La Figura C2.6 muestra las opciones tabulares y gráficas recomendadas.

En la opción tabular **Resumen estadístico** se obtienen tres medidas centrales como son la media, la mediana y la moda, dos medidas de posición, primer y tercer cuartiles, y cinco medidas de dispersión, varianza, desviación típica, rango, rango intercuartílico y coeficiente de variación.

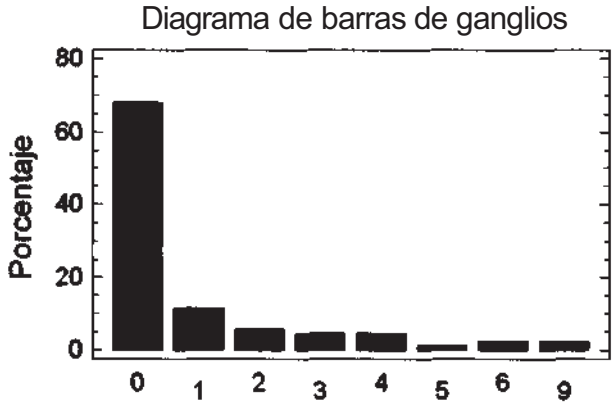


Figura C2.5. Diagrama de barras del número de ganglios infiltrados.

Algunos de los resultados obtenidos merecen un comentario más detallado. El hecho de que el valor 0 aparezca en 60 de los 88 valores registrados para esta variable explica el desplazamiento de todas las medidas centrales hacia dicho valor, esto cuestiona el apelativo de centrales de estas medidas; en rigor la única medida que obedece un criterio de posición central es la mediana, 0 ganglios infiltrados es el valor que aparece en el «medio» del conjunto de datos ordenados. La media aritmética, 0.948132, se desplaza por el efecto de datos atípicos, hay 2 pacientes con 9 ganglios infiltrados. La moda, el valor más frecuente de la variable, no obedece a un criterio de posición o promedio como las dos anteriores, en su obtención sólo interviene la frecuencia absoluta de los diferentes valores de la variable por lo que, como ocurre en este caso, puede ser un valor extremo de dicha variable.

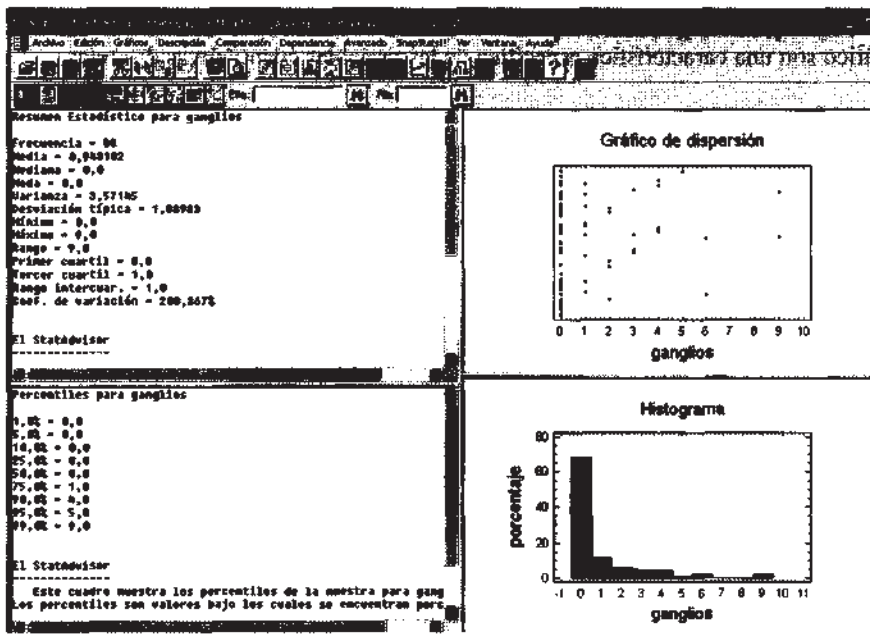


Figura C2.6. Pantalla de resultados correspondiente al número de ganglios infiltrados.

El coeficiente de variación CV es una medida de dispersión relativa, el valor 200.367%, indica el porcentaje que representa la desviación típica s = 1.8898 en relación a la media x = 0.9431. La primera interpretación de este resultado podría ser que la variable número de ganglios aparece muy dispersa en esta muestra, sin embargo un análisis más profundo nos debería llevar a concluir lo contrario. Una muestra de 88 datos, 60 de los cuales coinciden, es una muestra homogénea. En este caso los datos atípicos y sobre todo el orden de magnitud del valor más repetido, 0, explican este hecho. Si imaginamos una distribución idéntica a ésta pero desplazando los valores en dos unidas des, de 2 a 11 ganglios infiltrados, la media y desviación típica serían ahora x = 2.9431 y s = 1.8898 (la media se incrementa en dos unidades y la desviación típica permanece invariable) con lo que el CV = 64.21%.

En la opción tabular **Percentiles** se obtienen 9 percentiles, pudiéndose obtener otros en la opciones de ventana de este procedimiento. De especial interés son los percentiles 25, 50 y 75 que corresponden respectivamente al primer cuartil q_1, la mediana y el tercer cuartil q_3.

En las dos opciones gráficas se muestran las distribuciones de frecuencias absolutas de la variable. Para obtener el gráfico de dispersión, tal y como aparece en la Figura C2.6, abrir las **Opciones gráficas** de este resultado, seleccionar *eje X y* marcar *Por:* 1, de esta forma aparecerán los valores de la variable de uno en uno. Para obtener el histograma, tal y como aparece en la Figura C2.6, abrir las **Opciones de ventana** de este resultado, marcar *Nº de clases:* 10, *Límite inferior.* -0.5, *Límite superior.* 9.5, *Frecuencia:* **Relativa,** y *Tipo de gráfico:* **Histograma.** Las tres primeras marcas elegidas persiguen que en la base de cada rectángulo figure el valor entero de la variable, esto es debido a que el procedimiento elegido, **Datos numéricos,** está pensado en este programa para la descripción de variables de naturaleza continua. La ventaja de este gráfico respecto al gráfico de barras anterior es que ahora sí aparecen escalados los valores de la variable. Igual que en el gráfico de dispersión, debemos abrir las **Opciones gráficas** de este resultado, seleccionar *eje X y* marcar *Por.* 1 para que aparezcan los valores de la variable de uno en uno. Otro gráfico de interés en la descripción de una variable cuantitativa discreta es el diagrama acumulativo o acumulado. Para obtener esta gráfica con el STATGPRAPHICS Plus basta marcar las mismas opciones anteriores y poner *Frecuencia:* **Acumulada.** La Figura C2.7 muestra dicha gráfica. Se ha cambiado el título original, Histograma, por el de Diagrama acumulado para enfatizar la diferencia con el Histograma anterior, que suele representar frecuencias, absolutas o relativas, no acumuladas. El título se cambia abriendo, sobre el Histograma, las **Opciones gráficas,** eligiendo la opción *Título Principal.* Los escalones mostrados en este gráfico son una característica específica de las variables cuantitativas discretas.

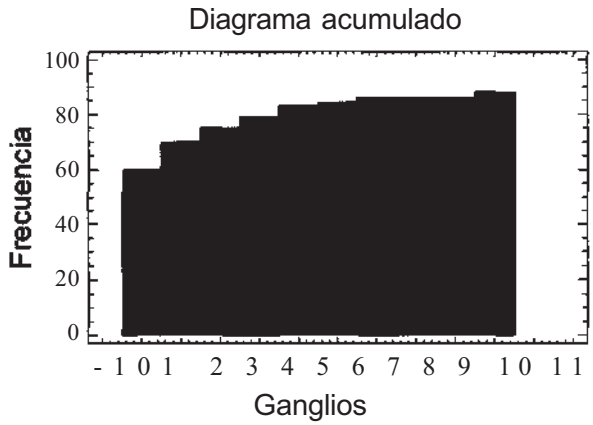


Figura C2.7. Diagrama acumulado del número de ganglios infiltrados.

C2.3. DESCRIPCIÓN DE UNA VARIABLE CUANTITATIVA CONTINUA

Para la descripción de la variable continua edad elegimos la anterior secuencia de procedimientos, **Descripción, Datos numéricos, Análisis unidimensional.** Seleccionamos la variable **edad** y la arrastramos al campo *Datos.* Pulsando **Aceptar** se obtienen los resultados. La Figura C2.8 muestra las opciones tabulares y gráficas recomendadas. Cuando se quiere describir una variable continua debemos definir el número de clases y el tamaño de las mismas. Sobre la opción tabular **Tabla de frecuencias** desplegamos las **Opciones de ventana** y elegimos *N.º de Clases:* 7, *Límite Inferior:* 32, *Límite Superior:* 88. Se han elegido estos 3 valores siguiendo *la regla de Sturges,* descrita en el apartado 1.3 del Capítulo 1, teniendo en cuenta el tamaño de la muestra n = 88, y las observaciones menor y mayor, 33 y 84 años, respectivamente. En general, el agrupamiento en clases de una variable continua persigue el poder visualizar la forma, simétrica, asimétrica a la derecha, apuntada, etc., de su gráfica de frecuencias, por lo que es habitual llevar a cabo diferentes agrupamientos en el mismo análisis. Abundaremos en este hecho al comentar los resultados gráficos obtenidos.

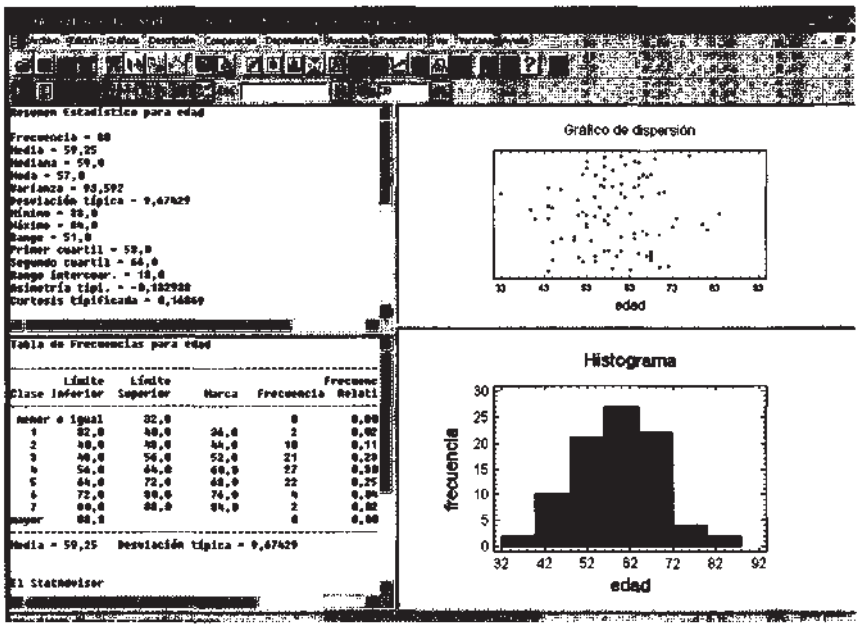
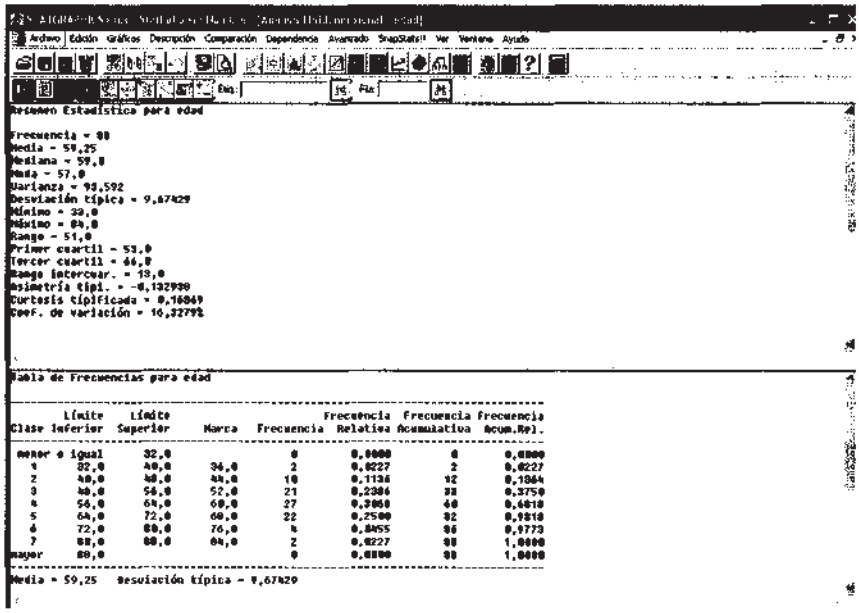


Figura C2.8. Pantalla de resultados correspondiente a la edad.

En la Figura C2.9 se muestran ampliadas las dos opciones tabulares seleccionadas, en la parte inferior la tabla de frecuencias y en la superior el resumen estadístico. La **Tabla de frecuencias** contiene 7 intervalos o clases de amplitud 8 años cada uno; este procedimiento no permite hacer intervalos de distinta amplitud. Asimismo, se muestran los puntos medios de los intervalos llamados marcas de clase. El **Resumen estadístico** presenta junto a las medidas centrales y de dispersión, anteriormente comentadas, dos medidas de forma, los coeficientes de asimetría y curtosis tipificados, que permiten determinar si la muestra procede de una distribución normal, en general valores próximos a 0 de ambos coeficientes, como en nuestro caso, apuestan por este modelo probabilístico. Otra opción tabular de interés es la obtención de **Percentiles,** que ya se comentó en el apartado de variable cuantitativa discreta.

En las dos opciones gráficas se muestran las distribuciones de frecuencias absolutas de la variable. Para obtener el histograma, tal y como aparece en la Figura C2.10, abrir las **Opciones de ventana** de este resultado, marcar *Frecuencia:* **Relativa,** y *Tipo de gráfico:* **Histograma.** El número de clases y los límites inferior y superior son los mismos que se eligieron en la opción tabular **Tabla**



Resumen Estadístico para edad

Frecuencia	88
Media	59,25
Mediana	59,0
Moda	57,0
Varianza	93,592
Desviación típica	9,67429
Mínimo	32,0
Máximo	84,0
Rango	51,0
Primer cuartil	53,0
Tercer cuartil	66,0
Rango intercuar.	13,0
Asimetría tipi.	-0,132930
Curtosis tipificada	0,16849
Coef. de variación	16,32779%

Tabla de Frecuencias para edad

Clase	Límite Inferior	Límite Superior	Marca	Frecuencia	Frecuencia Relativa	Frecuencia Acumulativa	Frecuencia Acum.Rel.
menor o igual	32,0	32,0		0	0,0000	0	0,0000
1	32,0	40,0	36,0	2	0,0227	2	0,0227
2	40,0	48,0	44,0	10	0,1136	12	0,1364
3	48,0	56,0	52,0	21	0,2386	33	0,3750
4	56,0	64,0	60,0	27	0,3068	60	0,6818
5	64,0	72,0	68,0	22	0,2500	82	0,9318
6	72,0	80,0	76,0	4	0,0455	86	0,9773
7	80,0	88,0	84,0	2	0,0227	88	1,0000
mayor	88,0			0	0,0000	88	1,0000

Media = 59,25 Desviación típica = 9,67429

Figura C2.9. Estadísticos y tablas de frecuencias de la edad.

de Frecuencias. Para editar los extremos de cada clase y sus marcas, mostrar las **Opciones gráficas** de este resultado, seleccionar *eje X* y marcar *Desde:* 32, *Hasta:* 88, *Por:* 4, de esta forma, en el eje X aparecen las edades de 4 en 4 años empezando en 32 y acabando en 88 años.

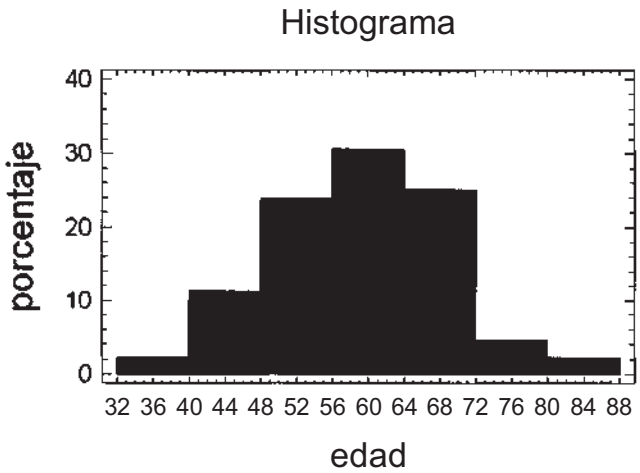


Figura C2.10. Histograma de porcentajes de la edad categorizada en 7 clases.

Si hacemos un agrupamiento en clases más fino que el anterior, eligiendo 10 intervalos, obtendríamos el segundo histograma que aparece en la Figura C2.11. Ahora, en las **Opciones de ventana** marcaríamos *Nº de Clases:* **10** y en las **Opciones gráficas** *Por:* **8,** siendo las demás opciones las mismas para ambos gráficos. La apariencia de este nuevo histograma parece alejar la idea de normalidad de la variable edad. Como se verá posteriormente, sólo un test estadístico de normalidad permitirá admitir o rechazar dicha conjetura.

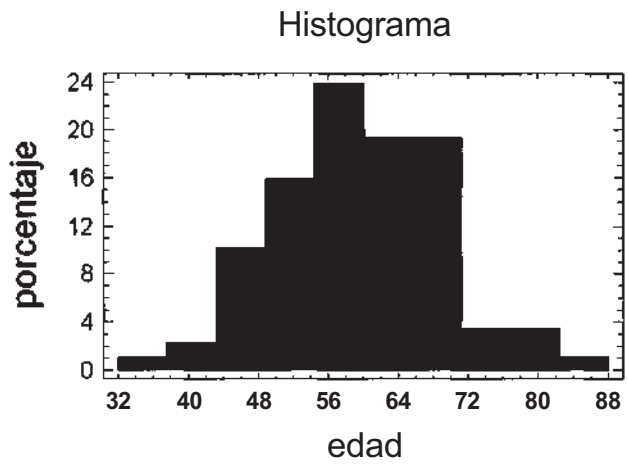


Figura C2.11. Histograma de porcentajes de la edad categorizada en 10 clases.

Otro gráfico de interés en la descripción de una variable cuantitativa continua es el polígono acumulado, que aparece en la Figura C2.12 en el lado inferior derecho. Para obtener esta gráfica con el STATGRAPHICS Plus basta marcar las mismas opciones anteriores, 7 clases, y poner *Frecuencia:* **Acumulada,** *Tipo de gráfico:* **Polígono.** Se ha cambiado el título original, Histograma, por el de Polígono acumulado para enfatizar la diferencia con el Histograma anterior, que suele representar frecuencias, absolutas o relativas, no acumuladas. El título se cambia abriendo, sobre el Histograma, las **Opciones graneas,** eligiendo la opción *Título principal.* Los tramos de esta línea poligonal «sustituyen» a los escalones del diagrama acumulado de la variable discreta, destacando el carácter continuo de esta variable.

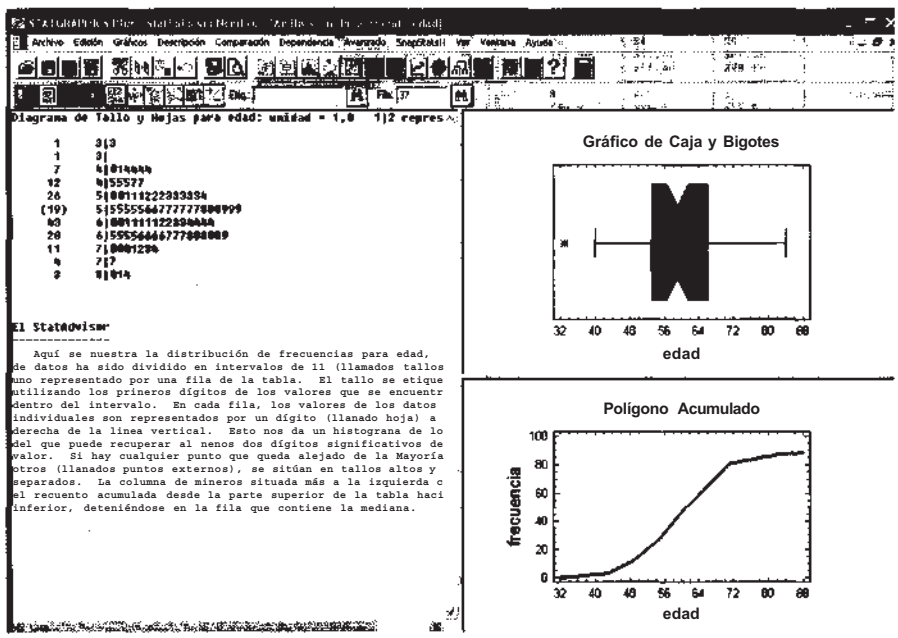


Figura C2.12. Diagrama de tallo y hojas, gráfico de caja y bigotes y polígono acumulado de la edad.

Se describen dos gráficos adicionales, válidos para variables cuantitativas, el diagrama de cajas y bigotes y el diagrama de tallo y hojas (véanse los apartados 1.6 y 1.2 del Capítulo 1, respectivamente). El primero de ellos, que se muestra en la parte superior derecha de la Figura C2.12, permite mostrar medidas centrales y de localización así como dar una idea de la dispersión y forma de la distribución de frecuencias. La caja central tiene por extremos q_1 *y* q_3, el primer y tercer cuartiles, la muesca en la caja se corresponde con el valor de la mediana, la cruz en el interior de la caja marca la posición de la media, las líneas horizontales, a ambos lados de la caja, llamadas bigotes, unen los valores adyacentes. Finalmente, en el exterior del bigote de la izquierda se marca el dato atípico 33 años. Este gráfico se selecciona entre las Opciones gráficas del procedimiento.

La última gráfica, más simple que el histograma, es el diagrama de tallo y hojas, para obtenerla hay que desplegar las Opciones tabulares de este procedimiento y seleccionar **Diagrama de tallo y hojas.** En la parte izquierda de la Figura C2.12 se muestra dicha gráfica.

C2.4. DESCRIPCIÓN CONJUNTA DE DOS VARIABLES

La descripción que se presenta a continuación se refiere a dos variables de cualquier naturaleza. Para obtener la tabla de frecuencias conjuntas de la localización del tumor según la aparición o no de recidiva local, elegimos el procedimiento **Descripción** de la barra de menú, a continuación **Datos cualitativos,** y finalmente **Tabulación cruzada.** Seleccionamos la variable **localización** y la arrastramos al campo *variable fila,* seleccionamos la variable **recidiva** y la arrastramos al campo *variable columna.* La Figura C2.13 muestra las opciones tabulares y gráficas recomendadas.

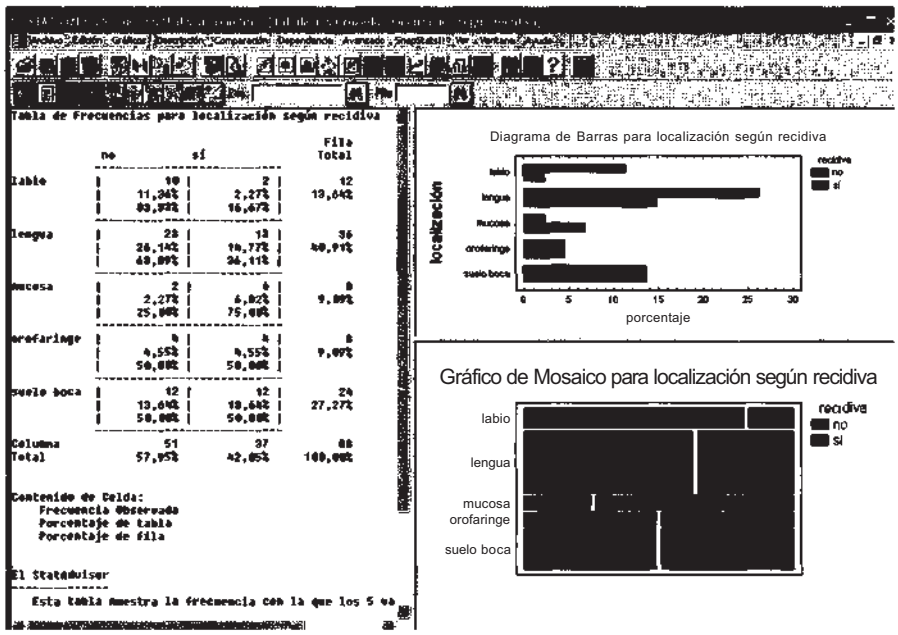


Figura C2.13. Pantalla de resultados correspondiente a la localización del tumor con respecto a la recidiva local.

La Opción tabular **Tabla de frecuencias** muestra el reparto de los 88 pacientes en cada una de las 10 celdas, en la última columna aparecen los totales de cada localización y el porcentaje correspondiente, en la última fila se muestran los totales con y sin recidiva y sus porcentajes. De especial interés resultan la segundas anotaciones de cada celda, los porcentajes que ahí figuran se refiereren,

en primer lugar, al reparto de los pacientes de cada fila en las dos celdas, porcentajes absolutos, y en segundo lugar, a los porcentajes que representan sobre el total de la fila, porcentajes relativos, por ejemplo, los 36 pacientes, que tienen el tumor localizado en la lengua, representan el 40.91% del total. Este porcentaje resulta de sumar el 26.14% que no tuvieron recidiva local y el 14.71% restante que sí tuvo recidiva. El reparto porcentual de los 36 pacientes atendiendo a la aparición o no de recidiva fue del 36.11% y del 63.89% respectivamente. La observación y comparación de estos porcentajes de celda, a lo largo de las filas, resulta de especial interés, así puede observarse cómo en los tumores de labio y lengua la no aparición de recidiva es más frecuente que la aparición de la misma; en los tumores de suelo de boca y orofaringe las frecuencias de aparición y no aparición de recidiva coinciden; finalmente, se observa que, en los tumores de mucosa, la aparición de recidiva es 3 veces más frecuente que la no aparición de la misma. Como se ha comentado anteriormente sólo un test estadístico conducirá a establecer, en este caso, la dependencia o no de los dos factores descritos en la tabla.

Las dos opciones gráficas, **Diagrama de barras y Gráfico de mosaico,** permiten visualizar lo que se acaba de comentar. El Diagrama de barras mostrado en la pantalla se obtiene eligiendo en las **Opciones de ventana,** *Escala:* **porcentajes,** *Dirección:* **horizontal,** *Tipo de diagrama:* **adosado,** y muestra tanto los porcentajes absolutos de las celdas como los totales de las filas. El Gráfico de mosaico muestra los porcentajes relativos de cada fila.

Como se comentó al principio de este apartado, el procedimiento mostrado sirve para la descripción conjunta de dos variables de cualquier naturaleza. Si una o las dos variables son cuantitativas continuas es preciso categorizar dicha o dichas variables para que resulte útil la tabla de frecuencias. Por ejemplo, si queremos describir la localización del tumor según la edad vamos a categorizar las edades agrupando las mismas en lustros, < 50 años, 50-55 años, 55-60 años,_, 75-80 años, > 80 años, llamaremos a esta nueva variable edad agrupada. La Figura C2.14 presenta el resultado obtenido al elegir la opción tabular **Tabla de frecuencias,** en ella se muestra la distribución conjunta de frecuencias y porcentajes de ambas variables.

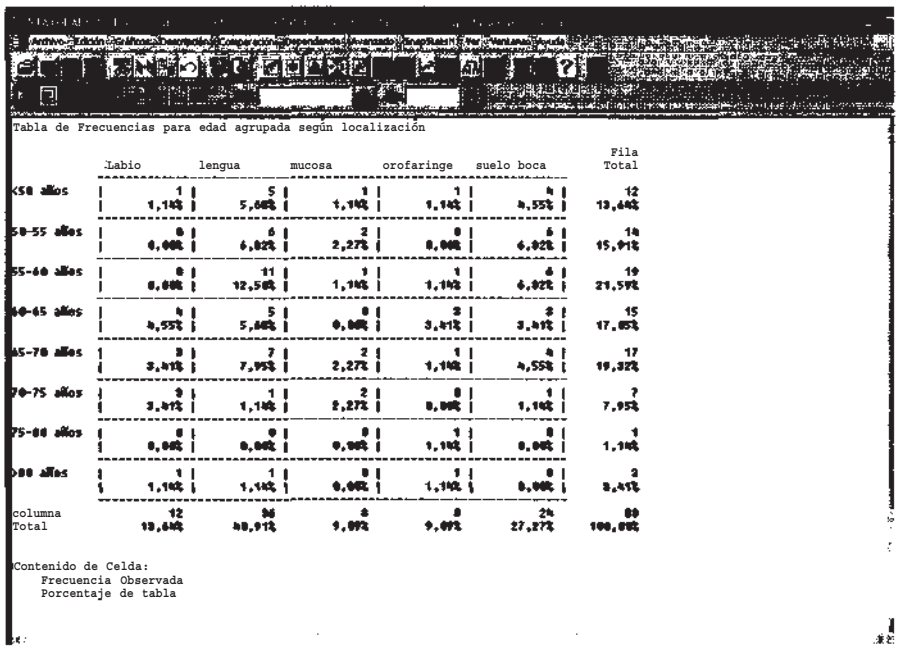


Tabla de Frecuencias para edad agrupada según localización

	Labio	lengua	mucosa	orofaringe	suelo boca	Fila Total
<50 años	1 1,14%	5 5,68%	1 1,14%	1 1,14%	4 4,55%	12 13,64%
50-55 años	0 0,00%	6 6,82%	2 2,27%	0 0,00%	6 6,82%	14 15,91%
55-60 años	0 0,00%	11 12,50%	1 1,14%	1 1,14%	6 6,82%	19 21,59%
60-65 años	4 4,55%	5 5,68%	0 0,00%	3 3,41%	3 3,41%	15 17,05%
65-70 años	3 3,41%	7 7,95%	2 2,27%	1 1,14%	4 4,55%	17 19,32%
70-75 años	2 2,27%	1 1,14%	2 2,27%	0 0,00%	1 1,14%	7 7,95%
75-80 años	0 0,00%	0 0,00%	0 0,00%	1 1,14%	0 0,00%	1 1,14%
>80 años	1 1,14%	1 1,14%	0 0,00%	1 1,14%	0 0,00%	3 3,41%
columna Total	12 13,64%	36 40,91%	8 9,09%	8 9,09%	24 27,27%	88 100,00%

Contenido de Celda:
Frecuencia Observada
Porcentaje de tabla

Figura C2.14. Tabla de frecuencias de doble entrada: edad y localización.

Teniendo en cuenta el carácter cuantitativo de la variable edad podemos continuar el análisis obteniendo, para cada localización, medidas centrales, de dispersión y forma, así como gráficos de frecuencias. Para ello, volvemos a la variable original, edad, y elegimos la secuencia de procedimientos, **Descripción, Datos numéricos, Análisis de subgrupo.** La Figura C2.15 muestra las opciones tabulares y gráficas recomendadas. De todas ellas ya se ha comentado su significado salvo de la opción gráfica **Gráfico de medias.** En esta última se muestra, para cada localización, una barra centrada en la media y con extremos los límites de confianza al 95%. La línea que une las barras se apoya en dichos valores medios. El concepto y significado de los límites de confianza se presenta en el Capítulo 6.

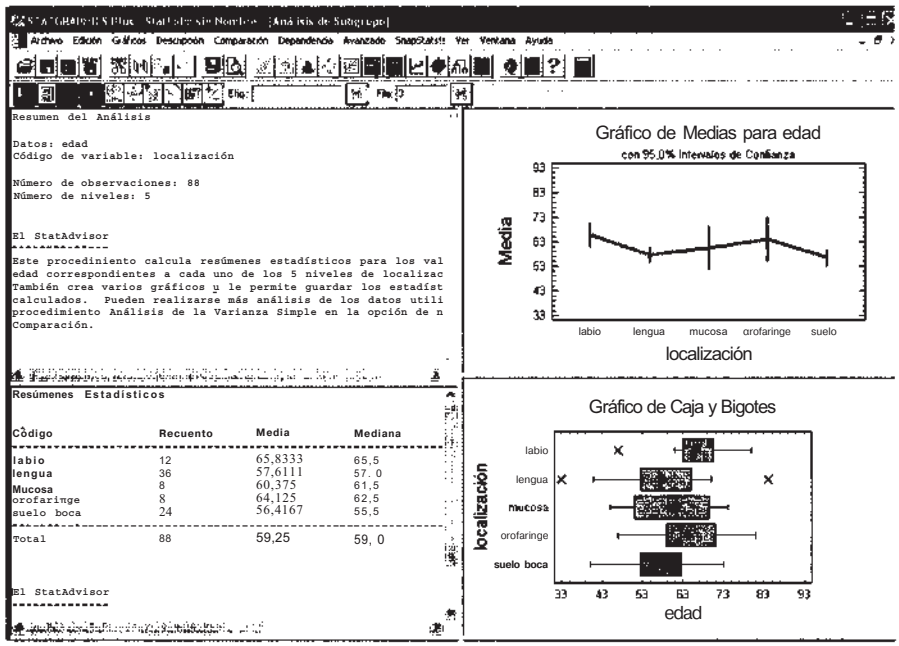


Figura C2.15. Pantalla de resultados correspondiente a la edad según localización.

C2.5. SELECCIÓN DE UNA SUBMUESTRA DE DATOS

En todos los procedimientos mostrados en este Apéndice C2 hemos utilizado la muestra completa de los 88 casos contenidos en la muestra original. Si deseamos describir, únicamente, una parte de estos datos, una vez abierta la ventana del procedimiento elegido, indicar en el campo *Selección:* la submuestra o grupo de datos que se desea describir. La Figura C2.16 ilustra, para el procedimiento **Descripción, Datos numéricos, Análisis unidimensional** y la variable **edad,** cómo seleccionar los pacientes cuya localización tumoral es la lengua. Es preciso añadir que la variable de selección debe estar codificada numéricamente, por ello, para llevar a cabo la selección que muestra la Figura C2.16, debemos partir de un fichero de datos donde la localización del tumor esté codificada como 1, 2,.., 6, siendo el valor 3 el asignado a la localización en la lengua.

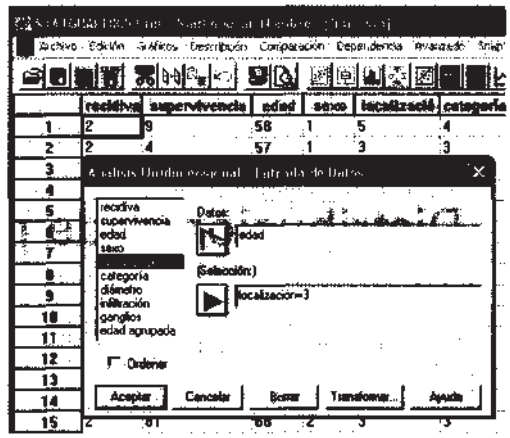
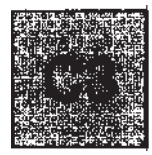


Figura C2.16. Ventana del procedimiento Análisis unidimensional de la variable edad, para el subgrupo de pacientes con Idealización del tumor en la lengua.



Distribuciones de probabilidad

Este apartado está dedicado al estudio de algunas de las principales distribuciones de probabilidad que suelen aparecer en las disciplinas relacionadas con las Ciencias de la Vida. El STATGRAPHICS Plus, no solo nos permite calcular valores de las funciones de distribución, densidad y/o masa, que incluyen y complementan las dadas en el Apéndice B, sino que permite un estudio gráfico detallado que creemos enriquece el conocimiento de tales distribuciones. Además incluimos, al final de este apartado, el estudio gráfico, y mediante test de hipótesis, del ajuste de unos datos a una distribución teórica dada.

C3.1. CARACTERÍSTICAS GENERALES DE LOS PROCEDIMIENTOS

El STATGRAPHICS Plus permite el estudio de las diferentes distribuciones de probabilidad, qué aparecen en los Capítulos 4 y 5, para ello bastará elegir la siguiente secuencia de procedimientos **Descripción, Distribuciones, Distribuciones de probabilidad** y emergerá un cuadro con las distribuciones de probabilidad disponibles (véase Fig. C3.1).

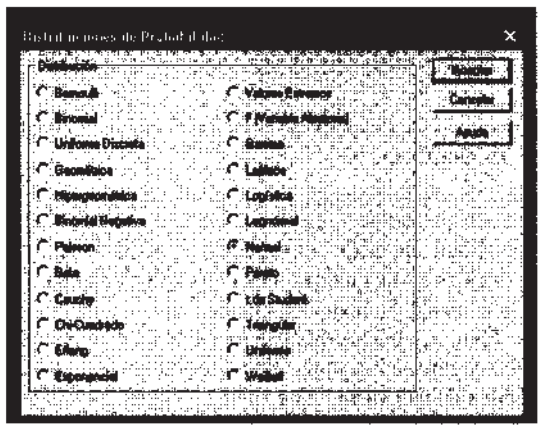


Figura C3.1. Distribuciones disponibles en STATGRAPHICS Plus.

En la Figura C3.1 vemos que, por defecto, viene marcada la distribución normal, pero cambiando el punto tendremos la distribución que nos interese en cada caso. Veamos primero las partes comunes a todas las distribuciones, para después pasar a estudiar las distribuciones binomial, de Poisson y normal vistas en el texto. Comenzamos describiendo las opciones de los principales iconos de la barra de ejecución una vez marcada la distribución elegida y pulsado **Aceptar.**

Dentro de las cuatro opciones tabulares **(Resumen del análisis, Distribución acumulada, CDF inverso, Números aleatorios),** las dos primeras se ejecutan por defecto:

- **Resumen del análisis:** Describe los parámetros de las distribuciones sobre las que se está haciendo el análisis. Si se pulsa el botón derecho del ratón y después **Opciones de análisis** se podrán cambiar dichos parámetros.
- **Distribución acumulada:** En esta opción se calculan tres valores que comentamos a continuación. Para clarificar nombres y contenido diferenciamos según la naturaleza de la variable.

Sea X una variable discreta, sea c un posible valor y F la función de distribución (véase Cap. 4):

- **Área de cola inferior** (<): Calcula $P(X < c)$, es decir, la probabilidad de ser menor estrictamente que dicho valor c. Hay que hacer notar que al ser X discreta, esta cantidad no es la función de distribución, pues la función de distribución es la probabilidad de ser menor o igual a c, $F(c) = P(X < c)$.
- **Probabilidad de masa** (=): Calcula $P(X = c)$, en el Capítulo 4 esta cantidad se denomina función de densidad, en otros textos se denomina función de masa, siendo desafortunada la denominación de probabilidad de masa, que parece más bien un error de traducción, pues hay que hacer notar que, un poco más abajo, en el StatAdvisor se utiliza el término de función de masa.
- **Área de cola superior** (>): Calcula $P(X > c)$, es decir, probabilidad de ser mayor que c.

Sea X una variable continua, sea c un posible valor, f la función de densidad y F la función de distribución (véase Cap. 5):

- **Área de cola inferior** (<): Calcula $P(X < c)$. Hay que hacer notar que al ser X continua, esta cantidad sí es la función de distribución, pues la probabilidad de un punto vale 0 con lo que $F(c) = P(X < c) = P(X < c)$.
- **Densidad de probabilidad:** Calcula el valor de la función de densidad en el valor c, f(c). Hay que hacer notar que esta cantidad no es una probabilidad. Esta cantidad se denomina habitualmente función de densidad, también en este caso parece desafortunada la denominación de Densidad de probabilidad pues en el mismo StatAdvisor se denomina esta cantidad con el término de función de densidad .
- **Área de cola superior** (>): Calcula $P(X > c) = 1\text{-}F(c)$, esta cantidad se denomina Función de supervivencia.

El valor, por defecto de c es 0, para cambiar el valor de c basta con pulsar el botón derecho del ratón y después **Opciones de ventana.**

- **CDF inverso:** Para una probabilidad que le señalemos nos devuelve la abscisa que deja a su izquierda dicha probabilidad, de ahí el nombre de inverso de la función de distribución, pues para una variable X y una probabilidad p nos da el valor de la abscisa c, de forma que $P(X < c) = p = F(c)$, siendo F la función de distribución.
- **Números aleatorios:** Genera 100 valores de la variable que le indiquemos, si queremos que no sean 100 los números generados podemos cambiar dicho número pulsando el botón derecho del ratón y **Opciones de ventana.** Para que dichos datos se graben en la pantalla de datos debemos pulsar Guardar resultados de la barra de ejecución e indicar el nombre de la variable donde se guardarán.

De las cinco opciones gráficas de la barra de ejecución, sólo nos interesan las dos primeras, que: aparecen por defecto y dibujan la función de densidad y de distribución (CDF).

C3.2. DISTRIBUCIÓN BINOMIAL B(n, p)

Opciones gráficas y cálculo de probabilidades

Supongamos que queremos dibujar las funciones de masa y distribución para una variable X Binomial de parámetros 8 y 0.25, así como la probabilidad de que valga 5, P(X = 5).

Elegimos el procedimiento **Descripción, Distribuciones, Distribuciones de probabilidad,** cambiamos la opción por defecto que es *Normal* por *Binomial,* para ello ponemos el cursor en Binomial, si ahora pulsamos **Aceptar** obtenemos una salida que corresponde a la distribución que viene por defecto (Binomial 10 y 0.1), para cambiar dichos parámetros debemos pulsar el botón derecho del ratón y **Opciones de análisis,** emergerá entonces una ventana con dos campos. En el campo *Probabilidad de evento* ponemos **0.25** y en el campo *Ensayos* ponemos 8, pulsando **Aceptar** tendremos ahora la salida de la Figura C3.2.

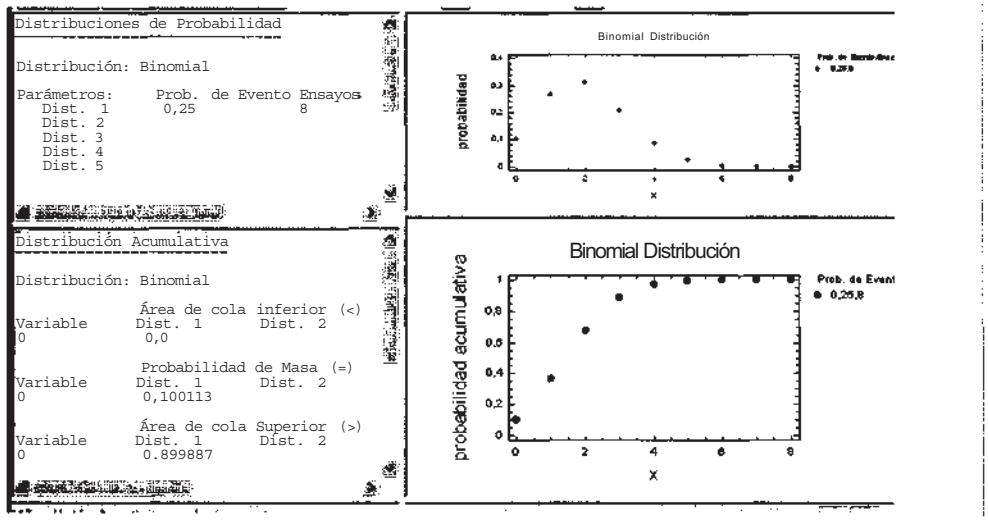


Figura C3.2. Resultados para la variable Binomial de parámetros 8 y 0.25.

Observando la Figura C3.2 vemos que la pantalla está dividida en cuatro partes, en la superior) izquierda hay una descripción de la distribución estudiada, en la superior derecha aparece un gráfico para la función de masa de la Binomial (8, 0.25), en el que se representa para cada valor de X cuánto vale su probabilidad, en la parte inferior derecha aparece el gráfico para la función de distribución. En la parte inferior izquierda se muestran para X, con distribución B(8,0.25), las siguientes probabilidades P(X < 0) = 0.0, P(X = 0) = 0.100113 y P(X > 0) = 0.899887.

Para calcular P(X = 5), tenemos que posicionarnos en la parte inferior izquierda de la pantalla, botón derecho del ratón y en **Opciones de ventana** cambiaremos el valor 0 por 5, obteniendo que P(X = 5) = 0.0230713.

C3.3. DISTRIBUCIÓN DE POISSON *P(λ)*

Opciones gráficas y cálculo de probabilidades

Supongamos que queremos dibujar las funciones de masa de las variables de Poisson de parámetros 1, 8 y 15 respectivamente, además calcular la probabilidad en 0, para las tres variables.

Elegimos el procedimiento **Descripción, Distribuciones, Distribuciones de probabilidad,** cambiamos la opción por defecto que es *Normal* por *Poisson,* para ello ponemos el cursor en Poisson, si ahora pulsamos **Aceptar** obtenemos una salida que corresponde a la distribución que viene por defecto (Poisson 10), para cambiar este parámetro debemos pulsar el botón derecho del ratón y **Opciones de análisis,** emergerá entonces una ventana, donde indicaremos que **Media** es 1, 8 y 15, pulsando **Aceptar** tendremos para **Distribución acumulada** de las opciones tabulares y **Densidad/función de masa** de la opciones gráficas, la salida de la Figura C3.3. Si nos fijamos en la ventana de la izquierda podemos ver los valores de la probabilidad en 0, para $\lambda = 1$ (Dist.l) es 0.367879, para $\lambda = 8$ (Dist.2) es 0.000335 y para $\lambda = 15$ (Dist.3) es 0.0000003. Si nos fijamos en la ventana de la derecha, vemos cómo la gráfica de la función de masa tiene forma más «acampanada» según aumenta el valor del parámetro.

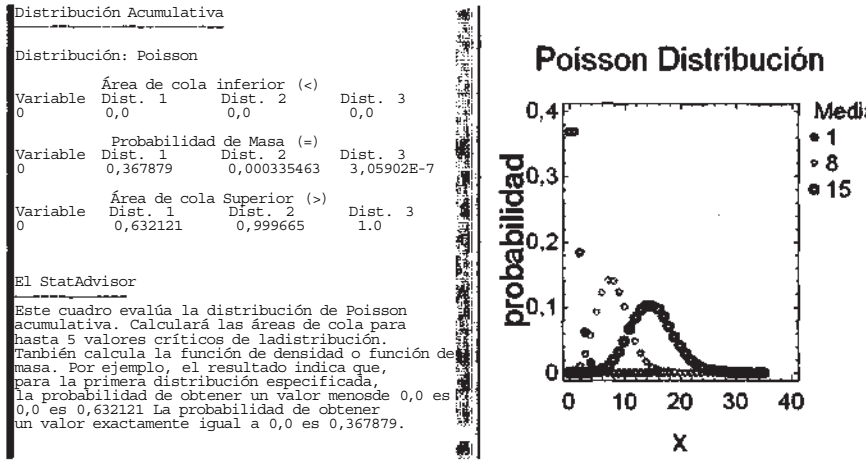


Figura C3.3. Resultados para las variables de Poisson de parámetros 1, 8 y 15.

Aproximaciones

La distribución binomial (n, p) puede aproximarse por la distribución de Poisson ($np = \lambda$), veamos en algunos ejemplos cómo es esta aproximación. Según las recomendaciones del texto si $n \geq 20$ y $p < 0.05$ es buena y si $n \geq 100$ y $np \leq 10$ muy buena. Calculemos con la binomial algunas probabilidades y luego con la correspondiente de Poisson. En la Tabla C3.1 damos los valores calculados

Tabla C3.1. Aproximación binomial por Poisson

x	B(20,0.05)	B(100,0.01)	B(1000,0.001)	B(10000,0.0001)	P(1)
0	0.35848	0.36603	0.36676	0.36786	0.36787
1	0.37735	0.36973	0.36806	0.36789	0.36787
2	0.18867	0.18486	0.18403	0.18394	0.18394
3	0.05958	0.06099	0.06128	0.06131	0.06131

con STATGRAPHICS Plus de las probabilidades, para cuatro binomiales (probabilidades exactas, B(n, p)) de algunos valores de la variable (0, 1, 2 y 3) y las probabilidades, para estos valores, con la correspondiente distribución de Poisson (probabilidades aproximadas, $P(\lambda = np)$).

Hay que observar que todas las binomiales de la Tabla C3.1 se aproximan por la de Poisson de parámetro $\lambda = 1$, siendo mejor la aproximación según crece n y decrece p.

Ajuste gráfico

Otro aspecto importante que se puede comprobar con STATGRAPHICS Plus es ver si unos datos se parecen a una distribución dada gráficamente, esto nos permitirá unir los conceptos vistos en los capítulos de descriptiva y distribuciones de probabilidad. Consideremos el siguiente ejemplo: se quiere saber si la distribución de ciertas bacterias en un cultivo sigue una distribución de Poisson. Para ello se divide el cultivo en 576 áreas (24 x 24) y se anota el número de bacterias por área, siendo los resultados dados en la Tabla C3.2:

Tabla C3.2. Número de bacterias por área

N.° bacterias	0	1	2	3	4
N.° **áreas**	229	211	93	35	7

Elegimos el procedimiento **Descripción, Distribuciones, Ajuste de distribuciones (datos no censurados)** y cambiamos la opción por defecto que es *Normal* por *Poisson,* ahora el cuadro superior izquierdo nos indica que la distribución de Poisson ajustada es de media 0.921739. Este valor es la media de los datos de la Tabla C3.2. Para poder visualizar el ajuste, basta seleccionar las opciones gráficas de la barra de ejecución y marcar **Histograma de frecuencias;** después, con el botón derecho en **Opciones de ventana** cambiamos *N.º de clases:* 5, *Límite inferior.* -0.5 y *Límite superior.* 4.5, obteniéndose el gráfico de la Figura C3.4. En la Figura C3.4, aparece un gráfico donde las alturas de los rectángulos representan las frecuencias de los datos de la Tabla C3.2 y el punto en cada rectángulo indica el valor de la función de masa para la distribución de Poisson P(0.921739). La similitud de la altura de los rectángulos y dichos puntos nos sugiere que la distribución del número de bacterias por área sigue una distribución de Poisson, aunque este hecho debería corroborarse con un test de hipótesis. Véase NOTA, al final de este Apéndice C3.

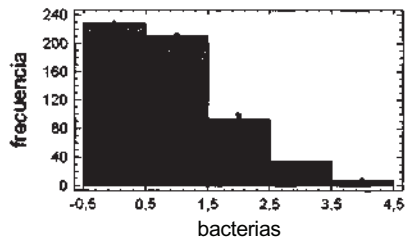


Figura C3.4. Ajuste gráfico.

C3.4. DISTRIBUCIÓN NORMAL (μ, σ)

Opciones gráficas

Para el estudio de las distribuciones normales comencemos por ver cómo varía la función de densidad de la normal al variar la media y desviación típica, para ello dibujamos la función de densi-

dad de tres distribuciones: N(5, 1), N(7, 1) y N(5, 3) (véase Fig. C3.5). Para obtenerla elegimos el procedimiento **Descripción, Distribuciones, Distribuciones de probabilidad,** por defecto es *Normal,* si ahora pulsamos **Aceptar** obtenemos la opción por defecto que es la N(0, 1) y entre los gráficos está la función de densidad, ahora debemos cambiar los parámetros de la normal, para ello pulsar el botón derecho del ratón y **Opciones de análisis,** allí cambiaremos los parámetros media y desviación típica, para las tres distribuciones.

Observando la Figura C3.5 vemos cómo al estar la distribución normal centrada en la media un cambio en el primer parámetro, la media, produce un «deslizamiento» sobre el eje X, mientras que una variación en el segundo parámetro, la desviación típica, produce un cambio en el apuntamiento en la función de densidad.

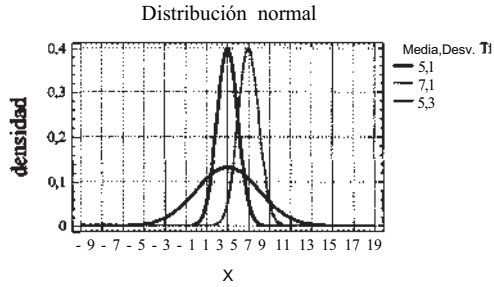


Figura C3.5. Funciones de densidad para normales variando media y desviación típica.

Cálculo de la función de distribución

Para calcular probabilidades de distribuciones normales con STATGRAPHICS Plus no necesitamos tipificar, por ejemplo si X sigue una distribución N(5,3) y queremos calcular $P(2 < X < 4)$ bastará con elegir el procedimiento **Descripción, Distribuciones, Distribuciones de probabilidad,** si ahora pulsamos **Aceptar** obtenemos la opción por defecto que es la N(0, 1) y entre las Opciones tabulares, está **Distribución acumulada,** que aparece en la parte inferior izquierda de la pantalla, para cambiar los parámetros de la normal debemos pulsar el botón derecho del ratón y **Opciones de análisis,** allí cambiaremos el valor de *Media* a 5 y *Desv.Tip.* a 3. Ahora pulsamos el botón derecho del ratón y en **Opciones de ventana** ponemos nuestros valores 2 y 4.

Obtendríamos ahora los valores de la Figura C3.6 con lo que estaríamos en condiciones de calcular la probabilidad pedida

$$P(2 < X < 4) = P(X < 4) - P(X < 2) = 0.369439 - 0.158655 = 0.210784$$

Otra forma alternativa de calcular la probabilidad anterior sería

$$P(2 < X < 4) = P(X > 2) - P(X > 4) = 0.841345 - 0.630561 = 0.210784$$

```
Distribución: Normal

              ¿rea de cola inferior (<)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
2             0.158655
4             0.369439

              Densidad de Probabilidad
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
2             0.0806569
4             0.125794

              Área de cola Superior (>)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
2             0.841345
4             0.630561
```

Figura C3.6. Probabilidades para N(5, 3).

Cálculo de abscisas

Supongamos ahora que queremos saber cuál es la abscisa de una variable X con distribución N(8, 2) que deja a su izquierda una probabilidad de 0.025, bastará con elegir el procedimiento **Descripción, Distribuciones, Distribuciones de probabilidad,** si ahora pulsamos **Aceptar** obtenemos por defecto la N(0, 1), para cambiar los parámetros de la normal debemos pulsar el botón derecho del ratón y **Opciones de análisis** y allí cambiaremos la *Media* a 8 y *Desv.Tip.* por 2, después entre las opciones tabulares, está **CDF inverso,** que calcula las abscisas para ciertas probabilidades fijadas por defecto, con el botón derecho del ratón y **Opciones de ventana** cambiamos una de las probabilidades por 0.025 y obtenemos $P(X < 4.08006) = 0.025$.

Ajuste gráfico

Al igual que vimos para las distribuciones discretas también podemos hacer un «ajuste visual» de unos datos a una distribución normal. Usaremos los datos del Ejemplo C2.1 del Apéndice C2, en concreto, veamos si la variable edad sigue gráficamente una distribución normal.

Al elegir el procedimiento **Descripción, Distribuciones, Ajuste de distribuciones (datos no censurados),** la opción por defecto es *Normal.* La normal ajustada por defecto es la que tiene la misma media y desviación típica que los datos, es decir, N(59.25, 9.67). La salida está en la Figura C3.7, donde se ve que el histograma de los datos se parece a una distribución normal, aunque este hecho debería corroborarse con un test de hipótesis. Véase NOTA al final de este Apéndice C3.

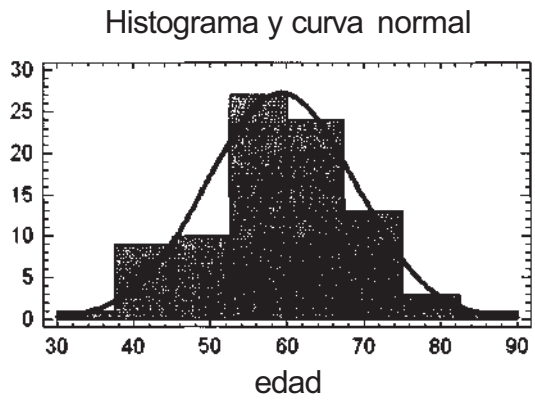


Figura C3.7. Ajuste normal gráfico.

NOTA: Test de bondad de ajuste

En el Capítulo 13, en el apartado 13.1, se comenta una prueba gráfica para ver si unos datos son normales. No se da ningún test estadístico para comprobar si los datos siguen una distribución dada, pero sí se comenta cómo los paquetes estadísticos realizan al menos uno de ellos. El STATGRAPHICS Plus realiza varios. Si elegimos **Descripción, Distribuciones, Ajuste de distribuciones (datos no censurados)** y nos fijamos en las opciones tabulares una opción es **Test de bondad de ajuste,** en dicha opción se realiza al menos un contraste para las distribuciones de la Figura C3.1, en el caso de la normal se pueden pedir hasta 7 test diferentes. Además otra opción de las opciones tabulares, **Test para la normalidad,** calcula otros tres diferentes. El estudio y comparación de estos 10 tests no son objetivo de este Apéndice. La prueba de Lilliefors, comentada en el Capítulo 13, se basa en la comparación de la función de distribución muestral con la teórica, esta misma idea es la que se utiliza, por ejemplo, en el test de Kolmogorov que se realiza en la opción **Test de bondad de ajuste.** En concreto para la variable edad de la Figura C3.7, el valor de p para el test de Kolmogorov es 0.971618, con lo que se aceptaría la hipótesis de normalidad de la variable edad.

Inferencia sobre los parámetros de una población

En este Apéndice vemos los resultados relacionados con la inferencia de los parámetros de la distribución binomial y normal, en el caso de una sola población. Dichos resultados se corresponden con los vistos en los Capítulos 6 y 7.

En el caso de una sola población el STATGRAPHICS Plus permite hacer inferencia sobre los parámetros que la caracterizan, tanto si tenemos el fichero con todos los datos como si los datos ya están resumidos en los valores de las estimaciones de dichos parámetros.

C4.1. PROCEDIMIENTOS PARA DATOS RESUMIDOS

Comencemos viendo el caso en el que los datos están resumidos. En esta situación, los parámetros para los que se puede hacer inferencia con STATGRAPHICS Plus, como se puede ver en la Figura C4.1, son los de la normal, binomial y Poisson. En este procedimiento, se nos permitirá hacer intervalos de confianza y test de hipótesis, tanto bilaterales como unilaterales. Resolvamos algunas situaciones.

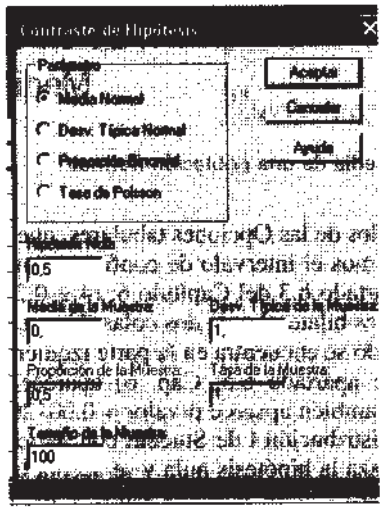


Figura C4.1. Parámetros de una población.

Parámetros de la distribución normal

Si revisamos el Ejercicio 7.1.6 del Capítulo 7, vemos que se refiere a un estudio sobre la edad de comienzo de la obesidad en niños menores de 12 años. Para ello, se toma una muestra de 100 niños obesos para los que se obtiene una media de 4 años y desviación típica muestral s = 1.5 años. Se piden intervalos de confianza al 95% para la media μ y varianza σ^2. Suponiendo que la edad sigue una distribución normal, añadamos una pregunta más que nos permita comparar estos valores con los de otro país, A, donde la edad media es de 3.5 años, utilizaremos como nivel de significación a= 0.05.

Media de la normal

Para resolver el ejercicio con STATGRAPHICS Plus elegimos el procedimiento **Descripción, Contraste de hipótesis** y aparece por defecto marcada la media de una distribución normal (véase Fig. C4.1). Como vamos a hacer un contraste, ponemos en el campo **Hipótesis nula** el valor 3.5. En el campo **Media de la muestra** ponemos el valor 4 y en el campo **Desviación típica de la muestra** el valor 1.5. Como nuestro tamaño muestral es 100, que es el que tiene por defecto el campo **Tamaño de la muestra,** ya podemos pulsar **Aceptar.** Se ejecutan las opciones por defecto de las Opciones tabulares y Opciones gráficas de la barra de ejecución (véase Fig. C4.2).

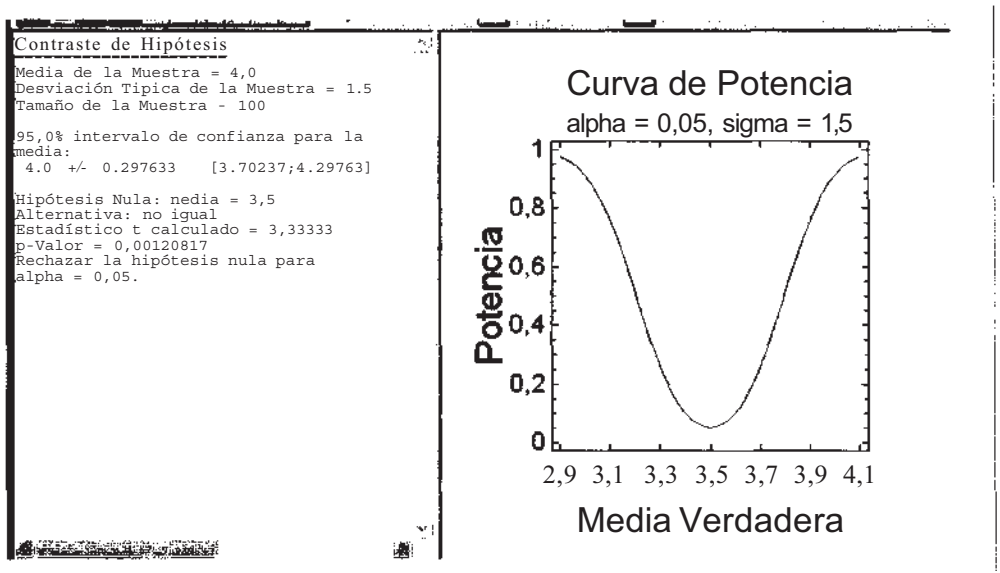


Figura C4.2. Inferencia sobre la media de una población normal.

Si nos fijamos en los resultados de las Opciones tabulares, que por defecto son un intervalo y el test bilateral para α = 0.05, tenemos el intervalo de confianza para la media al nivel de confianza del 95%, calculado según el apartado 6.3 del Capítulo 6: (4 ± 0.297633) = (3.70237, 4.29763). El contraste que queremos realizar es bilateral o de dos colas $(H_0 : \mu = 3.5, H_1 : \mu \neq 3.5)$ que es el que se realiza por defecto. El resultado se encuentra en la parte izquierda de la Figura C4.2, el valor del estadístico del contraste (véase apartado 6.5, Cap. 6) aparece en la ventana bajo el nombra Estadístico t calculado, 3.333. También aparece p-valor = 0.00120817, este valor correspondería a 2P(T > 3.3333), siendo T una distribución t de Student con 99 grados de libertad. Por lo tanto, al ser el valor de p < 0.05, se rechaza la hipótesis nula y se acepta la alternativa al nivel α = 0.05, es decir, existe diferencia significativa entre nuestros datos y los del país A. Estos resultados se comentan en el StatAdvisor, en la parte izquierda de la pantalla.

En la parte derecha de la Figura C4.2 aparece la curva de potencia para este ejercicio. Se define la potencia en un contraste como 1-p (error de tipo II) (véase apartado 6.6, Cap., 6, Tamaño de la muestra: contraste de hipótesis). Un comentario sobre el significado de la curva de potencia puede verse en el StatAdvisor. Para ver el StatAdvisor, basta con posicionarnos sobre la gráfica, hacer doble clic con el botón izquierdo del ratón y la ventana de la derecha ocupará toda la pantalla; ahora ya se puede maximizar el StatAdvisor que está en la parte baja de la pantalla.

Varianza de la normal

Elegimos el procedimiento **Descripción, Contraste de hipótesis** y aparece por defecto marcada la media de una distribución normal, que habrá que cambiar por desviación típica normal (véase Fig. C4.1). Como no vamos a hacer un contraste sobre la varianza, dejamos en el campo **Hipótesis nula** el valor por defecto (0.5). Ahora sólo está activo el campo **Desviación típica de la muestra,** donde pondremos el valor 1.5. Como nuestro tamaño muestral es 100, que es el que tiene por defecto el campo **Tamaño de la muestra,** no hay que cambiar nada. Si se pulsa **Aceptar,** se ejecutan las opciones por defecto de las Opciones tabulares y Opciones gráficas de la barra de ejecución; sólo nos interesan las Opciones tabulares. Se ha calculado un intervalo y realizado un contraste que no nos interesa ($H_0 : \sigma = 0.5, H_1 : \sigma \neq 0.5$); por lo tanto, sólo nos fijamos en los resultados que se corresponden con el intervalo de confianza del apartado 7.1 del Capítulo 7. El intervalo que se obtiene en la Figura C4.3 es para σ, (1.31701, 1.74251). Como el intervalo que nos piden es para la varianza σ^2, bastará con calcular los cuadrados de ambos extremos. Resulta, por lo tanto, que el intervalo de confianza al 95% para la varianza es $(1.31701^2, 1.74251^2) = (1.73451, 3.03634)$.

```
Contraste de Hipótesis
----------

Desviación Típica de la Muestra = 1.5
Tamaño de la Muestra = 100

95.0% intervalo de confianza para sigma: [1,31701;1.74251]

Hipótesis Nula: desviación típica = 0,5
Alternativa: no igual
Estadístico Chi-cuadrado calculado = 891.0
p-Valor =0,0
Rechazar la hipótesis nula para alpha = 0.05.
```

Figura C4.3. Inferencia para la desviación típica de una población normal.

Revisemos ahora un ejemplo de contraste unilateral (con una cola). El Ejercicio 6.5.7 del Capítulo 6 es otro ejemplo donde los valores de la media y desviación típica muestral vienen dados en el enunciado. Se estudia cómo el ejercicio físico puede tener un efecto beneficioso en la reducción del colesterol. Se piensa que el ejercicio reducirá la media del colesterol en más de 25 puntos. Los datos muestrales relativos a la reducción del colesterol dan unos valores para la media de 27 y para la desviación típica de 18, en una muestra de tamaño 80. Si pensamos que los datos corroboran la hipótesis del estudio haremos un test unilateral suponiendo la normalidad, $H_0 : \mu = 25$ frente a $H_0 : \mu \neq 25$.

Para hacer dicho test, elegimos **Descripción, Contraste de hipótesis** y aparece por defecto marcada la media de una distribución normal. Como vamos a hacer un contraste, ponemos en el campo **Hipótesis nula** el valor 25. En el campo **Media de la muestra** pondremos el valor 27; en el campo **Desviación típica de la muestra,** el valor 18 y en el campo **Tamaño de la muestra,** pondremos 80. Si pulsamos **Aceptar,** se ejecutan las opciones por defecto de las Opciones tabulares y Opciones gráficas de la barra de ejecución, como en la Figura C4.2. Por defecto, el test que se realiza es bilateral, así que para cambiarlo nos ponemos en la ventana de las Opciones tabulares (parte izquierda de la pantalla) y haciendo clic con el botón derecho, aparecen **Opciones de análisis** donde se puede cambiar la hipótesis alternativa; por defecto, la hipótesis alternativa es *No igual* (test bilateral o con

dos colas), y tendremos que cambiar a *Mayor que.* También en esa ventana se debería cambiar el nivel α si fuésemos a hacer un intervalo de confianza a nivel distinto del 95% (1-α), que es el valor por defecto. En nuestro caso, esto no importa para el test. Los resultados se dan en la Figura C4.4

```
Contraste de Hipótesis
----------  ----------

Media de la Muestra = 27,0
Desviación Tipica de la Muestra = 18,0
Tamaño de la Muestra = 80

95,0% inferior límite de confianza para la media:  27,0 - 3,34949    [23,6505]

Hipótesis Nula: media = 25,0
Alternativa: mayor que
Estadístico t calculado = 0,993808
p-Valor = 0,161675
Ho rechazar la hipótesis nula para alpha = 0.05.
```

Figura C4.4. Resultados del test unilateral.

Observando la Figura C4.4 vemos el valor del estadístico del contraste (véase apartado 6.5, Cap. 6) que aparece en la ventana como estadístico t calculado = 0.993808. Vemos también que aparece el valor de P = 0.161675, que correspondería a la P(T > 0.993808), siendo T una distribución t de Student con 79 grados de libertad.

Por lo tanto, al ser el valor de p > 0.05, no se rechaza la hipótesis nula al nivel $a = 0.05$, es decir, no existe evidencia estadísticamente significativa de que el ejercicio disminuya más de 25 unidades el nivel medio del colesterol.

Tamaño muestral

El STATGPvAPHICS Plus permite calcular el tamaño muestral adecuado para un contraste de hipótesis sobre los parámetros de una distribución, dichos parámetros son los de la Figura C4.5. Según se puede ver en el apartado 6.6 del Capítulo 6, sección Tamaño de la muestra: contraste de hipótesis, en un test intervienen cuatro elementos relacionados:

1) α, que es la probabilidad del error de tipo I,
2) potencia, que es 1- β, siendo β la probabilidad del error de tipo II,
3) D, diferencia que se quiere detectar en relación al valor de la hipótesis nula,
4) n, el tamaño muestral.

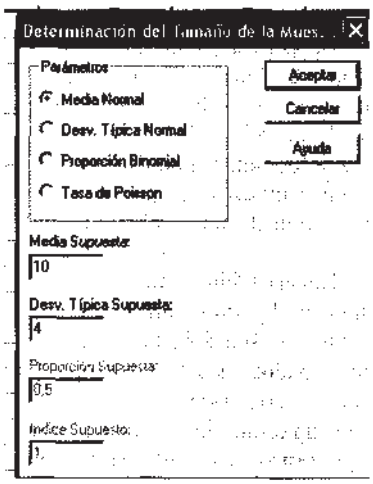


Figura C4.5. Parámetros y tamaño muestral.

Revisemos el Ejemplo 6.6.4 del Capítulo 6 y veamos cómo resolverlo con STATGRAPHICS Plus. En dicho ejemplo se quiere hacer un test unilateral $(H_0 : \mu = 10, H_1 : \mu \neq 10)$ a un nivel de significación $\alpha = 0.05$, potencia $1 - \beta = 0.90 = 1 - 0.1$, y la diferencia que queremos detectar con la hipótesis nula es de 2 unidades (pues el texto dice una media de 12, es decir 10 + 2).

Eligiendo el procedimiento **Descripción, Tamaño de la muestra,** aparece la ventana de la Figura C4.5, donde por defecto está marcada **Media normal,** que es lo que nos interesa. En el campo **Media supuesta** ponemos el valor de hipótesis nula, en este caso 10, en el campo **Desviación típica supuesta** ponemos la estimación de la desviación típica, en este caso 4, si ahora pulsamos **Aceptar** aparece la pantalla de la Figura C4.6. Marcamos ahora **Potencia;** en su campo ponemos 90 y en el campo **Diferencia** ponemos 2; el campo **Nivel de confianza** (1-α) no hay que cambiarlo pues el nivel de significación α es 0.05 y en el campo **Hipótesis alternativa** hay que poner *Mayor que* (véase Fig. C4.6).

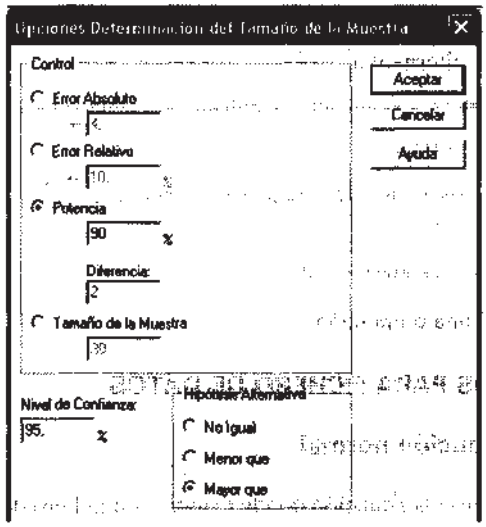


Figura C4.6. Cambios de opciones en tamaño muestral.

Al pulsar **Aceptar** en la Figura C4.6, llegamos a la Figura C4.7, donde vemos que el tamaño es 36, el mismo valor que se obtiene consultando la Tabla 6.7 del Capítulo 6.

Análogamente, se puede proceder para los otros parámetros de la Figura C4.5.

```
Determinación del Tamaño de la Muestra

Parámetro a estimar: media normal
Potencia deseada: 90.0% para la media = 10,0 frente a media = 12.0
Tipo de alternativa: mayor que
Riesgo para alpha: 5.0%
Sigma asumida: 4,0

El tamaño de la muestra requerido es de n=36 observaciones.
```

Figura C4.7. Obtención del tamaño muestral para la media de la normal.

Parámetro p de la distribución binomial

El STATGRAPHICS Plus permite hacer inferencia sobre el parámetro p de la binomial, cuando el tamaño muestral es grande, pues realiza el test e intervalo vistos en el Capítulo 8. Como regla de «suficientemente grande» se pueden dar diversas, en concreto nosotros lo aplicaremos según las reglas dadas en el apartado 13.9 del Capítulo 13: valores de n y p tales que $p < 0.5$ y $np > 5$ ó $p > 0.5$ y $n(1 - p) > 5$.

Resolvamos el Ejercicio 8.4.4 (Cap. 8), donde se plantea que el 90% de los pacientes con cáncer de pulmón mueren en 3 años. Se quiere estudiar un nuevo tratamiento para ver si se mejoran los resultados, es decir, si el porcentaje de muerte disminuye. Para ello se han tratado con el nuevo tratamiento 150 pacientes de cáncer de pulmón y se ve que murieron 128 (128/150 = 0.85333).

El contraste a realizar es $H_0 : p = 0.9$, $H_1 : p < 0.9$. Para hacer este contraste con STATGRAPHICS Plus, elegimos el procedimiento **Descripción, Contraste de hipótesis,** y aparecerá una ventana como la de la Figura C4.1, en la que está marcada **Media normal,** y habremos de cambiarla a **Proporción binomial.** Debemos rellenar tres campos: **Hipótesis nula** con **0.9, Proporción en la muestra** con 0.85333 y **Tamaño de la muestra** con 150. Ahora al pulsar **Aceptar,** se realizan las opciones por defecto, es decir, un test bilateral, por lo tanto, para cambiarlo, tendremos que ponernos en la parte izquierda de la pantalla, botón derecho del ratón, **Opciones de análisis** y cambiar a la alternativa de *Menor que,* al pulsar **Aceptar,** aparece en pantalla la Figura C4.8. En esta opción del procedimiento, no aparece en pantalla el valor del estadístico del contraste que calcula, sino sólo el valor de p correspondiente, que en este caso es 0.043936; como 0.043936 < 0.05, se rechaza la hipótesis nula y se acepta la alternativa.

```
Contraste de Hipótesis
-----------------------
Proporción de la Muestra = 0,85333
Tamaño de la Muestra = 150

Aproximado 95,0% superior limite de confianza para p:  [0,89857]

Hipótesis Nula: proporción = 0,9
Alternativa: menor que
p-Valor = 0,0439636
Rechazar la hipótesis nula para alpha = 0,05.
```

Figura C4.8. Inferencia sobre una proporción.

C4.2. PROCEDIMIENTOS PARA FICHERO DE DATOS

Parámetros de la distribución normal

Consideraremos a continuación la situación donde tenemos un fichero de datos como el de la Figura C2.1, del Ejemplo C2.1. Para este fichero de datos, consideremos la variable edad, que ha sido ampliamente descrita en el Apéndice C2 y cuya normalidad se ha comprobado en C3. Por lo tanto, para calcular intervalos de confianza para la media y la varianza utilizaríamos las fórmulas de los Capítulos 6 y 7.

Elegimos el procedimiento **Descripción, Datos numéricos, Análisis unidimensional,** indicamos que la variable edad es la que queremos estudiar y pulsamos **Aceptar.** Se ejecutan las opciones por defecto, entre ellas no están los intervalos de confianza, así que pulsamos las Opciones tabulares, de la barra de ejecución y marcamos **Intervalos de confianza.** Obtenemos los intervalos de confianza para la media y la desviación típica, por defecto al 95% de confianza (véase Fig. C4.9), Tenemos por lo tanto que el intervalo para la media es 57.2002, 61.2998 y para la varianza 8.42579^2, $11.3606^2 = 16.85158$, 129.06323, ambos al 95% de confianza.

Si queremos cambiar el nivel de confianza de los intervalos, habría que hacer clic en el botón derecho del ratón y en **Opciones de ventana,** cambiar el nivel de confianza.

```
Intervalos de Confianza para edad
---------------------------------

95,0% intervalo de confianza para la media:  59,25 +/- 2,04979    [57,2002:61,2998]
95,0% intervalo de confianza para la desviación tipica:  [8,42579;11,3606]
```

Figura C4.9. Intervalos de confianza para media y desviación típica de la normal.

Este mismo ejemplo nos servirá para ver los contrastes de hipótesis. Supongamos que queremos contrastar que la edad media de los pacientes que padecen este tipo de tumor es de 60 años, utilizaremos como nivel de significación $\alpha = 0.05$.

Elegimos el procedimiento **Descripción, Datos numéricos, Análisis unidimensional,** indicamos que la variable edad es la que queremos estudiar y pulsamos **Aceptar.** Se ejecutan las opciones por defecto, entre ellas no están los test de hipótesis, así que pulsamos el icono de las Opciones tabulares y marcamos **Contraste de hipótesis.** Obtenemos los resultados por defecto, entre ellos, un test de hipótesis bilateral (dos colas), con hipótesis nula $\mu = 0$; para variar estos valores, hay que hacer clic en el botón derecho del ratón y hacer los cambios en **Opciones de ventana** (véase Fig. C4.10).

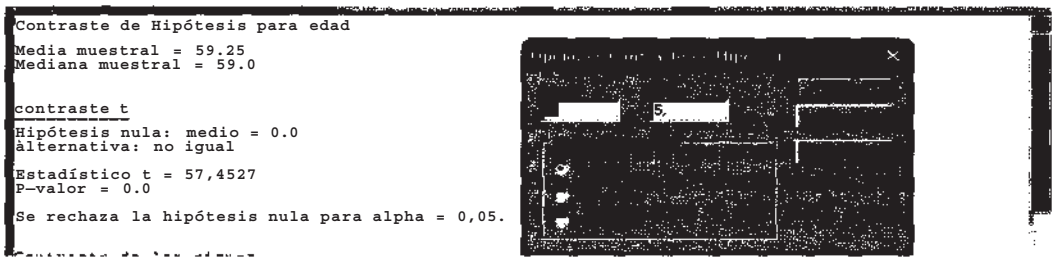


Figura C4.10. Opciones de cambio en los contrastes.

Observando la Figura C4.10, vemos que tenemos que cambiar **Media** por 60, el resto de los campos estaría correcto. Al pulsar **Aceptar** se obtienen los resultados de la Figura C4.11. Observando dicha figura vemos, en la parte superior, la estimación de la media (59.25) y de la mediana (59.0) y debajo aparece un test que se denomina **contraste t,** y que se corresponde con el contraste para la media de la normal visto en el Capítulo 6. El valor del estadístico del contraste aparece en la pantalla con el nombre Estadístico t = -0.727249, que se corresponde con el contraste $H_0 : \mu = 60$, $\mathbf{H_1 : \mu \neq 60}$. Como el valor de p = 0.469026 > 0.05, no se puede rechazar la hipótesis nula, es decir, no hay evidencia de que la edad media, para las personas que padecen este tipo de cáncer, sea distinta de 60.

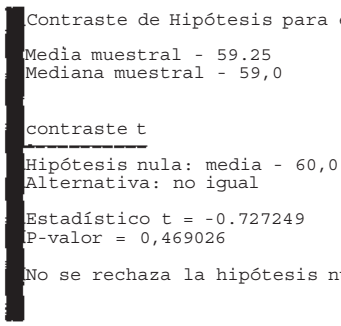


Figura C4.11. Test de la t para la media de la distribución normal.

Hay que señalar que, por defecto, también se calculan dos test no paramétricos que se comentarán en el Apéndice C9. Asimismo aparecen por defecto dos opciones gráficas, que son el **Gráfico de dispersión y** el **Gráfico de caja y bigotes** descritos en el Apéndice C2.

El STATGRAPHICS Plus no tiene ningún procedimiento específico para la inferencia sobre la varianza de una normal, cuando los datos están en un fichero. Por lo tanto, se debería utilizar el procedimiento visto anteriormente con datos resumidos. Así, los pasos a seguir son:

1) Calcular la desviación típica de los datos: elegimos el procedimiento **Descripción, Datos numéricos, Análisis unidimensional,** indicamos que queremos estudiar la variable edad y pulsamos **Aceptar.** En las opciones por defecto aparece **Resumen estadístico;** de esa pantalla obtenemos los valores frecuencia = 88 y desviación típica = 9.67429.
2) Inferencia sobre la desviación típica: una vez que tenemos el valor muestral de la desviación típica procederemos según el apartado anterior de este Apéndice: elegimos el procedimiento **Descripción, Contraste de hipótesis** y aparece por defecto marcada la media de una distribución normal, que habrá que cambiar por desviación típica normal.

Parámetro p de la distribución binomial

Cuando los datos están en un fichero, el STATGRAPHICS Plus no tiene ningún procedimiento específico para hacer inferencia sobre el parámetro p de una binomial. Supongamos que en el ejemplo del Apéndice C2 queremos ver si entre los afectados por este tipo de cáncer menos de una cuarta parte son mujeres. La forma de actuar será también en dos tiempos (H_0: p = 0.25, H_1: p < 0.23):

1) Calcular la proporción en la muestra de la característica (en el ejemplo mujeres): elegimos **Descripción, Datos Cualitativos, Tabulación,** indicamos que queremos estudiar la variable sexo y pulsamos **Aceptar.** En las opciones por defecto aparece **Tabla de frecuencias;** de esa ventana, obtenemos la frecuencia relativa de mujeres que es 0.1818, de un total de 88 casos (véase Fig. C4.12).
2) Inferencia sobre una proporción; una vez que tenemos el valor muestral de p, procederemos según el apartado C4.1. Elegimos el procedimiento **Descripción, Contraste de hipótesis** y aparece por defecto marcada la **media** de una distribución normal, que habrá que cambiar por **Proporción binomial,** y variar los valores de los diferentes campos como se recoge en la Figura C4.12. Si se pulsa **Aceptar,** se obtiene la opción por defecto que es el test bilateral, por lo tanto hay que pulsar el botón derecho del ratón, **Opciones de análisis** y cambiar la hipótesis alternativa por *Menor que.* El valor de p obtenido es 0.08463; por lo tanto, como 0.08463 > 0.05, no hay evidencia estadísticamente significativa para admitir que el porcentaje de mujeres afectadas sea inferior al 25%.

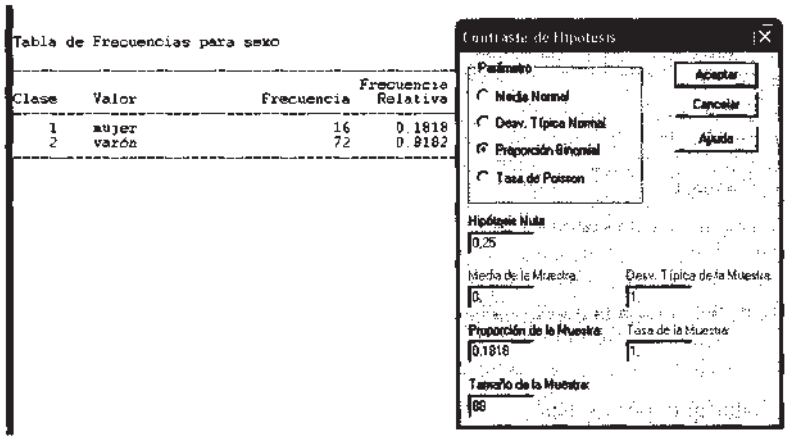
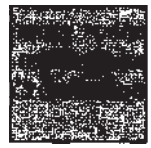


Figura C4.12. Contraste de una proporción.



Comparación de dos poblaciones

Cuando se aborda el problema de comparar dos poblaciones, y en particular la comparación de dos de sus parámetros, dos medias, dos varianzas o dos proporciones, resulta primordial preguntarse sobre el tipo de muestreo utilizado y por la distribución de la variable de interés. Así, si las dos muestras se obtuvieron de manera independiente, los métodos de inferencia estadística sobre los dos parámetros diferirán de los utilizados para el análisis de muestras apareadas o emparejadas. Si la distribución de los datos es, o puede suponerse, normal, los métodos de inferencia utilizados serán los conocidos como métodos paramétricos, mientras que si la distribución de los datos es desconocida y no puede admitirse la normalidad de los mismos, deberían utilizarse métodos no paramétricos.

C5.1. PROCEDIMIENTOS PARA DATOS RESUMIDOS

Al igual que ocurre en el caso de una sola población, el STATGRAPHICS Plus permite la comparación de dos muestras independientes, es decir la comparación estadística de dos de sus parámetros, a partir de ficheros que contengan todos los datos, o a partir, únicamente, de las estimaciones que, de dichos parámetros, se obtienen de tales ficheros completos. Comenzamos con esta última situación.

Comparación de dos medias y dos varianzas de poblaciones normales independientes

Para ilustrar este procedimiento vamos a revisar el Ejemplo 9.2.1, del Capítulo 9, en el que se desea comparar la dosis de digoxina en dos grupos de pacientes, > 64 años y < 64 años. Se desea contrastar la hipótesis estadística $\mu_1 < \mu_2$, es decir, si la dosis media de digoxina es menor en el grupo de más de 64 años. Fijamos el nivel de significación del contraste $\alpha = 0.05$. Los datos que se obtuvieron para dos muestras de 41 y 29 pacientes respectivamente, figuran en el Ejemplo 9.2.2 y son los siguientes:

Pacientes con más de 64 años

$n_1 = 41$
$x_1 = 0.265$ mg/día
$s_1 = 0.102$ mg/día

Pacientes con 64 años o menos

$n_2 = 29$
$x_2 = 0.268$ mg/día
$s_2 = 0.068$ mg/día

Antes de proceder, debemos suponer que la dosis de digoxina sigue una distribución normal en ambos grupos. Las posibles comparaciones de parámetros de dos poblaciones independientes se muestran en la Figura C5.1, que se obtiene seleccionando el procedimiento **Comparación, Dos muestras, Contraste de hipótesis...**

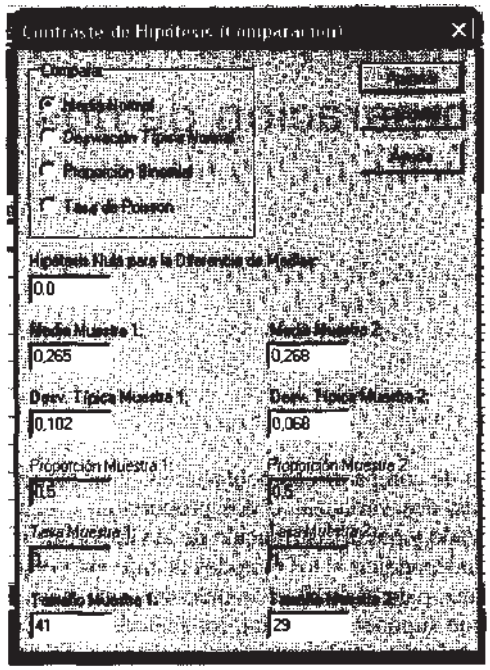


Figura C5.1. Comparación de parámetros de dos poblaciones independientes. Pantalla de datos correspondiente al Ejemplo 9.2.1.

La opción por defecto para este procedimiento es la comparación de dos medias de poblaciones normales, que es precisamente la que deseamos abordar. Rellenamos con los datos de ambas muestras los campos que aparecen activos, como muestra la Figura C5.1.

Antes de llevar a cabo dicha comparación, debemos contrastar una hipótesis estadística previa como es la de igualdad de varianzas poblacionales, $\mathbf{H_0 : \sigma_1^2 = \sigma_2^2;}$ para ello, en el procedimiento anterior marcar la opción *Comparar.* **Desviación típica normal, y Aceptar.** Los resultados se muestran en la Figura C5.2. El contraste de igualdad de varianzas que debemos abordar es un contraste bilateral (con dos colas), por lo que disponemos de dos maneras alternativas de llevar a cabo dicho contraste: a partir del intervalo de confianza para el cociente de varianzas o, a partir del valor de p del propio contraste bilateral. El intervalo de confianza al 95% para el cociente de varianzas es (1.09881, 4.41346), puesto que dicho intervalo no contiene el valor 1, se debe rechazar que las varianzas poblacionales sean iguales para un nivel de significación $\alpha = 0.05$.

Debajo de este intervalo, se describe el contraste de igualdad de las dos varianzas o de las dos desviaciones típicas, lo que equivale a contrastar que su cociente es 1. La hipótesis nula establece que el cociente σ_1/σ_2 es igual a 1, frente a la hipótesis alternativa de que dicho cociente es distinto de 1 (sigma 1/ sigma 2 <> 1 en el StatAdvisor), el estadístico del contraste es F = 2.25 y su valor p asociado (obtenido de la distribución F de Snedecor) es 0.02719, lo que conduce a rechazar la hipótesis nula para cualquier nivel de significación $\alpha > 0.028$. La pantalla de resultados de este procedimiento incluye la decisión del contraste: «Rechazar la hipótesis nula para *alpha* = 0.05». La Figura C5.2 incluye la ventana de opciones de este contraste. Se puede elegir la hipótesis alternativa correspondiente a los dos test unilaterales (con colas a la izquierda o a la derecha) y cambiar el nivel α que, por defecto, es del 5%.

```
Contraste de Hipótesis
----------------------

Desviaciones Típicas de la Muestra = 0,102 y 0,068
Tamaños de la Muestra = 41 y 29

95,0% intervalo de confianza para el ratio de varianzas:
  [1,09881;4,41346]

Hipótesis Nula: ratio de varianzas = 1.0
Alternativa: no igual
Estadístico F calculado = 2,25
p-Valor = 0,0271927
Rechazar la hipótesis nula para alpha = 0,05.
```

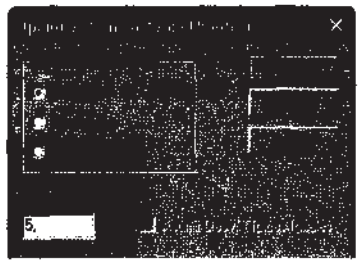


Figura C5.2. Comparación de varianzas para los datos del Ejemplo 9.2.1 y ventana de opciones de dicha comparación.

Ahora, estamos ya en condiciones de responder a la pregunta inicial ¿es la dosis media de digoxina en el grupo «> 64 años» estadísticamente menor que en el grupo «< 64 años»? Para ello, volvamos a la opción por defecto de este procedimiento *Comparar.* **Media normal** (véase Fig. C5.1) y **Aceptar.** Este procedimiento ejecuta, por defecto, un contraste bilateral, en el supuesto de que las desviaciones típicas sean iguales. Para modificar las condiciones de este procedimiento y adecuarlas a las de nuestro contraste, sobre la pantalla de resultados, debemos pulsar el botón derecho del ratón para abrir las **Opciones de análisis,** elegir *Hipótesis alternativa:* **Menor que,** y desactivar la opción *Asumir Desv. Típicas Iguales* (véase Fig. C5.3). La Figura C5.3 muestra la pantalla de resultados. La hipótesis nula establece que la diferencia de medias, $\mu_1 - \mu_2$, es igual a 0 frente a la hipótesis alternativa de que dicha diferencia es menor que 0, $\mu_1 - \mu_1 < 0$ (mu 1-mu 2 < 0 en el StatAdvisor). El estadístico del contraste es t = - 0.147584 y su valor p asociado (obtenido de la distribución t de Student aproximada) es 0.441555, lo que conduce a no rechazar la hipótesis nula para cualquier nivel de significación $\alpha < 0.44$. De nuevo, la decisión de este contraste aparece escrita en la pantalla de resultados: «No rechazar la hipótesis nula para *alpha* = 0.05».

```
Contraste de Hipótesis

Hedías de la Muestra - 0.265 y 0.268
Desviaciones Típicas de la Muestra = 0.102 y 0.068
Tamaños de la Muestra - 41 y 29

Aproximado 95. 0% superior limite de confianza para la
diferencia entre medias: -0.003 + 0.0338988    [0.0308988]

Hipótesis Nula: diferencia entre medias =0,0
Alternativa: menor que
Estadístico t calculado - -0.147584
p-Valor = 0.4415S5
No rechazar la hipótesis nula para alpha = 0.05.

(No se asumen varianzas iguales).
```

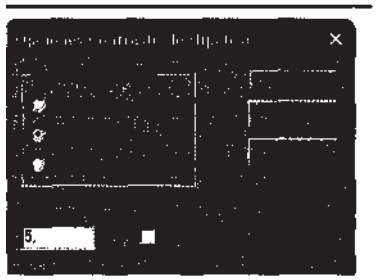


Figura C5.3. Comparación de medias para los datos del Ejemplo 9.2.1 y ventana de opciones de dicha comparación.

Comparación de dos proporciones independientes. Muestras grandes

Para ilustrar este procedimiento vamos a revisar el Ejemplo 8.6.2, del Capítulo 8, en el que se desea analizar si la ingesta de vitamina C es una ayuda en el tratamiento del cáncer. El suceso a evaluar es la mejoría en el plazo de 4 semanas. Se utilizó un grupo inicial de 150 pacientes que fue dividido en dos grupos de 75. Un grupo recibió 10 gramos de vitamina C diariamente, el otro recibió un placebo cada día. Al cabo de las 4 semanas de estudio, 47 de los pacientes que tomaron vitamina C y 43 de los que tomaron el placebo presentaron alguna mejoría.

Vamos a llevar a cabo un contraste bilateral para evaluar si existen diferencias significativas entre ambas proporciones muestrales, $p_1 = (47/75) = 0.63$ y $p_2 = (43/75) = 0.57$. Fijamos el nivel de significación del contraste $\alpha = 0.05$. Seleccionamos el procedimiento **Comparación** de la barra de

menú, a continuación **Dos muestras,** y finalmente **Contraste de hipótesis.** Si elegimos la opción *Comparar.* **Proporción binomial,** y rellenamos con los datos de ambas muestras los campos que aparecen activos en la ventana del procedimiento (véase Fig. C5.4), al pulsar **Aceptar,** se obtienen los resultados que muestra la Figura C5.5.

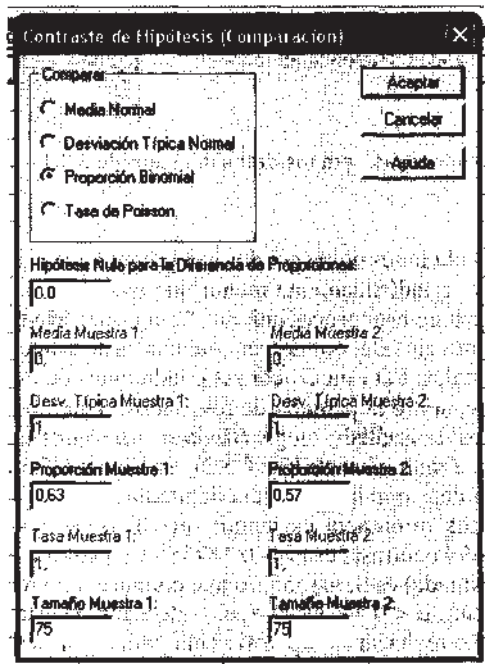


Figura C5.4. Pantalla de datos correspondiente al Ejemplo 8.6.2.

El contraste que vamos a llevar a cabo establece como hipótesis estadísticas las siguientes: H_0: $p_1 - p_2 = 0$ frente a $\mathbf{H_1}$: $\mathbf{p_1 - p_2 \neq 0}$. Disponemos de dos maneras alternativas para llevar a cabo este contraste bilateral: a partir del intervalo de confianza para la diferencia entre proporciones o, a partir del valor de p del propio contraste bilateral. El intervalo de confianza al 95% para la diferencia entre proporciones es (-0.0965, 0.2165), puesto que dicho intervalo contiene el valor 0, no se debe rechazar que las proporciones poblacionales, p_1 y p_2, sean iguales para un nivel de significación $\alpha = 0.05$.

Debajo de este intervalo, se describe el contraste de hipótesis, el estadístico del contraste es $z = 0.75$ y su valor p asociado (obtenido de la distribución normal) es 0.453252, lo que conduce a no rechazar la hipótesis nula para cualquier nivel de significación $\alpha < 0.45$. Podemos leer la decisión de este contraste en la pantalla de resultados: «No rechazar la hipótesis nula para *alpha* - 0.05». La Figura C5.5 incluye la ventana de opciones de este contraste, se puede elegir la hipótesis alternativa correspondiente a los dos test unilaterales (con colas a la izquierda o a la derecha) y cambiar el nivel α, que por defecto es del 5%.

Antes de finalizar con este procedimiento, es preciso puntualizar que el contraste que acabamos de explicar se basa en una distribución aproximada, la N(0, 1), por lo que, en rigor, sólo es adecuado para muestras grandes. La pregunta inmediata es qué se entiende por «muestra grande». Esta pregunta no tiene una única respuesta, los tamaños que hacen válida la aproximación normal dependen también del orden de magnitud de las proporciones muestrales, $\hat{p}_1$ y $\hat{p}_2$, cuanto mayores sean éstas menos tamaño se requiere. El STATGRAPHICS Plus muestra una advertencia, en la pantalla de resultados y en el StatAdvisor de este procedimiento, cuando los tamaños muestrales son escasos. Cuando no puede usarse la aproximación normal, lo razonable es usar el test de Fisher. El

STATGRAPHICS Plus calcula el test de Fisher cuando los tamaños muestrales son pequeños. Para ello, debemos elegir el procedimiento **Descripción, Datos cualitativos, Tablas de contingencia,** habiendo creado previamente un fichero de datos, con dos columnas y dos filas, que contenga las frecuencias observadas; en nuestro caso, las columnas serían 47, 43 y 28, 32. Hay que advertir que el test de Fisher no figura como opción tabular del procedimiento descrito, por lo que su ejecución depende de los tamaños totales de las dos muestras. En nuestro caso, el procedimiento anterior no calcula el test de Fisher. La Figura C5.5 incluye la ventana de opciones de este contraste; se puede elegir la hipótesis alternativa, correspondiente a los dos test unilaterales, y cambiar el nivel α que, por defecto, es del 5%.

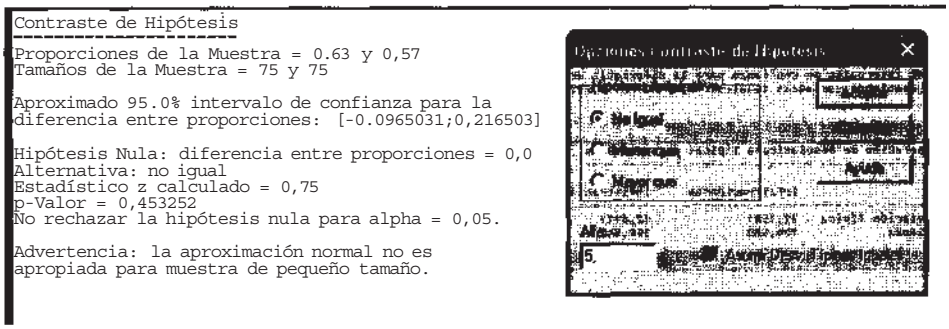


Figura C5.5. Comparación de proporciones para los datos del Ejemplo 8.6.2 y ventana de opciones de dicha comparación.

C5.2. PROCEDIMIENTOS PARA FICHERO DE DATOS

Comparación de dos medias y dos varianzas de poblaciones normales independientes

Considerando los datos contenidos en el Ejemplo C2.1, del Apéndice C2, vamos a comparar el diámetro del tumor en los dos grupos que define la variable infiltración ganglionar, más en concreto vamos a comparar, estadísticamente, las medias del diámetro del tumor de los pacientes con y sin infiltración ganglionar. Las dos muestras son claramente independientes. Vamos a suponer que el diámetro del tumor se distribuye normalmente en ambas poblaciones. Fijamos el nivel de significación del contraste $\alpha = 0.05$

Para llevar a cabo este análisis, elegimos el procedimiento **Comparación, Dos muestras, Comparación de dos muestras...,** una vez abierta la ventana del procedimiento seleccionar la opción *Entrada:* **Columnas de código y datos,** y elegir *Datos:* **diámetro,** y *Código de Muestra:* **infiltración.** La otra opción de *Entrada* de datos, **Dos columnas de datos,** requiere que las dos muestras de datos a comparar figuren en dos columnas separadas en el fichero de datos. Los resultados se muestran en la Figura C5.6. Hemos seleccionado entre las opciones tabulares, **Resumen del procedimiento, Resumen estadístico, Comparación de medias y Comparación de desviaciones típicas,** y entre las opciones gráficas, **Gráficos de caja y bigotes, Función de distribución.**

El Resumen del procedimiento nombra como Muestra 1 al subgrupo sin infiltración ganglionar, y como Muestra 2 al subgrupo con infiltración. Es importante recordar este orden ya que cuando se analiza posteriormente la diferencia de medias, el minuendo será la media de la muestra 1; igualmente, al analizar el cociente de varianzas, el numerador será la varianza de la muestra 1. Vemos, además, que la muestra 1 está formada por 60 datos y la muestra 2 por 28. También se muestran los valores mínimo y máximo de los datos de ambas muestras.

El Resumen estadístico, que se muestra ampliado en la Figura C5.7, nos va a proporcionar la primera aproximación a la comparación que queremos llevar a cabo. Observando la media y la

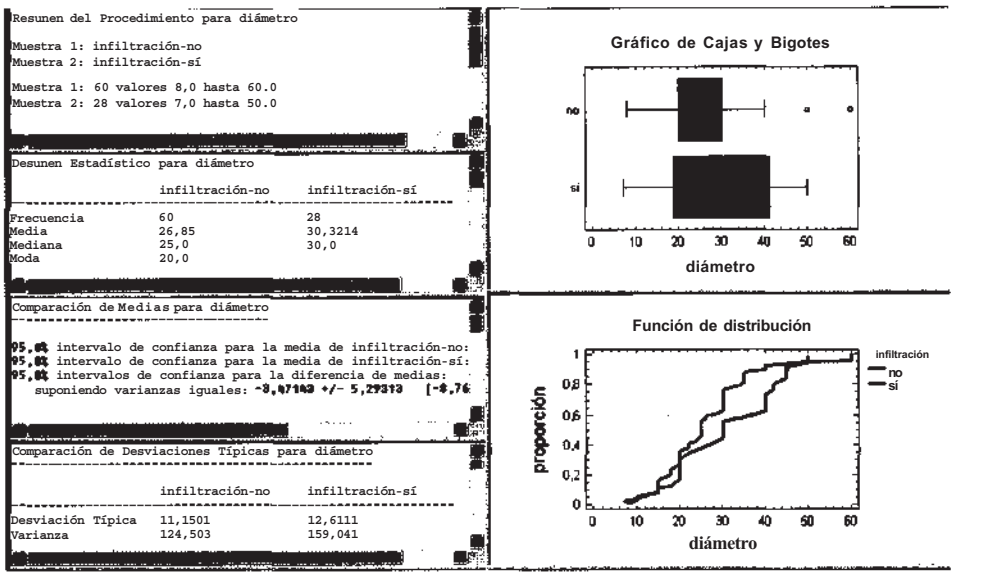


Figura C5.6. Pantalla de resultados correspondiente a la comparación del diámetro del tumor entre pacientes con y sin infiltración ganglionar.

mediana de ambos grupos, se aprecia que el grupo con infiltración tiene un diámetro ligeramente superior. En cuanto a la moda, sólo se muestra el valor 20 mm, correspondiente al grupo sin infiltración; en el otro grupo hay dos modas, 30 mm y 40 mm (ambos valores aparecen en 4 pacientes), por lo que no se muestra ningún valor para este grupo. Una medida de especial importancia en la comparación de dos medias de poblaciones independientes es la varianza, pues de sus valores en los dos grupos, y más en concreto de su cociente, se derivará una u otra prueba estadística. Finalmente, la asimetría y la curtosis tipificadas resultarán importantes para contrastar la normalidad de la variable en ambos grupos. En este caso, hemos supuesto que el diámetro del tumor sigue una distribución normal en ambos grupos, por lo que no tendremos en cuenta estos estadísticos.

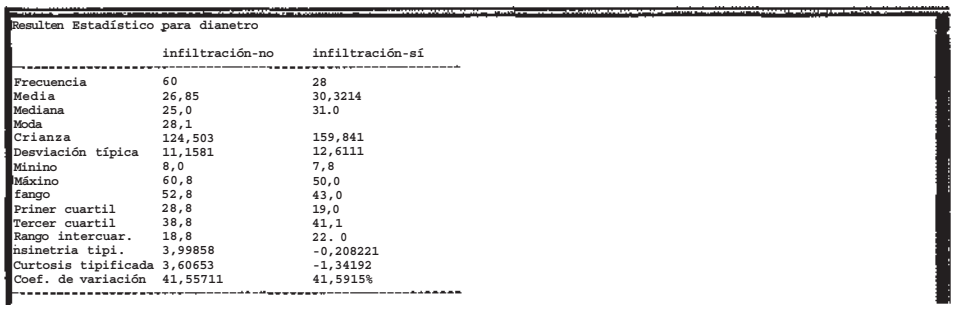


Figura C5.7. Estadísticos correspondientes al diámetro del tumor en los grupos con y sin infiltración ganglionar.

La tercera de las opciones tabulares que debe examinarse antes de llevar a cabo la comparación de medias es la **Comparación de desviaciones típicas,** que se muestra ampliada en la parte inferior de la Figura C5.8. De especial importancia es el estadístico cociente de varianzas, 0.782833. Este valor, que se obtiene dividiendo la varianza de la muestra 1 entre la de la muestra 2, es un esti-

mador del mismo cociente de varianzas poblacionales. Se debe contrastar la hipótesis estadística de igualdad de ambas varianzas poblacionales o, lo que es equivalente, que su cociente sea 1. Se muestran dos maneras de llevar a cabo dicho contraste: a partir del intervalo de confianza para el cociente de varianzas, o a partir del valor de p del propio contraste bilateral. El intervalo de confianza al 95% para el cociente de varianzas es (0.390609, 1.44785). Puesto que dicho intervalo contiene el valor 1, no se puede rechazar que las varianzas poblacionales sean iguales.

Debajo de este intervalo, se describe el contraste de igualdad de las dos varianzas, lo que equivale a contrastar la igualdad de las dos desviaciones típicas. La hipótesis nula establece que ambas desviaciones típicas son iguales frente a la hipótesis alternativa de que son diferentes (sigma 1 <> sigma 2); el estadístico del contraste es F = 0.782833 y su valor p asociado (obtenido de la distribución F de Snedecor) es 0.428393, lo que conduce a no rechazar la hipótesis nula para cualquier nivel de significación $\alpha < 0.42$.

Estamos ya en condiciones de responder a la pregunta inicial ¿existen diferencias significativas entre los diámetros medios de ambos grupos? El contraste que vamos a llevar a cabo establece como hipótesis estadísticas: $H_0 : \mu_1 - \mu_2 = 0$ (diámetros medios iguales) frente a H_1: $\mu_1 - \mu_2 \neq 0$ (diámetros medios distintos). Se trata pues de un contraste bilateral (con dos colas). Los resultados se muestran en la parte superior de la Figura C5.8, **Comparación de medias** para diámetro. Este procedimiento ejecuta, por defecto, un contraste bilateral, con diferencia de medias 0, en el supuesto de que las desviaciones típicas sean iguales y para un nivel $\alpha = 0.05$. Como éstas son nuestras condiciones, no es preciso modificar nada. Para otras condiciones del contraste, sobre la pantalla de resultados, pulsar el botón derecho del ratón para abrir las **Opciones de ventana,** elegir *Hipótesis nula, Alternativa,* nivel *Alpha,* activar/desactivar la opción *Asumir Desv. Típicas Iguales.* Estas **Opciones de ventana** se muestran en la parte derecha de la Figura C5.8.

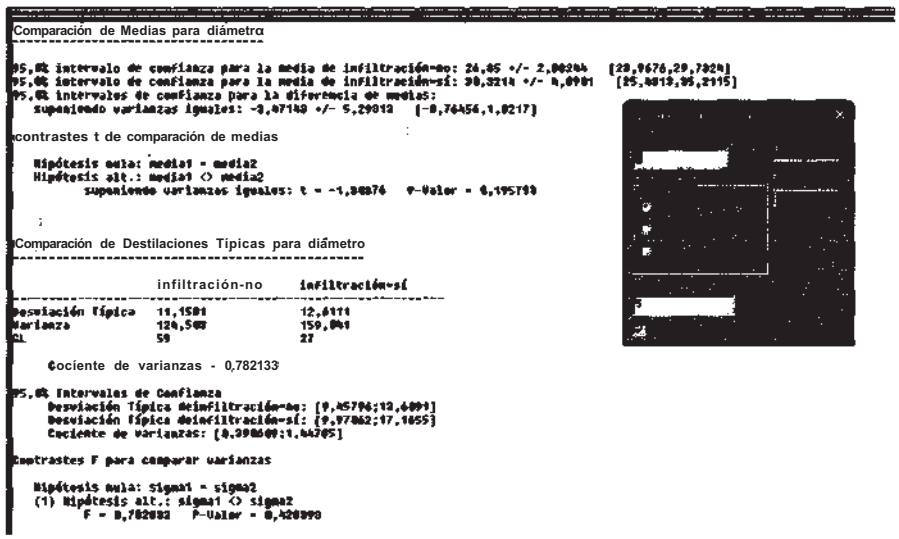


Figura CS.8. Comparación de medias y desviaciones típicas del diámetro del tumor en pacientes con y sin infiltración ganglionar. Opciones para la comparación de medias.

Al igual que antes disponemos de dos maneras alternativas de llevar a cabo dicho contraste: a partir del intervalo de confianza para la diferencia de medias o, a partir del valor de p del propio contraste bilateral. El intervalo de confianza al 95% para la diferencia de medias es (- 8.76456, 1.8217). Puesto que dicho intervalo contiene el valor 0, no se puede rechazar que las dos medias poblacionales sean iguales. La hipótesis nula establece que ambas medias son iguales (media 1 = media 2), frente a la hipótesis alternativa de que son diferentes (media 1 **o** media 2). El estadísti-

co del contraste es t = -1.30376 y su valor p asociado (obtenido de la distribución t de Student) es 0.195793, lo que conduce a no rechazar la hipótesis nula para cualquier nivel de significación $\alpha < 0.19$. Es decir, las diferencias entre los diámetros del tumor en ambos grupos no son estadísticamente significativas para $\alpha = 0.05$.

Con las dos opciones gráficas seleccionadas (véase Fig. C5.6) se pretende visualizar la comparación que queremos abordar. El **Gráfico de cajas y bigotes** parece anticipar el resultado: las difeerencias no son estadísticamente significativas. El otro gráfico, **Función de distribución,** muestra las distribuciones de frecuencias relativas acumuladas de ambos grupos, también conocidas como funciones de distribución empíricas. El cruce o solapamiento entre ambas apuesta, de nuevo, por la no significación de las diferencias. Conviene recalcar que ambos gráficos no pueden sustituir a los test analíticos que acabamos de comentar; en particular hay un contraste estadístico que usa las funciones de distribución empíricas para comparar dos muestras independientes, el test de Kolmogorov-Smirnov, que no tiene cabida en este apartado por tratarse de un test no paramétrico.

Si comparamos los dos procedimientos mostrados para comparar dos medias y dos varianzas de poblaciones normales independientes, para datos resumidos y para datos completos, parece evidente la mayor capacidad de análisis de este último. En particular, podemos elegir más opciones tabulares y gráficas.

Comparación de dos medias para datos apareados. Poblaciones normales

En el Ejemplo 9.5.1, del Capítulo 9, se analiza el efecto del ejercicio físico en el nivel de colesterol en plasma. Para 11 sujetos, se mide el nivel de colesterol antes y después de un período de ejercicios. Los datos aparecen en dicho ejemplo. Nuestro objetivo es contrastar la hipótesis estadística H_1: $\mu_A - \mu_D > 0$, siendo μ_A y μ_D los niveles medios de colesterol antes y después del ejercicio, respectivamente. Las dos muestras de datos son claramente apareadas o emparejadas. Vamos a suponer que el nivel de colesterol se distribuye normalmente. Fijamos el nivel de significación del contraste $\alpha = 0.1$.

Para llevar a cabo este análisis, elegimos el procedimiento **Comparación, Dos muestras. Comparación de muestras pareadas.** Una vez abierta la ventana del procedimiento, elegimos *Muestra 1:* **colesterol antes,** y *Muestra 2:* **colesterol después.** Los resultados se muestran en la Figura C5.9. Hemos seleccionado, entre las opciones tabulares, **Resumen estadístico, Intervalos de confianza y Contraste de hipótesis.**

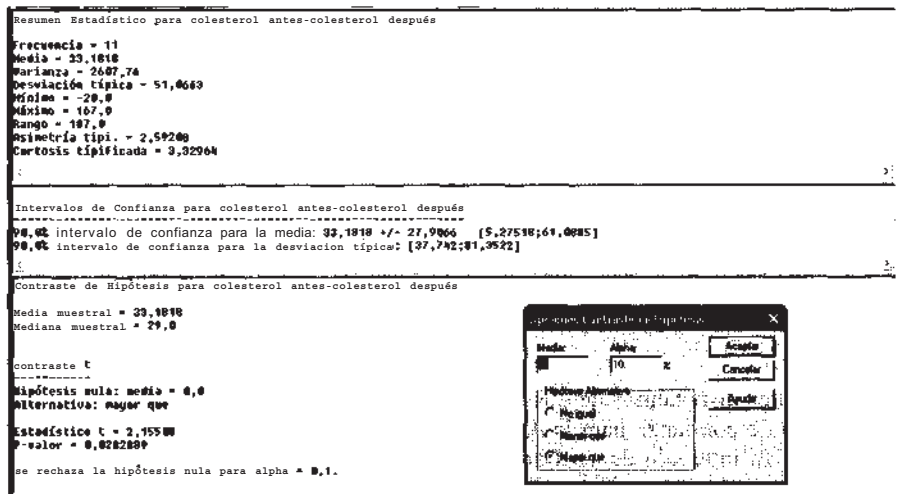


Figura C5.9. Pantalla de resultados correspondiente a la comparación del colesterol antes y después del ejercicio físico. Opciones de ventana de la comparación.

Del **Resumen estadístico** destacamos dos características: el valor medio del descenso de colesterol en los 11 sujetos, que fue de 33.1818 mg/dL; y la dispersión de los datos, desviación típica, 51.06 mg/dL, o rango, 187 mg/dL. Lo elevado de estos dos últimos valores «cuestiona» la significación estadística del descenso observado.

El Intervalo de confianza al 90% para el descenso medio de colesterol es (5.275, 61.088). Si elevamos el nivel de confianza al 95% (abriendo las **Opciones de ventana),** el nuevo intervalo sería (-1.125, 67.488). Una diferencia entre los dos intervalos es que el segundo contiene el valor 0, por lo que en un test bilateral para la diferencia de medias con $\alpha \equiv 0.05$, la decisión sería no rechazar la hipótesis nula que establece «que no hay descenso medio de colesterol».

El Contraste de hipótesis unilateral, H_0: $\mu_A - \mu_D = 0$ frente a $\mathbf{H_1: \mu_A - \mu_D > 0}$, se lleva a cabo eligiendo en las opciones de ventana del procedimiento las que se muestran en la parte inferior derecha de la Figura C5.9. El estadístico del contraste es t = 2.155 y su valor p asociado (obtenido de la distribución t de Student) es 0.02828, lo que conduce a rechazar la hipótesis nula para cualquier nivel de significación $\alpha > 0.029$. Por tanto, la misma decisión se habría obtenido si hubiésemos fijado $\alpha = 0.05$.

La aparente contradicción entre las decisiones adoptadas con el intervalo y el contraste, cuando $\alpha = 0.05$, no es tal ya que no son equivalentes los dos contrastes. El intervalo de confianza sólo permite el contraste de hipótesis bilaterales. En realidad, siempre ocurrirá lo que acabamos de comentar y su explicación formal estriba en que los test bilaterales son «más conservadores» que los unilaterales.

Comparación de dos proporciones independientes. Muestras grandes

Considerando los datos contenidos en el Ejemplo C2.1, del Apéndice C2, vamos a comparar la proporción de recidiva local en hombres y mujeres, considerando únicamente los tumores localizados en la lengua. El fichero de datos reducido a los 36 tumores de lengua se muestra en la Figura C5.10. Vamos a llevar a cabo un contraste bilateral para evaluar si existen diferencias significativas entre ambas proporciones muestrales. Fijamos el nivel de significación del contraste $\alpha = 0.05$. Seleccionamos el procedimiento **Descripción, Datos cualitativos, Tabulación cruzada,** la ventana de este procedimiento, con los campos completados, se muestra dentro de la Figura C5.10, pulsando **Aceptar** se obtienen los resultados que muestra la Figura C5.11.

Hemos seleccionado, entre las opciones tabulares, **Resumen del procedimiento, Tabla de frecuencias y Contraste de chi-cuadrado** y, entre las opciones gráficas, **Gráfico mosaico.**

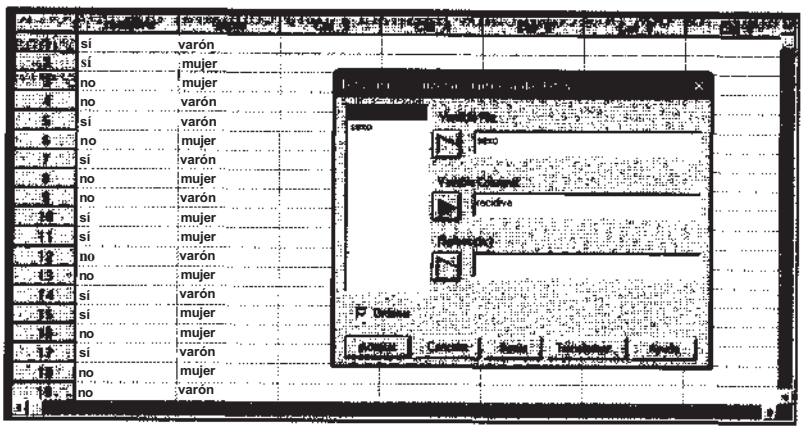


Figura C5.10. Fichero de datos correspondiente a la comparación de proporciones de recidiva local en varones y mujeres. Ventana del procedimiento.

La Tabla de frecuencias 2 x 2 contiene en cada casilla la frecuencia observada y el porcentaje que representa sobre el total de la fila, así, el reparto porcentual de las 12 mujeres atendiendo a la aparición o no de recidiva fue del 41.67% y del 58.33% respectivamente. Dentro de la Figura C5.11 aparecen las **Opciones de ventana** de este procedimiento con la opción elegida. Las proporciones muestrales que vamos a comparar son $\hat{p}_M = (5/12) = 0.4167$ y $\hat{p}_V = (8/24) = 0.3333$.

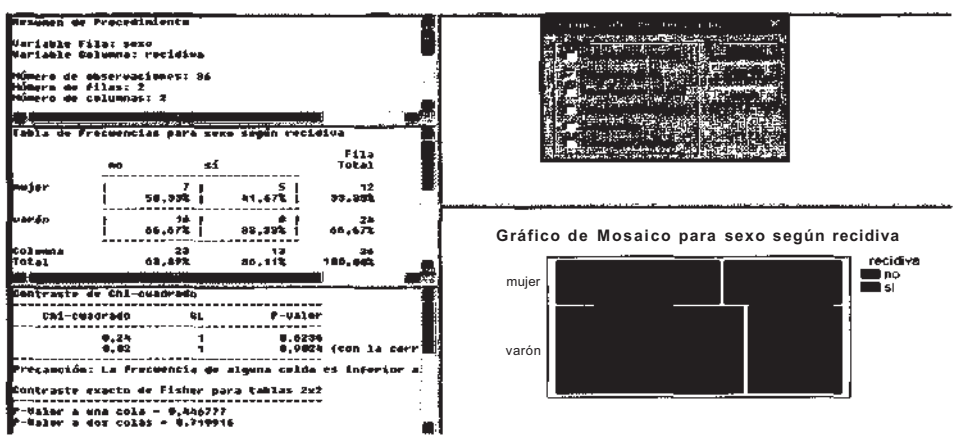


Figura C5.11. Comparación de proporciones de recidiva local en varones y mujeres.

El Contraste de chi-cuadrado es el test de comparación de las dos proporciones independientes para muestras grandes. El estadístico del contraste es $x^2 = 0.24$, y su valor de p asociado (obtenido de la distribución x^2 de Pearson) es 0.6236, lo que conduce a no rechazar la hipótesis nula para cualquier nivel de significación $\alpha < 0.62$, es decir, no existen diferencias significativas entre las proporciones de recidiva local en hombres y mujeres. Conviene destacar que el contraste que acabamos de explicar se basa en una distribución aproximada, la x^2 de Pearson, por lo que, en rigor, sólo es adecuado para muestras grandes. El STATGRAPHICS Plus muestra una advertencia en la pantalla de resultados. Cuando no puede usarse la aproximación normal, lo razonable es usar el test de Fisher, que aparece debajo del test chi-cuadrado.

Finalmente, el gráfico de mosaico, al mostrar los porcentajes relativos de cada fila, permite visualizar la comparación que se va a llevar a cabo.

Si comparamos los dos procedimientos mostrados para comparar dos proporciones independientes, para datos resumidos y para datos completos, el primero de ellos proporciona mayor capacidad de análisis numérico mientras el segundo proporciona mayores opciones gráficas.

Análisis de la varianza

El ejemplo siguiente servirá para ilustrar las técnicas de análisis de la varianza expuestas en el Capítulo 10. Consideraremos los modelos con uno y dos factores de clasificación, también conocidos como ANOVA de una y dos vías, respectivamente. Todos los modelos requieren, básicamente, dos hipótesis estructurales acerca de la variable respuesta: la normalidad e igualdad de varianzas de dicha variable en todas las poblaciones. Supondremos, al objeto de simplificar la exposición, que la variable respuesta sigue una distribución normal en todos los casos. En el apartado C3.4 del Apéndice C3 se muestra cómo comprobar dicho supuesto.

Ejemplo C6.1

La tabla siguiente muestra la información relativa a 74 recién nacidos a término en un hospital. Se describen 6 variables de diferente naturaleza. De tipo discreto: **Nacional.,** nacionalidad (1: española, 2: africana, 3: oriental), **Sexo** (1: niño, 2: niña), ambas de naturaleza cualitativa nominal. De tipo continuo: **Peso** (gramos), **Talla (mm), Per. Cran.,** perímetro craneal (mm), **Edad Gest,** semanas de gestación (semanas).

N.º	Nacional	Sexo	Peso	Talla	Per. Cran.	Edad Gest
1		2	2915	465	340	40
2		2	4235	540	355	40
3			3100	485	355	40
4			2905	475	335	40
5			3150	500	355	40
6			2680	455	345	40
7			2995	490	340	39
8			3355	515	355	40
9		2	3560	515	355	39
10		2	2840	485	340	39
11		2	3040	490	345	40
12		2	3140	495	330	40
13		2	3100	490	320	37
14		2	2805	475	340	40

N.°	Nacional	Sexo	Peso	Talla	Per. Cran.	Edad Gest.
15	1	1	3375	495	345	41
16	1	1	2725	470	345	40
17	1	2	3105	485	340	38
18	1	1	2390	450	335	38
19	1	1	2475	465	335	40
20	1	2	2440	450	325	39
21	1	2	3145	495	340	39
22	1	1	3000	490	345	40
23	1	2	3015	490	340	38
24	1	1	3605	515	360	39
25	1	1	3150	480	355	38
26	1	2	2990	475	340	39
27	1	1	3810	510	345	39
28	1	1	3340	505	350	39
29	2	1	3000	510	360	41
30	2	2	2585	470	325	37
31	2	1	3275	495	340	39
32	2	1	3080	490	340	40
33	2	1	3425	510	360	40
34	2	1	2580	460	335	39
35	2	1	3315	500	360	39
36	2	1	3015	490	345	42
37	2	1	3325	500	350	37
38	2	2	3395	505	365	39
39	2	1	3330	505	360	40
40	2	2	2640	465	320	37
41	2	2	3500	495	360	39
42	2	2	3505	505	370	38
43	2	2	3100	480	355	40
44	2	2	3040	490	340	39
45	2	1	3670	510	365	40
46	2	1	4195	525	360	41
47	2	2	3640	495	360	39
48	2	2	3590	525	345	41
49	2	1	4015	515	370	39
50	2	1	3470	515	365	39
51	3	2	3370	490	345	38
52	3	1	3515	490	350	35
53	3	2	3610	495	345	39
54	3	1	3795	490	330	36
55	3	1	3770	530	350	39
56	3	1	2966	515	345	39
57	3	1	3215	500	355	38
58	3	1	3030	480	340	40
59	3	2	2840	475	340	36
60	3	1	3975	550	355	42
61	3	2	3380	495	350	39
62	3	1	3405	505	355	40
63	3	2	3235	500	350	39
64	3	2	2645	450	320	38
65	3	1	2810	475	340	37
66	3	1	3475	520	350	40
67	3	2	2820	470	330	38
68	3	2	3065	500	340	40
69	3	1	3040	515	345	41
70	3	2	2995	485	345	40

N.°	Nacional	Sexo	Peso	Talla	Per. Cran.	Edad Gest
71	**3**	**2**	2810	505	330	**40**
72	**3**	**2**	3210	500	350	**38**
73	**3**	**1**	3470	505	340	**40**
74	**3**	**2**	3115	505	335	**39**

C6.1. ANÁLISIS DE LA VARIANZA DE UNA VIA

Deseamos comparar el perímetro craneal de los recién nacidos de las tres nacionalidades descritas en el Ejemplo C6.1. En concreto, queremos contrastar la hipótesis de que los perímetros medios de las tres poblaciones coinciden.

El contraste de igualdad de medias

El contraste estadístico se formula así: $\mathbf{H_0}: \mu_i = \mu_2 = \mu_3 = \mu$, frente $\mathbf{H_1}: \mu_i \neq \mu$ r a algún i = 1, 2, 3.Vamos a suponer que los datos provienen de seleccionar tres muestras aleatorias de recién nacidos, una de cada nacionalidad. El diseño experimental corresponde, pues, a un diseño de una vía, completamente aleatorio con efectos fijos (véase apartado 10.1 del Cap. 10).

Para llevar a cabo este contraste con el STATGRAPHICS Plus debemos elegir el procedimiento **Comparación, Análisis de la varianza, ANOVA simple.** Una vez que emerge la ventana del procedimiento, elegimos *Variable dependiente:* **perímetro craneal** y *Factor,* **nacionalidad.** La Tabla C6.1 muestra los resultados de las opciones tabulares **Tabla ANOVA y Contraste de varianza.** La primera de las opciones muestra los resultados del contraste solicitado. El valor del estadístico del contraste es 5.29 y su valor p asociado (obtenido de la distribución F de Snedecor con 2 y 71 grados de libertad) es 0.0072, lo que conduce a rechazar la hipótesis nula para cualquier nivel de significación $\alpha > 0.008$. Es decir, existen diferencias, estadísticamente significativas, entre los perímetros craneales de los grupos.

Tabla C6.1. Tabla de ANOVA y contrastes de homogeneidad de varianzas

Tabla ANOVA para perímetro craneal según nacionalidad

Análisis de la varianza

Fuente	Sumas de cuad.	Gl	Cuadrado medio	Cociente-F	Valor de p
Entre grupos	1280.2	**2**	640.101	**5.29**	**0.0072**
Intra grupos	8587.7	**71**	120.954		
Total (Corr.)	9867.91	73			

Contraste de **varianza**

Contraste C de Cochran: 0.532253	Valor de p = 0.0239954
Contraste de Bartlett: 1.08038	Valor de p = 0.0676399
Contraste de Hartley: 2.42516	
Test de Levene: 1.64421	Valor de p = 0.200434

Gl, grados de libertad.

El contraste de igualdad de varianzas

El contraste de varianza comprueba la hipótesis previa de homogeneidad de las tres varianzas, $\mathbf{H_0}: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$, frente a $\mathbf{H_1}: \sigma_i^2 \neq \sigma^2$para algún i = 1, 2, 3. Se muestran los resultados de

4 contrastes diferentes, los test de Cochran y Hartley (éste no muestra el valor p del contraste) requieren tamaños de muestras iguales o aproximadamente iguales. Los test de Barlett (véase apartado 13.8 del Capítulo 13) y Levene siempre son aplicables. Los tamaños muestrales son, en nuestro caso, 28, 22 y 24, por lo que deben preferirse los contrastes de Barlett y Levene, cuyos valoras de p son 0.067 y 0.20. Puesto que el menor de estos valores de p es mayor de 0.05, no deberíamos rechazar la hipótesis de igualdad de las tres varianzas, por lo que el contraste de igualdad de medias mostrado en la Tabla C6.1 resulta adecuado.

Para despejar las dudas acerca de la validez del contraste de igualdad de medias en caso de obtener, en los contrastes previos de igualdad de varianzas, valores de p pequeños, podemos decir que si los tamaños muestrales son semejantes, el contraste F de igualdad de medias es igualmente exacto aunque las varianzas difieran mucho. No ocurre lo mismo si los tamaños muestrales son muy diferentes **($[\text{máx } n_i / \text{mín } n_i] > 2$).** Finalmente, conviene recordar que el contraste F también es válido cuando la variable respuesta se desvía de la normalidad, siempre que los tamaños muestrales no sean pequeños. La gran perjudicada, en caso de vulnerarse la normalidad y/o la igualdad de varianzas, será la estimación de la varianza (la varianza residual) y consiguientemente los resultados que dependen de ella como las comparaciones *a posteriori* o comparaciones múltiples (véase apartado 10.2 del Cap. 10). Finalmente, y en relación a los test de Barlett y Levene, podemos decir que el de Barlett es muy eficaz pero extremadamente sensible a la falta de normalidad de la variable, déficit que no tiene el test de Levene. Por este motivo, si tenemos dudas respecto a la normalidad de la variable, debemos elegir el test de Levene.

Comparaciones múltiples

Una vez que encontramos diferencias estadísticamente significativas entre las medias de los perímetros craneales de las tres nacionalidades, el análisis debe proseguir para encontrar la/s media/s responsable/s de dicha significación. Para llevar a cabo las comparaciones de medias por parejas, debemos seleccionar la opción tabular **Contraste múltiple de rango.** Una vez obtenidos los resultados, debemos pulsar el botón derecho del ratón para abrir las **Opciones de ventana** y elegir en *Método:* **Bonferroni** y cambiar el *Nivel de confianza* que, por defecto, es del 95%. La Tabla C6.2 presenta los resultados de los tres contrastes de hipótesis, H_0: $\mu_i = \mu_j$, H_1: $\mu_i \neq \mu_j$ para i, j = 1, 2,3, utilizando el método de Bonferroni (véase apartado 10.2 del Capítulo 10). Este método lleva a cabo todos los posibles contrastes entre pares de medias.

En la primera parte de la Tabla C6.2, bajo la columna titulada Grupos homogéneos, se identifican dos grupos entre cuyas medias no hay diferencias estadísticamente significativas. Esto se representa mediante la alineación de la letra X en dicha columna, siendo aquellos grupos los correspondientes a las nacionalidades 1 y 3. Por el contrario, existen diferencias significativas entre el perímetro medio de la nacionalidad 2 y los de las nacionalidades 1 y 3. Podemos concluir, pues, al 95%, que $\mu_2 \neq \mu_1, \mu_3$. En la parte inferior de la Tabla C6.2 se muestra la diferencia estimada entre cadi par de medias marcando con un asterisco (*) las que resultaron estadísticamente significativas. Por último, la información contenida en la columna +/- Límites nos permite obtener intervalos de confianza para las diferencias de medias que resultaron significativas; por ejemplo, el intervalo de confianza para $\mu_2 - \mu_3$, al 95%, es: $(9.14773 \pm 7.9597) = (1.188, 17.107)$

Un método alternativo al de Bonferroni, menos conservador que éste (se requiere menor diferencia observada para declarar significativa dicha diferencia) es el método de Duncan (véase apartado 10.2 del Cap. 10), que no requiere llevar a cabo todas las comparaciones. Para obtener lo¡5 resultados de este método basta cambiar Bonferroni por Duncan en las **Opciones de ventana** *de* la opción tabular **Contraste múltiple de rango.** El STATGRAPHICS Plus permite otros 4 métodos para abordar las comparaciones múltiples. Conviene puntualizar que si los tamaños muestrales son iguales y queremos llevar a cabo todas las comparaciones, podemos usar el método de Tukey, que proporciona resultados más precisos que el de Bonferroni. Finalmente, si deseamos construir un único intervalo de confianza para la diferencia de dos medias debemos elegir en *Método:* **LSD.**

Tabla C6.2. Comparaciones múltiples. Método de Bonferroni

Contraste múltiple de rango para perímetro craneal según nacionalidad

Método: 95 porcentaje de Bonferroni

Nacionalidad	Frec.	Media	Grupos homogéneos
3	24	343.125	**X**
1	28	343.214	**X**
2	22	352.273	**X**

Contraste	Diferencias	+/- Límites
1-2	*-9.05844	7.68297
1-3	0.0892857	7.50156
2-3	*9.14773	7.9597

* Indica una diferencia significativa.

La Figura C6.1 muestra los resultados de la opción gráfica Gráfico **de medias;** se muestran los intervalos de confianza al 95% para las medias de las tres nacionalidades.

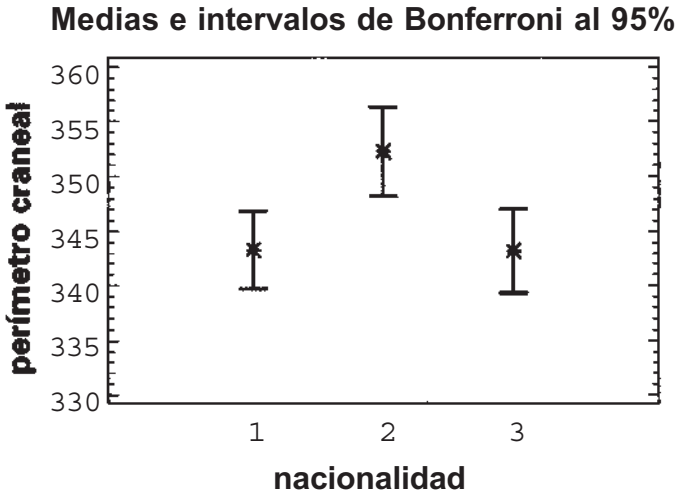


Figura C6.1. Gráfico de medias para el perímetro craneal de las tres nacionalidades.

NOTA

Si el fichero de datos contiene los valores de las diferentes muestras en columnas separadas (en nuestro ejemplo, los perímetros relativos a cada nacionalidad en una columna diferente), el análisis anterior se puede llevar a cabo eligiendo el procedimiento **Comparación, Muestras múltiples, Comparación de varias muestras.** Debemos seleccionar la opción *Entrada:* **Columnas de datos múltiples** y, una vez abierta la ventana del procedimiento, seleccionar las tres muestras y arrastrarlas al campo *Muestras;* pulsando **Aceptar,** se obtienen los mismos resultados y opciones que en el procedimiento **ANOVA simple.** Este mismo procedimiento, **Comparación de varias muestras,** permite el análisis de los datos, tal y como aparecen en el Ejemplo C6.1; bastaría modificar únicamente la opción *Entrada* anterior, elegir en este caso **Columnas de código y datos y,** una vez abier-

ta la ventana del procedimiento, elegir *Datos:* **perímetro craneal** y *Códigos de nivel:* **nacionalidad.**

C6.2. ANÁLISIS DE LA VARIANZA DE DOS VÍAS

Vamos a comparar ahora el perímetro craneal de los recién nacidos, atendiendo al sexo y a la nacionalidad. Para ello, vamos a seleccionar, entre los nacidos de cada nacionalidad, dos muestras de tamaño 10, una de niños y otra de niñas, de forma que dispondremos de 6 muestras de tamaño 10 cada una, correspondientes a las 6 combinaciones de niveles de ambos factores. El diseño experimental corresponde, pues, a un diseño de dos vías, completamente aleatorio con efectos fijos (véase apartado 10.5 del Cap. 10). El hecho de elegir todas las muestras de igual tamaño obedece a facilitar el método de análisis. Para llevar a cabo este estudio con el STATGRAPHICS Plus, debemos elegir el procedimiento **Comparación, Análisis de la varianza, ANOVA factorial** y, una vez que emerge la ventana del procedimiento, elegir *Variable dependiente:* **perímetro craneal** y *Factores:* **nacionalidad** y sexo. Pulsando **Aceptar** se obtienen las opciones por defecto de este procedimiento. Al pulsar, sobre cualquiera de ellas, con el botón derecho del ratón para abrir las **Opciones de análisis** y elegir *Máximo orden de interacción: 2,* estamos solicitando la inclusión en el modelo de un término de interacción entre el sexo y la nacionalidad. La Tabla C6.3 muestra los resultados de la opción tabular **Tabla ANOVA.**

La primera hipótesis que debemos contrastar es la de no interacción sexo-nacionalidad, que se formula como sigue: H_0: $(\alpha\beta)_{ij} = 0$, frente a $\mathbf{H_1}$: $(\alpha\beta)_{ij} \neq 0$ para algún i = 1,2; j =1, 2, 3 (véase apartado 10.5 del Capítulo 10). El valor del estadístico del contraste es 0.22 y su valor p asociado es 0.8055, lo que conduce a no rechazar la hipótesis nula, es decir, no hay evidencia estadística de interacción sexo-nacionalidad. Por consiguiente, podemos proseguir el análisis y llevar a cabo los contrastes para los efectos principales, que se formulan como sigue: (a) $\mathbf{H_0}$: $\mu_{1..} = \mu_{2..} = \mu_{sexo}$ y (b) H_0: $\mu_{.1.} = \mu_{.2.} = \mu_{.3.} = \mu_{nacional.}$. Los estadísticos de los contrastes (a) y (b) son, respectivamente, 22.26 y 7.03, y sus valores p asociados, $p < 0.0001$ y $p = 0.0019$, por lo que se rechazan ambas hipótesis nulas, es decir, existen diferencias significativas entre los perímetros craneales medios de ambos sexos, y entre los de las nacionalidades.

Tabla C6.3. Tabla de ANOVA para el modelo con dos vías e interacción

Análisis de la varianza para perímetro craneal • Sumas de cuadrados tipo **III**					
Fuente	SC	GL	CM	Cociente-F	Valor de p
EFECTOS PRINCIPALES					
A: sexo	2220.42	1	2220.42	33.26	0.0000
B: **nacionalidad**	1403.33	2	701.667	7.03	0.0019
INTERACCIONES					
AB	43.3333	2	21.6667	0.22	0.8055
RESIDUOS	5387.5	54	99.7685		
TOTAL (CORREGIDO)	9054.58	59			

El análisis prosigue abordando las comparaciones múltiples para las tres nacionalidades. En el caso del sexo, por tener sólo 2 modalidades, el análisis finaliza con el contraste (a) anterior. Si deseamos obtener intervalos de confianza para las diferencias de medias que resultaron significativas, procederíamos como en el caso de una vía, utilizando la información contenida en la columna +/- Límites, de la opción tabular **Contraste múltiple de rango** (véase Tabla C6.2). Se mostrarán por defecto los resultados relativos al sexo (el primer factor); pulsando sobre dichos resultados, con el botón derecho del ratón en las **Opciones de ventana,** seleccionaríamos *Factor,* **nacionalidad** para

obtener los resultados correspondientes al segundo factor. En estas últimas opciones se puede modificar el nivel de confianza y el método de análisis, Duncan, Bonferroni, etc. En el caso del sexo, el intervalo de confianza más preciso para la diferencia media de perímetros se obtendría eligiendo *Método:* **LSD,** al haber un único intervalo posible.

Una opción tabular recomendada, especialmente cuando la hipótesis de no interacción se rechaza, es la **Tabla de medias** (véase Tabla C6.4). Observando la columna de medias se pueden anticipar, empíricamente, los resultados que hemos obtenido: el perímetro craneal medio es 346.4 mm, 352.5 mm para los niños y 340.3 mm para las niñas; por nacionalidades, el mayor perímetro medio se da en los nacidos africanos, 353.2 mm. Estas ordenaciones se mantienen si contemplamos los perímetros medios de los sexos por nacionalidades: los niños tienen mayores valores que las niñas en las tres nacionalidades; el mayor perímetro, para ambos sexos, se da en la nacionalidad africana, lo que muestra de manera empírica la no interacción sexo-nacionalidad.

Entre las opciones gráficas de este procedimiento merecen destacarse dos, el **Gráfico de medias y** el **Gráfico de interacción.** El primero ya ha sido comentado, así que mostramos el segundo en la Figura C6.2.

Tabla C6.4. Tabla de medias para el perímetro craneal

Tabla de medias por mínimos cuadrados para el perímetro craneal con intervalos de confianza del 95%

Nivel		Frecuencia	Media	Error estándar	Límite inferior	Límite superior
Total		**60**	346.417			
Sexo						
1		**30**	**352.5**	1.82363	348.844	356.156
2		**30**	340.333	1.82363	336.677	343.989
Nacionalidad						
1		**20**	342.75	2.23348	338.272	347.228
2		**20**	**353.25**	2.23348	348.772	357.728
3		**20**	343.25	2.23348	338.772	347.728
Sexo según nacionalidad						
1	1	**10**	350.0	3.15862	343.667	356.333
1	2	**10**	**359.0**	3.15862	352.667	365.333
1	3	**10**	348.5	3.15862	342.167	354.833
2	1	**10**	335.5	3.15862	329.167	341.833
2	2	**10**	**347.5**	3.15862	341.167	353.833
2	3	**10**	338.0	3.15862	331.667	344.333

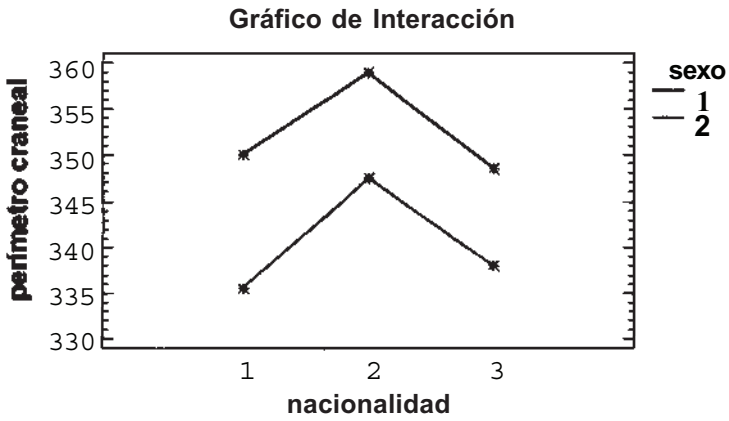


Figura C6.2. Gráfico de interacción sexo-nacionalidad.

El paralelismo de las poligonales del gráfico resulta una confirmación geométrica de la no interacción de los dos factores estudiados, el cambio del perímetro craneal de una nacionalidad a otra es prácticamente el mismo en los dos sexos. Para finalizar conviene añadir que si la interacción resulta significativa, el análisis de los efectos principales debería llevarse a cabo de la siguiente forma: analizar las posibles diferencias entre los perímetros de las tres nacionalidades para los recién nacidos niños, y estudiar lo mismo para las niñas. El procedimiento estadístico consistiría en dos Anova de una vía, uno para niños y otro para niñas.

C6.3. BLOQUES COMPLETOS ALEATORIZADOS

Si el diseño experimental se corresponde con un diseño de Bloques completos aleatorizados (véase apartado 10.4 del Cap. 10), el análisis estadístico se lleva a cabo utilizando el modelo de Anova con dos vías, efectos fijos sin interacción, con una observación por casilla. Para ello, debemos elegir el procedimiento **Comparación, Análisis de la varíanza, ANOVA factorial.** Una vez emerge la ventana del procedimiento, seleccionamos la *Variable dependiente* y dos *Factores,* el factor de interés y la variable de bloque. Pulsando **Aceptar,** se obtienen las opciones por defecto de este procedimiento, entre ellas **Tabla ANOVA.** Abriendo las **Opciones de análisis,** que emergen pulsando con el botón derecho del ratón, encontramos *Máximo orden de interacción:* 1, que significa que no hay interacción entre el factor de interés y el bloque; debemos, pues, mantener este valor. El contraste deseado, la comparación de los diferentes niveles del factor de interés, se lleva a cabo, en la opción tabular **Tabla ANOVA,** a partir del valor de p correspondiente al estadístico del contraste del efecto principal relativo al factor de interés. En la tabla de Anova aparece también el contraste relativo a la variable de bloque, que en este diseño tiene un interés menor.

Regresión y correlación

C7.1. REGRESIÓN LINEAL SIMPLE

Para ilustrar este procedimiento vamos a utilizar el Ejemplo C6.1, del Apéndice C6. Deseamos estudiar las posibles relaciones lineales entre las variables de crecimiento y/o desarrollo de los recién nacidos, es decir, entre las 4 variables continuas de los datos: peso, talla, perímetro craneal y semanas de gestación. En particular, vamos a analizar la relación entre el peso y la talla. Consideraremos el peso como variable dependiente o respuesta y la talla como variable independiente o predictora. Se trata de un estudio observacional, al escapar al control del investigador los valores de la variable predictora (véase apartado 11.1, del Cap. 11). Para obtener la recta de regresión mínima cuadrática del peso sobre la talla, debemos elegir el procedimiento **Dependencia, Regresión simple.**.. Una vez emerge la ventana del procedimiento, hemos de elegir la variable dependiente, en este caso el peso, y arrastrarla al campo Y; elegir la variable independiente, en este caso la talla, y arrastrarla al campo X (véase Fig. C7.1). Pulsando **Aceptar** se obtienen, por defecto, los resultados que se muestran en la Tabla C7.1 y la Figura C7.2.

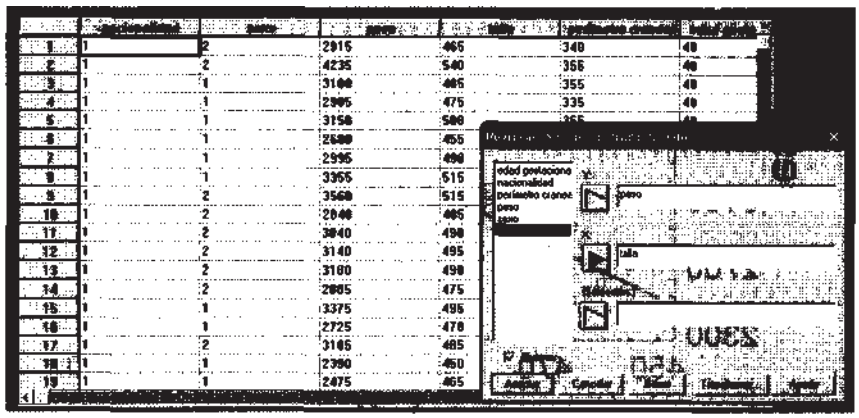


Figura C7.1. Fichero de datos correspondiente al Ejemplo C6.1 y ventana del procedimiento de regresión simple del peso sobre la talla.

La Tabla C7.1 muestra los resultados de la opción tabular **Resumen del procedimiento,** si bien se ha reducido su contenido original para ayudar a comprender los diferentes pasos de este procedimiento. Así, tras indicar las variables dependiente e independiente, se destacan los valores que definen la recta de regresión, la ordenada en el origen y la pendiente de la recta. Después, se destaca el valor R-cuadrado, que corresponde al coeficiente de determinación, medida ésta de la bondad del ajuste de la recta a los datos. Finalmente, hemos incluido la parte del StatAdvisor donde se muestra la ecuación de la recta de regresión.

Tabla C7.1. Ecuación de regresión del peso sobre la talla

Análisis de regresión - Modelo lineal Y = a + b * X

Variable dependiente: peso
Variable independiente: talla

Parámetro	**Estimación**	Error estándar	**Estadístico** T	**Valor** de P
Ordenada	**-4724.63**	643.977	-7.33664	0.0000
Pendiente	**16.0399**	1.30272	12.3126	0.0000

Coeficiente de correlación = 0.823405
R-cuadrado = 67.7996 porcentaje
El StatAdvisor
La salida muestra los resultados del ajuste al modelo lineal para describir la relación entre peso y talla. La ecuación del modelo ajustado es: peso = **-4724.63 + 16.0399 * talla**

La Figura C7.2 muestra el resultado de la opción gráfica **Gráfico del modelo ajustado.** Puede apreciarse la nube de puntos o diagrama de dispersión y la recta de regresión ajustada. Este gráfico es la primera aproximación para razonar y validar (gráficamente) la hipótesis de linealidad en la relación entre el peso y la talla. Observando dicha nube de puntos, parece razonable el ajuste lineal, si bien se aprecia una gran dispersión de los pesos para tallas predeterminadas. Esto último tiene su efecto en el valor del coeficiente de determinación, 67.79: sólo el 67.79% de la variación total del peso se explica por su relación lineal con la talla.

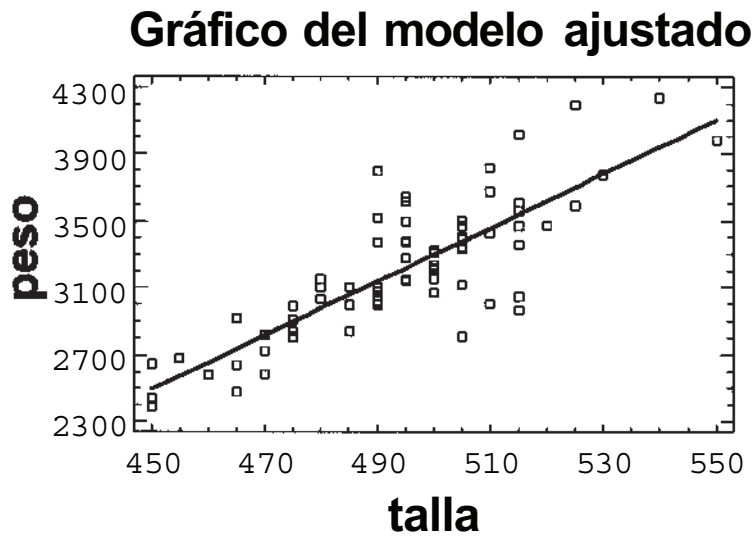


Figura C7.2. Nube de puntos y recta de regresión del peso sobre la talla.

Hasta aquí, lo que hemos intentado resolver es un problema de ajuste matemático de una ecuación a unos datos. Si deseamos hacer inferencias acerca de los parámetros que definen la recta de regresión teórica, debemos comprobar previamente algunos supuestos (normalidad, igualdad de varianzas del peso para diferentes niveles de la talla, y linealidad de los pesos medios para dichos niveles). El cuarto supuesto relativo a la toma de la muestra, la independencia entre los diferentes pesos no requiere contraste alguno, quedando garantizado por el tipo de muestreo utilizado. En el apartado **Análisis de los residuos,** que se verá posteriormente, se presenta una forma sencilla de comprobar estos supuestos.

Inferencias respecto a los parámetros

En la Tabla C7.1 aparecen dos contrastes de hipótesis acerca de los dos parámetros de la recta de regresión, la ordenada en el origen α y la pendiente β. Los dos contrastes formulan como hipótesis nulas las siguientes: $H_0 : \alpha = 0$, $H_0 : \beta = 0$. El primero de estos contrastes carece de interés en la mayoría de los casos, ya que raramente el punto de corte de la recta de regresión con el eje de ordenadas (ordenada en el origen) será el punto (0,0). Además, dicho punto de corte carece de significado casi siempre. En nuestro caso, la interpretación de α indica que sería el peso medio que correspondería a un recién nacido con talla 0 mm. El segundo contraste, el contraste de regresión, sí tiene importancia. El estadístico del test aparece en la columna «estadístico T» y vale 12.3126 (resultado de dividir el valor de la estimación entre su error estándar) y su valor p asociado (obtenido de la distribución t de Student con n-2 grados de libertad) es $p < 0.0001$, por lo que se rechaza la hipótesis nula y se acepta la alternativa $H_1 : \beta \neq 0$, es decir, existe una relación lineal significativa entre el peso y la talla. Lógicamente, podemos someter a prueba otras hipótesis acerca del valor de β, que representa el aumento medio de peso por cada incremento unitario de la talla. Por ejemplo, para contrastar que dicho incremento es de 15 gramos, $H_0 : \beta = 15$, el valor del estadístico del test sería (16.0399-15)/1.30272 = 0.798 cuyo valor p asociado (t de Student con 72 grados de libertad) es 0.4274, por lo que debería aceptarse el valor propuesto para β, en la hipótesis nula.

El STATGRAPHICS Plus no proporciona, de manera explícita, intervalos de confianza para ambos parámetros; sin embargo, resulta muy sencillo obtenerlos. La expresión genérica de dichos intervalos es «estimación ± *t* (error estándar)», siendo *t* la abscisa de una distribución t con 72 grados de libertad relativa al nivel de confianza establecido. Por ejemplo, los intervalos de confianza al 95% para a y β son, respectivamente:

$$-4724.63 \pm 1.993\ (643.9779) = (-6008.07, -3441.18) \text{ y}$$
$$16.0399 \pm 1.993\ (1.30272) = (13.44, 18.64)$$

El contraste de regresión, $H_0 : \beta = 0$ puede llevarse a cabo, de forma alternativa, mediante el análisis de la varianza. Este contraste, que se basa en la descomposición de la variabilidad total en dos sumandos, la variación explicada por la regresión y la no explicada, se presenta en la opción tabular que acabamos de comentar, **Resumen del procedimiento.** La Tabla C7.2 muestra esta parte de los resultados.

Tabla C7.2. Contraste de regresión. Análisis de la varianza

Análisis de la varianza					
Fuente	Suma de cuadrados	Gl	Cuadrado medio	Cociente-F	**Valor** de p
Modelo	7.74755E6	1	7.74755E6	151.60	0.0000
Residuo	3.67959E6	72	51105.4		
Total (Corr.)	1.14271E7	73			

Gl, grados de libertad.

El valor del estadístico del contraste es 151.60, resultado de dividir el cuadrado medio de la regresión, 7.74755E6, entre el cuadrado medio del error, 51105.4. El valor p asociado (F de Snedecor con 1 y 72 grados de libertad) es < 0.0001, por lo que, al igual que antes, se rechaza la hipótesis nula y podemos afirmar que existe una relación lineal significativa entre el peso y la talla. Este contraste es equivalente al anterior; los valores de p de ambos test son el mismo, debido a que el estadístico F de Snedecor es el cuadrado del estadístico t de Student. En nuestro caso, se puede comprobar que 151.60 es el cuadrado de 12.3126.

Predicciones

El STATGRAPHICS Plus proporciona predicciones del peso, para cualquier valor de la talla. Teniendo en cuenta que la capacidad predictiva de la recta de regresión se basa en el rango de valores observado para la variable predictora, en nuestro caso la talla, no deberíamos obtener predicciones del peso para valores de la talla que no estén comprendidos entre el mínimo y el máximo de la misma.

Eligiendo la opción tabular **Predicciones,** se muestran los valores predichos para el peso correspondientes a los valores mínimo, 450 mm, y máximo, 550 mm, de talla. Igualmente se proporcionan, para estas tallas, los intervalos de predicción del peso individual y de confianza del peso medio, ambos al 95% (véase apartado 11.5, del Cap. 11). Para obtener predicciones correspondientes a otras tallas comprendidas entre las anteriores, hemos de pulsar el botón derecho del ratón para abrir las **Opciones de ventana** y añadir en el campo *Predicción en X* los valores de la talla deseados. También se puede modificar aquí el nivel de confianza de los intervalos. La Tabla C7.3 incorpora la predicción correspondiente a la talla 500 mm, junto a las dos que obtiene, por defecto, este procedimiento. En la opción gráfica comentada anteriormente, **Gráfico del modelo ajustado,** se muestran, por defecto, las bandas de predicción y confianza. En la Figura C7.2 se suprimieron ambas bandas para destacar la recta de regresión ajustada. El gráfico completo se muestra en el apartado **Análisis de los residuos** de este Apéndice.

Tabla C7.3. Predicciones e intervalos de predicción y confianza al 95%

Valores predichos

		95% Límites de predicción		95% Límites de confianza	
X	Predicho Y	Inferior	Superior	Inferior	Superior
450.0	2493.33	2025.52	2961.13	2367.82	2618.84
500.0	3295.32	2841.36	3749.29	3240.61	3350.04
550.0	4097.32	3620.83	4573.81	3942.54	4252.09

Comparación del modelo lineal con otros no lineales

Una tercera opción tabular de interés, cuando el ajuste lineal resulta inadecuado, es **Comparación de modelos alternativos.** Si elegimos esta opción, podemos comparar hasta 11 modelos de regresión, no lineales, ajustados a nuestros datos. La comparación se hace a partir de los valores del coeficiente de determinación: en base a dichos valores, se ordenan los diferentes modelos, incluido el lineal, de mayor a menor R-cuadrado, es decir, del modelo que mejor se ajusta a los datos al modelo que peor lo hace. La Tabla C7.4 muestra los resultados de esta opción.

Como puede apreciarse en la Tabla C7.4, ninguno de los modelos alternativos mejora sustancialmente el ajuste lineal. Conviene puntualizar que la elección de un modelo de regresión debe tener en cuenta no sólo la bondad del ajuste numérico sino también la adecuación gráfica de los datos al mismo y, finalmente, su adecuación o explicación biológica.

Tabla C7.4. Comparación de modelos de regresión mediante el coeficiente de determinación

Comparación de modelos alternativos		
Modelo	**Correlación**	**R-cuadrado**
Doble inverso	0.8423	70.95%
Inverso-Y	-0.8358	69.86%
Curva-S	-0.8350	69.73%
Multiplicativo	0.8340	69.55%
Exponencial	0.8322	69.26%
Raíz cuadrada-Y	0.8285	68.63%
Raíz cuadrada-X	0.8234	67.80%
Lineal	**0.8234**	**67.80%**
Logarítmico-X	0.8233	67.78%
Inverso-X	-0.8225	67.64%
Logístico	Sin ajuste	
Log Probit	Sin ajuste	

Análisis de los residuos

El análisis de los residuos es una forma sencilla de contrastar, *aposteriori,* las hipótesis estructurales o supuestos del modelo de regresión lineal que resultan necesarios para validar las inferencias respecto a los parámetros. Este método, que siempre es aplicable, resulta especialmente útil cuando disponemos de pocos valores de la variable dependiente para algún o algunos valores de la variable independiente. En este último caso, no se pueden utilizar los métodos analíticos, vistos en los Apéndices C4 y C6, para contrastar la normalidad y la igualdad de varianzas, respectivamente.

Cada uno de los pesos contenidos en la base de datos proporciona un residuo, que se define como la diferencia entre dicho peso observado y el peso estimado según la recta de regresión. Observando la Figura C7.2, resulta sencillo comprobar cómo los puntos que quedan por encima de la recta de regresión proporcionarán residuos positivos, siendo negativos los correspondientes a los puntos que quedan por debajo de dicha recta.

La comprobación de la hipótesis de normalidad requiere el uso de los test de normalidad vistos en el Apéndice C4 pero aplicados a los residuos. Las hipótesis de igualdad de varianzas y linealidad se pueden comprobar con el análisis gráfico de los residuos. Para ilustrar el método, vamos a reducir la muestra limitando el estudio a los recién nacidos de nacionalidad española (nacionalidad = 1 en la base de datos) y vamos a analizar, para esta submuestra, la relación entre el peso y la talla. La recta de regresión obtenida es:

$$\textbf{peso} = \textbf{-5636.17 + 17.8899 * talla,}$$

y el coeficiente de determinación $R^2 = 87.17\%$, por lo que la relación lineal entre el peso y la talla es más fuerte para los recién nacidos españoles que para toda la muestra completa. La nube de puntos, la recta de regresión y las bandas de predicción y confianza al 95% se muestran en la Figura C7.3.

La igualdad de varianzas de las distribuciones de los pesos para las diferentes tallas se comprueba observando el gráfico de los residuos frente a los valores predichos del peso, o de manera equivalente, frente a las tallas. Para obtener ambos gráficos, debemos seleccionar las opciones gráficas **Residuos frente a Predicho y Residuos frente a X,** respectivamente. La primera de estas gráficas se muestra en la Figura C7.4. No se observa que la variabilidad de los residuos aumente o disminuya con la magnitud de las predicciones. Este mismo gráfico, residuos frente a valores predichos, resulta muy útil para detectar posibles desviaciones de la hipótesis de linealidad. La distribución

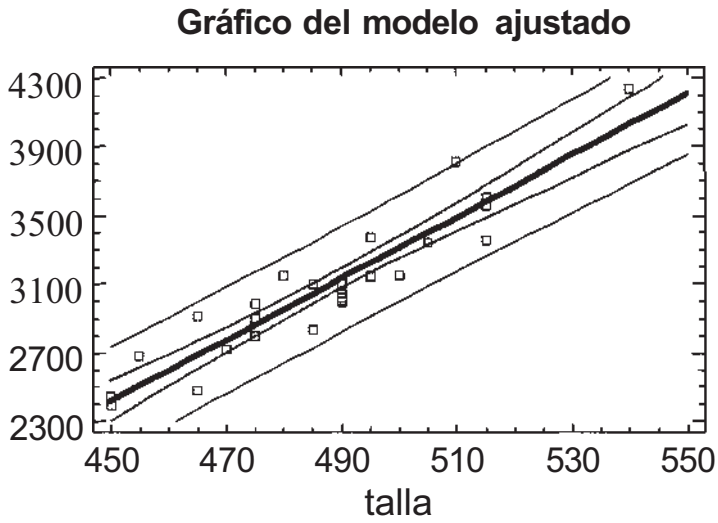


Figura C7.3. Nube de puntos, recta de regresión y predicciones del peso sobre la talla en los recién nacidos españoles.

aleatoria de los residuos alrededor de la línea horizontal revela la adecuación del modelo lineal para explicar la relación del peso con la talla. En las opciones de ventana de este gráfico podemos elegir *Residuos o Residuos estudentizados,* siendo esta última opción la que muestra la Figura C7.4. El aspecto de la gráfica es el mismo con cualquiera de las dos opciones; sólo cambian las unidades de medida de los residuos, que en el caso de los residuos estudentizados son unidades t de Student, por lo que resultan útiles, además, para detectar observaciones atípicas, que son aquellas cuyos residuos estudentizados son superiores a 2 en valor absoluto.

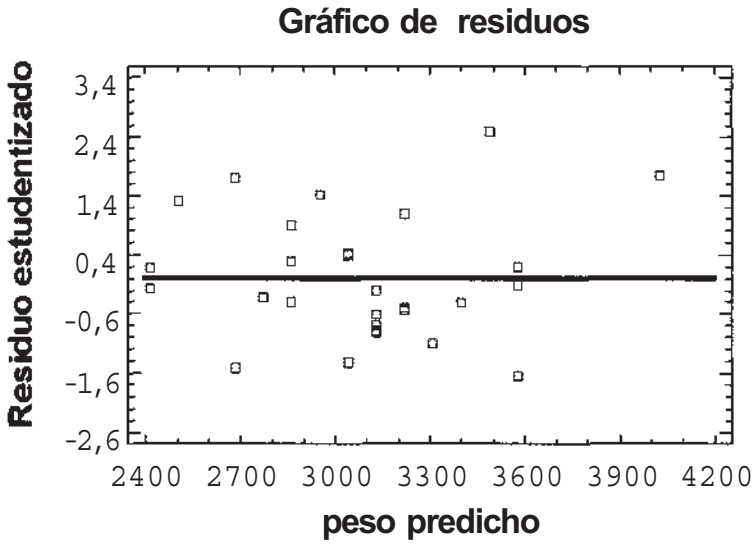


Figura C7.4. Gráfico de residuos frente a valores predichos.

C7.2. CORRELACIÓN LINEAL

Dado que las dos variables que estamos estudiando son aleatorias, podemos medir la asociación lineal entre ambas mediante el coeficiente de correlación lineal. El valor de este coeficiente, para la muestra completa, se presenta en la Tabla C7.1, antes del coeficiente de determinación R-cuadrado, y es 0.8234, lo que indica que existe, entre el peso y la talla, una correlación moderada positiva (véase Fig. 11.14 del Cap. 11). En el caso de la regresión lineal simple (una sola variable predictora), el coeficiente de determinación es el cuadrado del coeficiente de correlación lineal.

Conviene también recordar que, como en todos los estudios «observacionales» (las dos variables son aleatorias), tiene sentido preguntarse por la otra recta de regresión, la de la talla en función del peso. En los estudios «controlados» (los valores de la variable predictora se pueden preseleccionar), sólo tiene sentido la regresión de la variable dependiente sobre la predictora, y tampoco debe calcularse el coeficiente de correlación lineal.

El contraste de incorrelación o independencia lineal, $\mathbf{H_0}$: $\rho = 0$, $\mathbf{H_1}$: $\rho \neq 0$, es equivalente al contraste de regresión, $\mathbf{H_0}$: $\beta = 0$, $\mathbf{H_1}$: $\beta \neq 0$, por lo que su resultado ya ha sido comentado. Basta volver a la Tabla C7.1 y recordar que el estadístico del contraste vale 12.3126, con un valor p asociado menor que 0.0001, por lo que se rechaza la hipótesis de incorrelación y admitimos que existe una relación lineal entre el peso y la talla.

C7.3. REGRESIÓN LINEAL MÚLTIPLE

Para ilustrar este procedimiento vamos a utilizar, de nuevo, el Ejemplo C6.1 anterior. Deseamos estudiar la posible relación lineal entre el peso y la talla, el perímetro craneal y la edad gestacional, contemplando estas tres últimas como regresores o variables predictoras. El peso será la variable respuesta.

Como ocurre con cualquier modelo de regresión múltiple, el objetivo del análisis será la obtención de un modelo, lineal en este caso, que explique razonablemente la respuesta observada, y esto quiere decir que se debe elegir el modelo que mejor se ajuste a los datos con el menor número de regresores.

Para obtener la ecuación de regresión, debemos elegir el procedimiento **Dependencia, Regresión múltiple...** Una vez emerge la ventana del procedimiento, completaremos los campos como muestra la Figura C7.5.

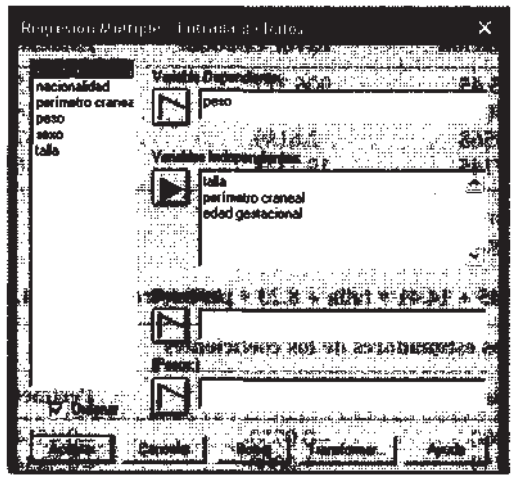


Figura C7.5. Pantalla de datos del procedimiento de regresión lineal múltiple.

La Tabla C7.5 muestra la parte sustancial de las opciones tabulares **Resumen del procedimiento y Matriz de correlaciones.** En la primera parte de dicha Tabla se destacan los valores de los coeficientes de regresión. Las tres variables predictoras tienen coeficientes de regresión significativos. El mayor valor de p corresponde a la edad gestacional, p = 0.0071; después se destaca el valor del coeficiente de determinación, ajustado por grados de libertad, R^2 = 72.7102%, lo que significa que el modelo de regresión obtenido explica el 72.71% de la variación total del peso; finalmente se muestra la ecuación de regresión. Debemos puntualizar que se llega al mismo modelo, en este caso, sea cual sea la estrategia de selección de variables utilizada, *Selección hacia adelante, hacia atrás, Todas las variables,* que podemos encontrar en las **Opciones de análisis** de este procedimiento. En la segunda parte de la Tabla C7.5 se muestra la matriz de correlaciones estimadas de los coeficientes del rnodelo de regresión. Su finalidad es detectar el problema de la multicolinealidad, la alta correlación entre algunas o todas las variables predictoras. El problema de la multicolinealidad es bastante complejo y escapa al contenido de este Apéndice. Los signos típicos de la multicolinealidad son el aumento considerable de los errores estándar de los coeficientes de regresión y valores muy altos de dichos coeficientes aun después de ser estandarizados. El STATGRAPHICS Plus aborda este problema en 1a opción **Matriz de correlaciones,** incluyendo el StatAdvisor de dicha opción una advertencia cuando alguna de las correlaciones es superior a 0.5 en valor absoluto. En nuestro caso, esto ocurre entre los coeficientes de la talla y el perímetro craneal, cuyo coeficiente de correlación vale -0.5634. La solución más simple a este problema es eliminar del modelo una de las dos variables. Eliminamos el perímetro craneal por ser la que presenta menor asociación lineal con el peso (comparando los coeficientes de correlación). El nuevo ajuste proporciona la siguiente ecuación, que llamaremos **modelo I:**

peso = -3338.93 + 17.1277 * talla - 49.1203 * edad gestacional

y el coeficiente de determinación ajustado es, ahora, R^2 = 69.2935%. Comparando ambos coeficientes R^2, podemos concluir que la eliminación del modelo de una variable asociada linealmente con la respuesta no causa una pérdida de ajuste significativa. Otra manera sencilla de ratificar el comentario anterior consiste en comparar los valores de cada coeficiente de regresión en ambos modelos En este caso, si comparamos los coeficientes de talla y edad gestacional, se observa que el orden de magnitud es parecido para ambos, por lo que el modelo I es el modelo final propuesto.

Tabla C7.5. Ecuación de regresión del peso sobre la talla, el perímetro craneal y la edad gestacional. Matriz de correlaciones

Análisis de regresión múltiple

Variable dependiente: peso

Parámetro	Estimación	Error estándar	Estadístico T	Valor de P
CONSTANTE	**-4625.45**	906.317	-5.10357	**0.0000**
Talla	**14.4086**	1.53481	9.38787	**0.0000**
Perímetro craneal	**8.23568**	2.6189	3.14471	**0.0024**
Edad gestacional	**-54.7145**	19.7338	-2.77263	**0.0071**

R-cuadrado = 73.8317 porcentaje
R-cuadrado (ajustado para g.l.) = 72.7102 porcentaje

peso = -4625.45 + 14.41 * talla + 8.23 * perímetro cran. - 54.71 * edad gestac.

Matriz de correlación de los estimadores de los coeficientes

	Constante	Talla	Perímetro	Edad gesta.
CONSTANTE	1.0000	-0.0759	-0.4514	-0.5686
Talla	-0.0759	1.0000	**-0.5634**	-0.2316
Perímetro	-0.4514	**-0.5634**	1.0000	-0.0901
Edad gestacional	-0.5686	-0.2316	-0.0901	1.0000

La Figura C7.6 muestra el resultado de la opción gráfica **Observado frente a Predicho.** Cuanto más próximos se encuentren los puntos a la línea diagonal, mejor será la capacidad predictiva del modelo propuesto, resultando pues, una alternativa gráfica para validar el ajuste del modelo a los datos.

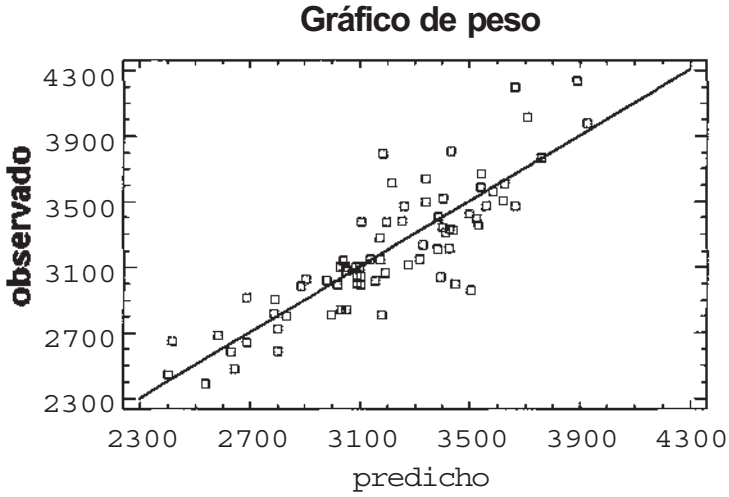
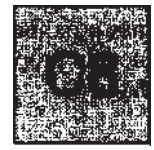


Figura C7.6. Gráfica de pesos observados frente a los pesos predichos por el modelo de regresión múltiple.



Contrastes para datos cualitativos

En el Capítulo 12, se ven los test para resolver los problemas de homogeneidad y de independencia de dos caracteres cualitativos, basados en el estadístico ji-cuadrado, X^2. Estos dos test tienen el mismo estadístico del contraste. Es por ello que los paquetes estadísticos no los distinguen a la hora de su cálculo. Deberá, por tanto, ser el investigador el que plantee las hipótesis adecuadas a los dos tipos de problemas.

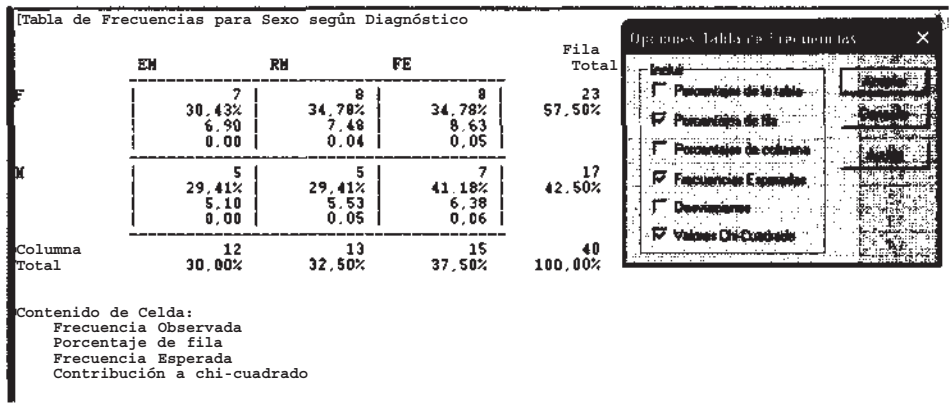
El STATGRAPHICS Plus permite analizar estos problemas cuando los datos vienen dados en dos formas: fichero de datos y datos resumidos en una tabla de contingencia. Veamos ambos casos a continuación. Hay que señalar que el STATGRAPHICS denomina chi-cuadrado en lugar de ji-cuadrado al estadístico, y en este Apéndice se seguirá esta denominación.

C8.1. FICHERO DE DATOS

Consideremos que tenemos los datos del Ejemplo 1.1.1 del Capítulo 1, en el que se dan los valores referentes a un geriátrico en el que se evalúa el estado mental de los pacientes. Entre otras variables medidas se tienen el sexo y el diagnóstico. Los datos presentados son parte de un estudio más amplio, y no se especifica cómo se han recogido. Supongamos que se ha fijado un grupo de hombres y otro de mujeres. En el Ejemplo 1.1.2, referente a estos mismos datos, se plantea el interés por ver si la distribución de los diagnósticos es la misma en hombres y en mujeres. Las hipótesis a contrastar en esta situación serían H_0: Las dos muestras son homogéneas, frente a H_1: Las dos muestras no son homogéneas.

Primero obtengamos la tabla de contingencia: elegiremos el procedimiento **Descripción, Datos cualitativos, Tabulación cruzada;** indicaremos qué variable queremos que esté en las filas (p. ej., sexo) y cuál en las columnas (p. ej., diagnóstico). Al pulsar **Aceptar,** aparecen por defecto las opciones tabulares **(Resumen del procedimiento y Tabla de frecuencias)** y opciones gráficas, comentadas detalladamente en el Apéndice C2. Si nos posicionamos en la ventana correspondiente a la opción **Tablas de frecuencias,** podemos pulsar el botón derecho del ratón y tendremos las **Opciones de ventana** (véase Fig. C8.1). Si marcamos **Frecuencias esperadas,** podremos ver qué frecuencia se espera en caso de ser cierta la hipótesis de homogeneidad. Esto permitirá ver si todas las celdillas tienen una frecuencia esperada al menos de 5, condición necesaria para la correcta aplicabilidad del test. Si marcamos **Desviaciones,** tendremos la diferencia entre la frecuencia observada y esperada en cada celdilla. Si marcamos **Valores chi-cuadrado,** veremos qué celdillas aportan

más al valor del estadístico del contraste, pues calcula el valor de desviaciones al cuadrado, dividido por la frecuencia esperada, es decir, qué celdillas difieren más de la hipótesis nula.



[Tabla de Frecuencias para Sexo según Diagnóstico

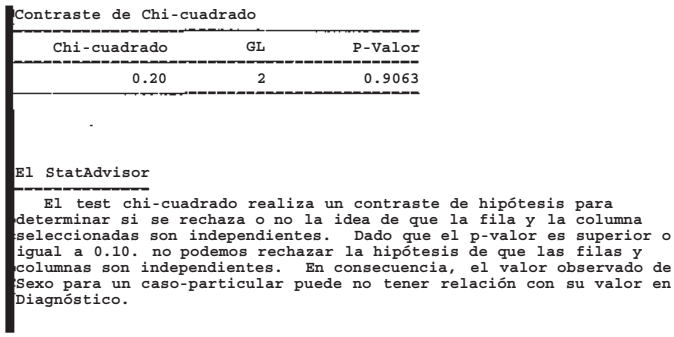
	EM	RM	FE	Fila Total
F	7 30,43% 6,90 0,00	8 34,78% 7,48 0,04	8 34,78% 8,63 0,05	23 57,50%
M	5 29,41% 5,10 0,00	5 29,41% 5,53 0,05	7 41,18% 6,38 0,06	17 42,50%
Columna Total	12 30,00%	13 32,50%	15 37,50%	40 100,00%

Contenido de Celda:
Frecuencia Observada
Porcentaje de fila
Frecuencia Esperada
Contribución a chi-cuadrado

Figura C8.1. Opciones de una tabla de frecuencias.

El valor del estadístico del contraste se calcula al marcar en las opciones tabulares la opción **Contraste de chi-cuadrado.** Observando la Figura C8.2, vemos que el valor del estadístico del contraste es 0.20 y su valor p asociado es 0.9063; luego no podemos rechazar la hipótesis nula de homogeneidad, es decir, no existe diferencia significativa en la distribución de las enfermedades mentales entre hombres (M) y mujeres (F).

Hay que hacer notar que, como ya indicamos al comienzo, el STATGRAPHICS no diferencia entre los problemas de independencia y homogeneidad. Por ello, el comentario del StatAdvisor (véase Fig. C8.2) siempre considera que el contraste es de independencia.



Contraste de Chi-cuadrado

Chi-cuadrado	GL	P-Valor
0.20	2	0.9063

El StatAdvisor

El test chi-cuadrado realiza un contraste de hipótesis para determinar si se rechaza o no la idea de que la fila y la columna seleccionadas son independientes. Dado que el p-valor es superior o igual a 0.10. no podemos rechazar la hipótesis de que las filas y columnas son independientes. En consecuencia, el valor observado de Sexo para un caso-particular puede no tener relación con su valor en Diagnóstico.

Figura C8.2. Test basado en el estadístico chi-cuadrado.

C8.2. DATOS EN TABLAS DE CONTINGENCIA

Consideremos ahora el caso donde los datos ya están resumidos en una tabla de contingencia, como la Tabla 12.17 del Ejercicio 12.2.2, del Capítulo 12. Allí se quiere estudiar si el tipo de sujeción de los ocupantes de un coche está relacionado con la magnitud de la lesión en los accidentes de tráfico. Para ello, se revisan 1000 accidentes y se estudian las dos variables, siendo las hipótesis a contrastar H_0: Las dos variables son independientes, frente a H_1: Las dos variables están relacionadas.

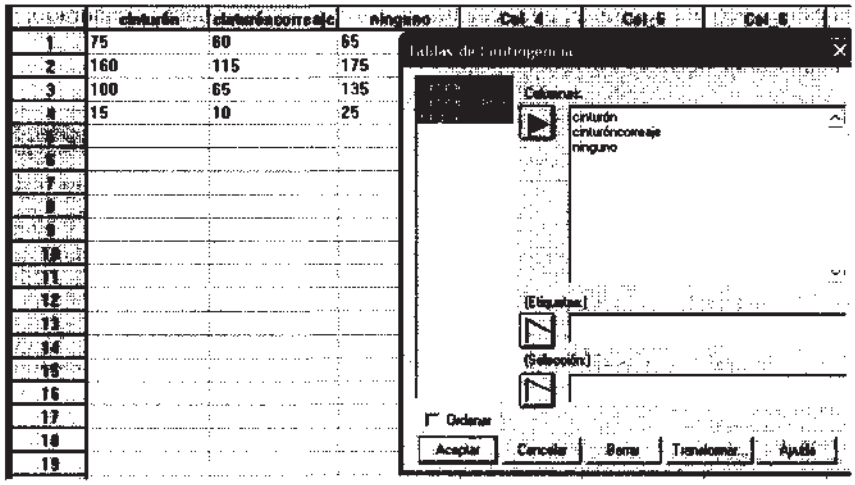


Figura C8.3. Entrada de datos para tablas de contingencia.

El STATGRAPHICS permite dar como entrada de datos dicha tabla. Para ello, debemos introducirla en la pantalla de datos según se observa en la Figura C8.3. Elegimos el procedimiento **Descripción, Datos cualitativos, Tablas de contingencia** y ahora en la ventana emergente debemos indicar qué columnas queremos analizar (véase Fig. C8.3); ya podemos pulsar **Aceptar.** Por defecto, se calcula el test basado en el estadístico chi-cuadrado. Las opciones gráficas y tabulares son las mismas que en el caso anterior. El valor del estadístico es 10.96 y su valor p asociado es 0.0897, con lo que no se rechazaría la hipótesis nula de independencia para **a** = 0.05. En la Figura C8.4 vemos el resultado del test y la tabla de frecuencias en la que se han pedido algunos datos de interés. El valor de p es muy próximo a 0.05, con lo que para otros niveles de significación, por ejemplo 0.1, sí se aceptaría que están relacionadas las dos variables. Cobra interés detenernos en las frecuencias esperadas. Así, si nos fijamos por ejemplo, en la categoría de «Muerte», vemos que las frecuencias esperadas, calculadas bajo el supuesto de independencia, para las categorías cinturón y cinturón con correaje, son mayores que las frecuencias observadas, es decir, se han producido menos muertes de las que cabría esperar en caso de ser independientes. En el caso de «Ninguna» sucede lo contrario.

	cinturón	cinturón correaje	ninguno	Total
Ninguna	75 70.00 0.36	60 50.00 2.00	65 80.00 2.81	200 20.00*
Menor	160 157.50 0.04	115 112.50 0.06	175 180.00 0.14	450 45,00%
Mayor	100 105.00 0.24	65 75.00 1.33	135 120.00 1,88	300 30,00%
Muerte	15 17.50 0.36	10 12.50 0.50	25 20.00 1.25	50 5,00%
Columna Total	350 35.00%	250 25.00%	400 40,00%	1000 100,00%

Contenido de Celda:
Frecuencia Observada — frecuencia Esperada — Contribución a chi-cuadrado

Contraste de Chi-cuadrado

Chi-cuadrado	GL	P-Valor
10.96	6	0,0897

Figura C8.4. Tabla de contingencia y contraste chi-cuadrado.

NOTAS

I) La distribución del estadístico Chi-cuadrado es asintótica. Es por ello que, en el caso de las tablas 2×2, algunos autores creen conveniente la corrección de dicho estadístico. El nombre que recibe dicha corrección es de Yates y se calcula, por defecto, en el STATGRAPHICS Plus, a la vez que el test chi-cuadrado. La utilización o no de dicha corrección es tema actual de controversia entre los estadísticos.

En el apartado 12.1 del Capítulo 12, al final de PRUEBA DE INDEPENDENCIA, se comenta que un test alternativo para muestras pequeñas, en tablas 2 x 2, es el test de Fisher. Dicho test se calcula, por defecto, con el STATGRAPHICS Plus al pedir el test chi-cuadrado cuando las muestras son pequeñas.

II) Hay que señalar que si tenemos una tabla 2 x 2, el contraste chi-cuadrado para homogeneidad de muestras coincide con el contraste bilateral de igualdad de dos proporciones independientes visto en el apartado C5.2 del Apéndice C5.

Contrastes no paramétricos

En este Apéndice se tratan los métodos no paramétricos de contraste de hipótesis. Se utilizan cuando tenemos en estudio una variable continua, y no puede suponerse que siga una distribución normal. Tratamos el caso de una, dos y k poblaciones, así como el caso de muestras independientes y relacionadas.

C9.1. UNA POBLACIÓN

En el caso de una población, cuando no hay normalidad, se proponen en el apartado 13.2 del Capítulo 13 dos test no paramétricos. El STATGRAPHICS Plus calcula estos dos test pero en su versión aproximada a una normal, válida sólo para muestras grandes. El test denominado «Contraste de los signos para la mediana», en el Capítulo 13, se llama en el STATGRAPHICS Plus «contraste de los signos». El test «Contraste de los rangos de signos de Wilcoxon» se llama en el STATGRAPHICS Plus «contraste de los rangos con signo». Veamos cómo calcular estos test, que serían las alternativas no paramétricas al contraste de la media de una distribución normal basado en la t de Student.

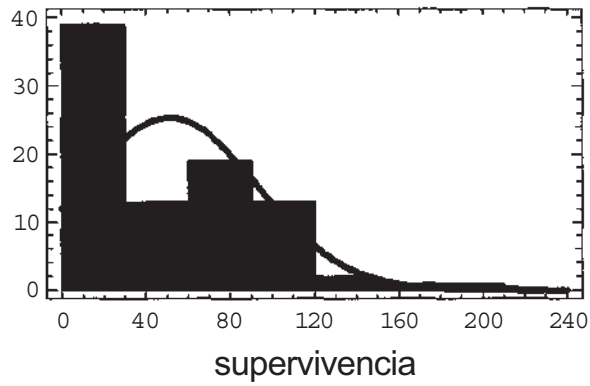
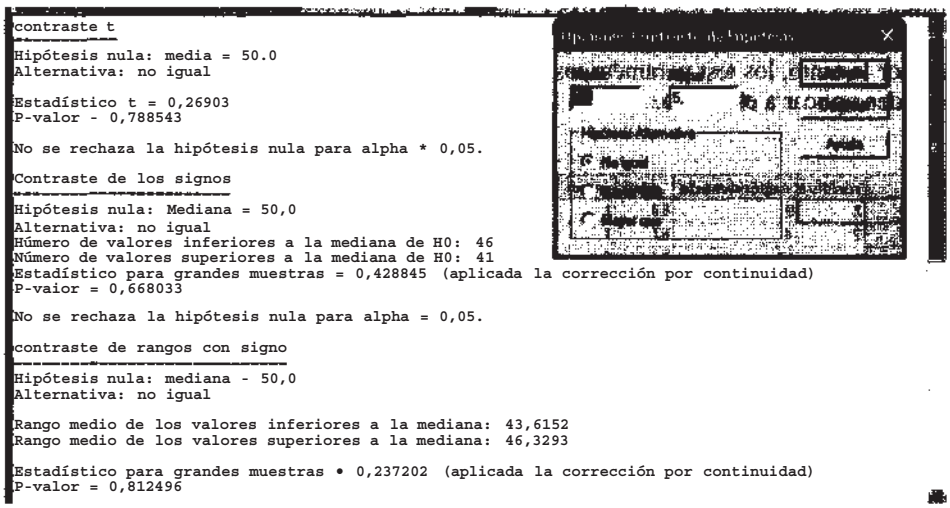


Figura C9.1. Ajuste gráfico de los datos.

Consideremos que tenemos los datos del Ejemplo C2.1 del Apéndice C2. Si nos fijamos en la variable supervivencia global y estudiamos su normalidad (según Apéndice C3), vemos que no sigue una distribución normal (p = 0.01892). La Figura C9.1 muestra el ajuste gráfico (véase el apartado C3.4 del Apéndice C3). Supongamos ahora que queremos contrastar que su mediana es 50 meses, para **a** = 0.05.

Elegimos el procedimiento **Descripción, Datos numéricos, Análisis unidimensional,** indicamos que la variable es **supervivencia** y pulsamos **Aceptar.** Los contrastes no se ejecutan por defecto, así que buscaremos, en las opciones **tabulares,** la **opción Contraste de hipótesis.** Se calculan, entonces, tres contrastes: el primero es el test **de** la t de **Student para** contrastar la media de una distribución normal y otros dos test no paramétricos, comentados anteriormente. Por defecto, el valor que se contrasta para la mediana es 0. Por ello, pulsamos el botón derecho del ratón, **Opciones de ventana** y en el campo **Media** pondremos 50. Nótese que nuestro contraste es sobre la mediana pero, al ejecutarse conjuntamente con el test de la media, en el campo pone genéricamente **Media.** En la Figura C9.2 vemos parte de la salida; en este caso, al no ser simétrica la variable, según las recomendaciones del apartado 13.2 del Capítulo 13 parece más adecuado utilizar el test «contraste de los signos». Por lo tanto, como el valor de p es 0.668033 no se rechaza la hipótesis nula, es decir, no hay evidencia estadística para decidir que la mediana sea distinta de 50.



```
contraste t
-----------
Hipótesis nula: media = 50.0
Alternativa: no igual

Estadístico t = 0,26903
P-valor - 0,788543

No se rechaza la hipótesis nula para alpha * 0,05.

Contraste de los signos
-----------------------
Hipótesis nula: Mediana = 50,0
Alternativa: no igual
Húmero de valores inferiores a la mediana de H0: 46
Número de valores superiores a la mediana de H0: 41
Estadístico para grandes muestras = 0,428845  (aplicada la corrección por continuidad)
P-vaior = 0,668033

No se rechaza la hipótesis nula para alpha = 0,05.

contraste de rangos con signo
-----------------------------
Hipótesis nula: mediana - 50,0
Alternativa: no igual

Rango medio de los valores inferiores a la mediana:  43,6152
Rango medio de los valores superiores a la mediana:  46,3293

Estadístico para grandes muestras • 0,237202  (aplicada la corrección por continuidad)
P-valor = 0,812496
```

Figura C9.2. Contrastes para una población.

C9.2. DOS POBLACIONES

Muestras independientes

Consideremos el Ejemplo C2.1 del Apéndice C2; supongamos que estamos interesados en ver si varía la supervivencia global entre hombres y mujeres. En la Figura C9.3 se ha representado un gráfico conjunto (que podríamos denominar histograma doble) del tiempo de supervivencia en hombres y en mujeres. La observación de dicha figura nos dice que no hay normalidad. Si esto se quiere comprobar mediante un contraste de hipótesis, tendríamos que hacer dos test de bondad de ajuste a la normal, uno para hombres y otro para mujeres.

Para contrastar la normalidad en hombres procederíamos según se indica en el apartado 3.4 del Apéndice C3, seleccionando la muestra según el Apéndice C2, es decir, con la base de datos numérica. Pulsando el procedimiento **Descripción, Distribuciones, Ajuste de distribuciones (Datos no censurados)** y seleccionando por **sexo** (hombres está codificado como 1), indicamos la variable a analizar, que es **supervivencia** (véase Fig. C9.4). El test de Kolmogorov se ejecuta por defecto, y

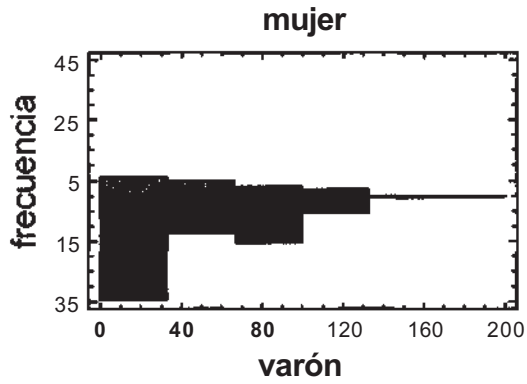


Figura C9.3. Histograma para edad en hombres y mujeres.

observamos que el valor de p es 0.0319353, con lo que se rechaza la normalidad para $\alpha = 0.05$. No haría ya falta hacer un test de normalidad para mujer, pues al menos para hombres la edad no es normal; por lo tanto, los test paramétricos para contrastar dos muestras, vistos en el Apéndice C5, no se pueden aplicar a este caso.

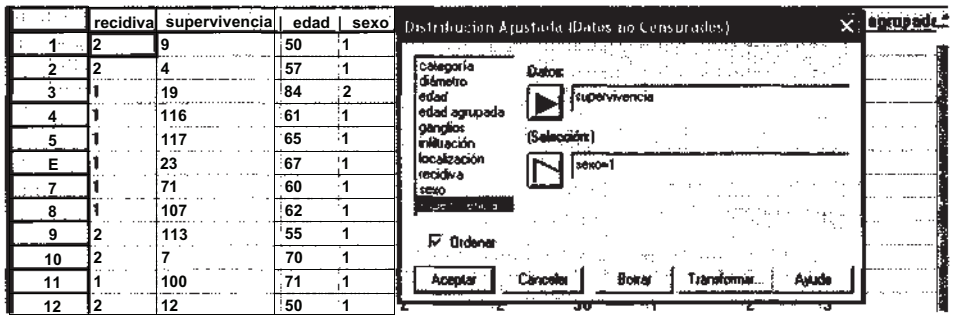


Figura C9.4. Elección de una submuestra.

Para comparar el tiempo de supervivencia entre hombres y mujeres utilizaremos la alternativa no paramétrica vista en el apartado 13.4, del Capítulo 13, «Contraste de la suma de los rangos de Wilcoxon». El STATGRAPHICS Plus realiza una aproximación normal del test de Wilcoxon, sólo válida para muestras grandes. Elegiremos el procedimiento **Comparación, Dos muestras, Comparación de dos muestras.** Para indicar las variables a utilizar, tenemos que marcar en el campo *Entrada:* **Columnas de código y datos,** y ya podremos indicar como *datos* la variable **supervivencia** y como *código de muestra* el **sexo** (véase Fig. C9.5). Por defecto, en las opciones gráficas, sale el histograma doble de la Figura C9.3. Al pulsar las opciones tabulares, marcaremos la opción **Comparación de medianas.** El test que se calcula es el de Wilcoxon, y el valor de p es 0.7372 (véase Fig. C9.5), con lo que podemos decir que no existe diferencia significativa entre las medianas de tiempo de supervivencia de hombres y mujeres.

En el apartado 13.4 del Capítulo 13, se comenta la equivalencia entre el test de Wilcoxon y el test de Mann-Whitney, equivalencia que justifica el nombre «contraste W de Mann-Whitney (Wilcoxon)», con el que aparece en el STATGRAPHICS Plus.

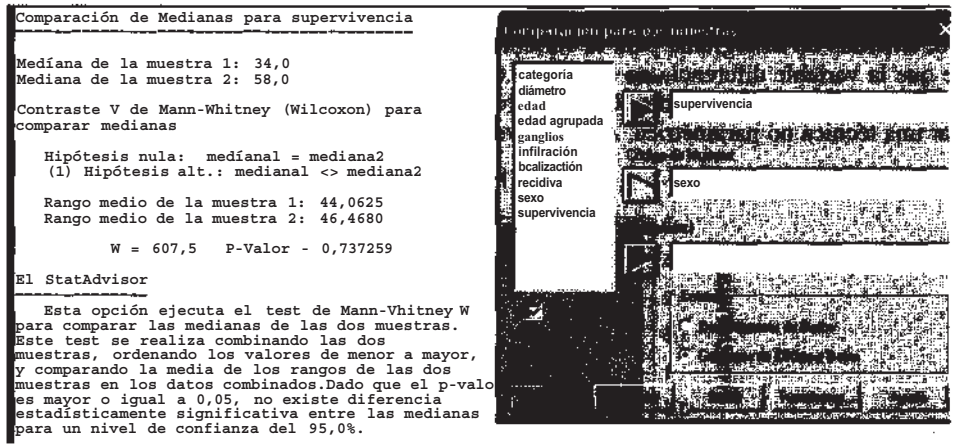


Figura C9.5. Test no paramétrico de comparación de dos poblaciones.

Muestras apareadas

Como se dice en el apartado 13.3 del Capítulo 13, el caso de muestras apareadas es una extensión del caso de una muestra, considerando la variable diferencia como única variable. Por ello, tendremos aquí las mismas consideraciones que para una muestra, es decir, se realizan los dos contrastes vistos en el Capítulo 13, pero en versión aproximada a la normal, sólo válida para muestras grandes.

Resolvamos el Ejercicio 13.3.3 del Capítulo 13, donde se quiere estudiar si la tensión arterial disminuye después de intervenir para eliminar una obstrucción renal en enfermos con función renal deteriorada. Para ello, se toma la tensión antes y después de la operación a 10 pacientes.

Los datos deben introducirse según la Figura C9.6; después, elegiremos el procedimiento **Comparación, Dos muestras, Comparación de muestras pareadas,** e indicaremos los datos a utilizar según se ve en la Figura C9.6.

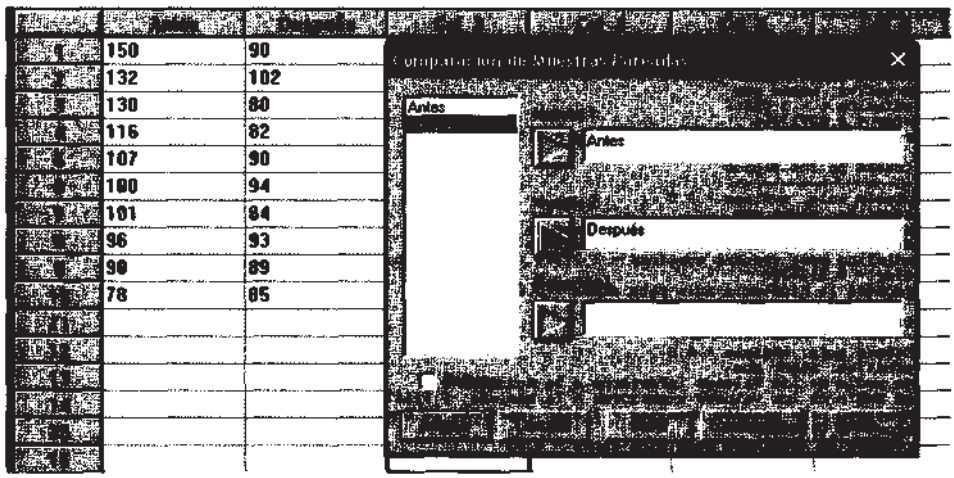


Figura C9.6. Entrada de datos para muestras apareadas.

Si pulsamos las opciones gráficas, una de ellas es el **Histograma de frecuencias.** Al pedir esta opción, tenemos un histograma para la variable diferencia. En la Figura C9.7, se aprecia gráficamente que la variable diferencia no sigue una distribución normal (10 son pocos datos para hacer un test de los vistos en el Apéndice C3, y también para dibujar un histograma), de ahí lo adecuado de usar una técnica no paramétrica. Además, vemos la falta de simetría de dicha distribución.

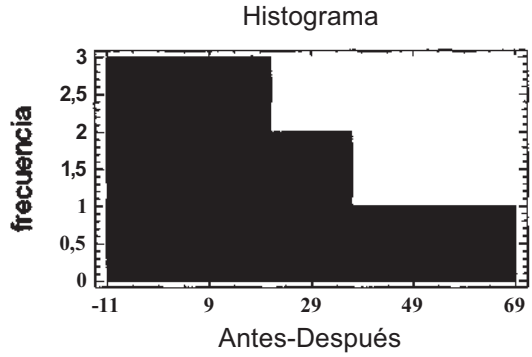


Figura C9.7. Representación gráfica para la variable diferencia.

Para realizar los contrastes no paramétricos, abrimos las opciones tabulares y marcamos la opción **Contraste de hipótesis.** Se calculan por defecto los test bilaterales (1 paramétrico y 2 no paramétricos), siendo la hipótesis nula la igualdad a 0 (para media y mediana, alternativamente!). Pulsamos el botón derecho del ratón y en **Opciones de ventana** marcamos *Mayor que,* ya que al estar trabajando con la variable diferencia «Antes-Después», si ha disminuido la tensión la diferencia será positiva, es decir, mayor de 0. La salida se recoge en la Figura C9.8. Como puede verse, el primer test es el de la t de Student para muestras apareadas (véase Apéndice C5). El test llamado «Contraste de los signos» es una aproximación normal, del test explicado en «Contraste de los signos para la mediana de la diferencia», en el apartado 13.3 del Capítulo 13. El test llamado «contraste de rangos con signo» es una aproximación normal del test explicado en «Contraste de los rangos de signo de Wilcoxon: datos emparejados», en el apartado 13.3 del Capítulo 13, donde se ofrecen recomendaciones de utilización.

```
Contraste de Hipótesis para Antes-Después
Media muestral = 21.1
Mediana muestral = 17.0

contraste t
-----------
Hipótesis nula: media = 0.0
Alternativa: mayor que
Estadístico t = 3.02566
P=valor - 0.00717356
Se rechaza la hipótesis nula para alpha = 0.05.

Contraste de los signos
-----------------------
Hipótesis nula: Mediana = 0.0
Alternativa: mayor que
Número de valores inferiores a la mediana de H0: 1
Húmero de valores superiores a la mediana de H0: 9
Estadístico para grandes muestras = 2.21359 (aplicada la corrección por continuidad)
P-valor - 0.0134283
Se rechaza la hipótesis nula para alpha - 0.05.

contraste de rangos con signo
-----------------------------
Hipótesis nula: mediana = 0,0
Alternativa: mayor que
Rango medio de los valores inferiores a la mediana: 4.0
Rango medio de los valores superiores a la mediana: 5.66667
Estadístico para grandes muestras = 2.3459 (aplicada la corrección por continuidad)
P-valor = 0.00949053
Se rechaza la hipótesis nula para alpha = 0.05.
```

Figura C9.8. Contraste, no paramétrico, para dos muestras apareadas.

Dada la falta de simetría de la variable diferencia (véase Fig. C9.7), se debería aplicar el contraste de los signos. Su valor de p es 0.0134283, luego se aceptaría la eficacia de la operación para $\alpha = 0.05$. Sin embargo, hay que hacer notar que el pequeño tamaño de la muestra no hace aconsejable la utilización de la aproximación normal. Sería más correcto resolver el problema según el apartado 13.3 del Capítulo 13, es decir, como en la salida vemos que sólo un dato está por debajo del valor de la mediana que figura en la hipótesis nula, al consultar la tabla de la binomial (10,0.5), vemos que su valor de p es 0.0107, con lo que llegaríamos a la misma conclusión.

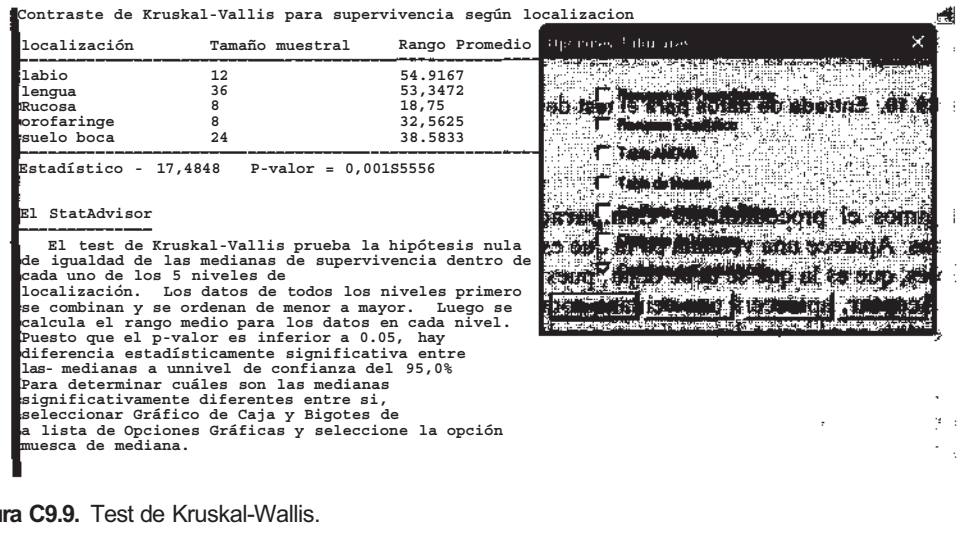
C9.3. K POBLACIONES

Muestras independientes

En caso de k poblaciones independientes, cuando no se puede admitir la hipótesis de normalidad previa del ANOVA, se deberá utilizar la técnica no paramétrica vista en el apartado 13.5 del Capítulo 13, es decir, el test de Kruskal-Wallis. Si estamos analizando nuestros datos con STATGRAPHIS Plus, y es la hipótesis (previa a un ANOVA) de igualdad de varianzas la que no se verifica, también utilizaremos el test de Kruskal-Wallis. Otros paquetes estadísticos incluyen generalizaciones al test de la t de Welch (visto en el apartado C5.1 del Apéndice C5) para más de 2 muestras (véase Apéndice D6).

Consideremos el Ejemplo C2.1 del Apéndice C2, y supongamos que queremos ver si la supervivencia global varía con la localización del tumor. Para poder aplicar un ANOVA de una vía, lo primero a comprobar es si el tiempo sigue una distribución normal en cada tipo de localización. Aquí, la variable localización tiene 5 categorías y para dos de ellas (mucosa y bucofaringe) hay tan sólo 8 datos con lo que no podríamos comprobar la normalidad con un test. Si pintamos el tiempo de supervivencia para cada localización, vemos que no parece darse el supuesto de normalidad. Usaremos por tanto la alternativa no paramétrica, que es el test de Kruskal-Wallis.

Elegimos el procedimiento **Comparación, Análisis de la varíanza, ANOVA simple;** en la ventana que emerge indicamos como *variable dependiente* la **supervivencia,** y *como factor* la **localización.** Al pulsar **Aceptar,** el test que buscamos no se ejecuta por defecto, así que en las opciones tabulares marcamos **Contraste de Kruskal-Wallis** (véase Fig. C9.9). En la Figura C9.9 se muestran también los resultados para esta opción: observamos que el valor de p es 0.0015, con lo que podemos concluir que el tiempo de supervivencia mediano difiere en al menos una de las localizadones con respecto a las otras, para cualquier $\alpha > 0.002$.



Contraste de Kruskal-Vallis para supervivencia según localizacion

localización	Tamaño muestral	Rango Promedio
labio	12	54.9167
lengua	36	53,3472
Rucosa	8	18,75
orofaringe	8	32,5625
suelo boca	24	38.5833

Estadístico - 17,4848 P-valor = 0,001S5556

El StatAdvisor

El test de Kruskal-Vallis prueba la hipótesis nula de igualdad de las medianas de supervivencia dentro de cada uno de los 5 niveles de localización. Los datos de todos los niveles primero se combinan y se ordenan de menor a mayor. Luego se calcula el rango medio para los datos en cada nivel. Puesto que el p-valor es inferior a 0.05, hay diferencia estadísticamente significativa entre las- medianas a unnivel de confianza del 95,0% Para determinar cuáles son las medianas significativamente diferentes entre si, seleccionar Gráfico de Caja y Bigotes de a lista de Opciones Gráficas y seleccione la opción muesca de mediana.

Figura C9.9. Test de Kruskal-Wallis.

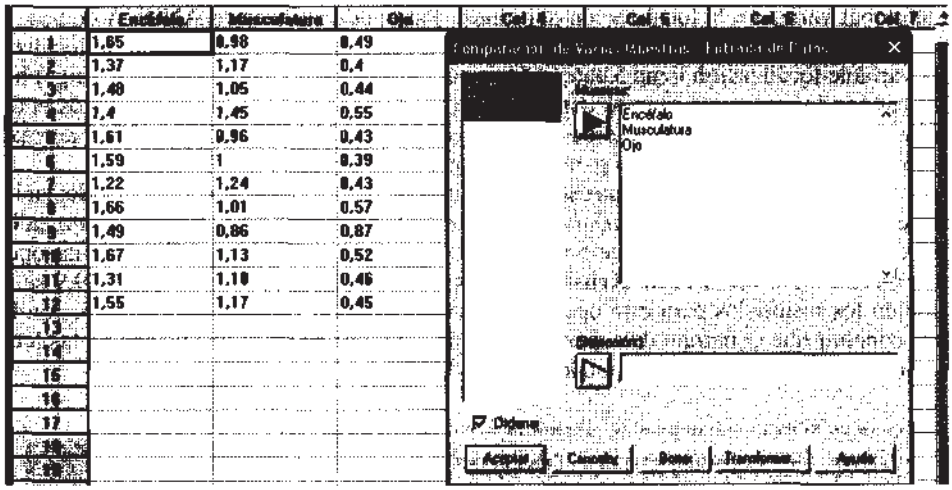
Ahora deberíamos encontrar cuáles de las localizaciones son las que producen la diferencia. El STATGRAPHICS Plus no realiza ningún test *a posteriori* del test de Kruskal-Wallis. En las opciones gráficas se sugiere (véase StatAdvisor correspondiente) un método gráfico basado en el diagrama de cajas y bigotes explicado en el Apéndice C2.

Hay que hacer notar otra forma alternativa de entrada de datos para poder realizar el test (de Kruskal-Wallis, que se comentará en la NOTA del test de Friedman (véase a continuación).

Muestras apareadas

Cuando tenemos k muestras que están relacionadas y no se cumple la hipótesis de normalidad del ANOVA correspondiente (Bloques completos aleatorizados, apartado 10.4 del Capítulo 10), la alternativa no paramétrica vista en el apartado 13.6 del Capítulo 13 es el test de Friedman.

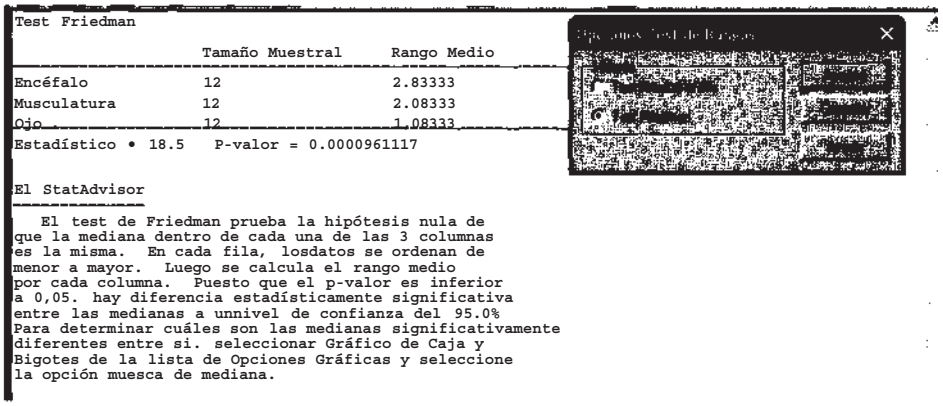
Resolvamos el Ejercicio 13.6.2 del Capítulo 13. Se toman 12 truchas sometidas a dosis subletales de metilo de mercurio. Se quiere ver si existen diferencias significativas entre las medianas de concentración del mercurio en tres localizaciones del cuerpo. Para ello, se estudia dicha concentración en tres localizaciones (encéfalo, musculatura y ojo) de cada trucha. La pantalla de datos debe tener, en este caso, la estructura de la Figura C9.10.



	Encéfalo	Musculatura	Ojo
1	1,65	0,98	0,49
2	1,37	1,17	0,4
3	1,40	1,05	0,44
4	1,4	1,45	0,55
5	1,61	0,96	0,43
6	1,59	1	0,39
7	1,22	1,24	0,43
8	1,66	1,01	0,57
9	1,49	0,86	0,87
10	1,67	1,13	0,52
11	1,31	1,10	0,46
12	1,55	1,17	0,45
13			
14			
15			
16			
17			
18			
19			

Figura C9.10. Entrada de datos para el test de Friedman.

Elegimos el procedimiento **Comparación, Muestras múltiples, Comparación de varias muestras.** Aparece una ventana en la que está marcada, por defecto, la opción **Columnas de datos múltiples,** que es la que se debe dejar, pues sólo en dicha opción se calcula el test de Friedman. Al pulsar **Aceptar,** aparecerá una ventana como la de la Figura C9.10. Al marcar las muestras y pulsar **Aceptar,** se ejecutan las opciones por defecto, y el test que buscamos no está entre ellas. Entre las opciones tabulares marcamos **Test de Kruskal-Wallis y Friedman, y** al pulsar **Aceptar** se cal cula por defecto el test de Kruskal-Wallis, así que habrá que pulsar el botón derecho del ratón y en **Opciones de ventana** marcar **Test de Friedman.** Los resultados se encuentran en la Figura C9.11. Observamos que el valor de p es 0.00009, de lo que se concluye que la concentración mediana de metilo de mercurio es distinta en al menos una de las localizaciones.



Test Friedman

	Tamaño Muestral	Rango Medio
Encéfalo	12	2.83333
Musculatura	12	2.08333
Ojo	12	1.08333

Estadístico • 18.5 P-valor = 0.0000961117

El StatAdvisor

El test de Friedman prueba la hipótesis nula de que la mediana dentro de cada una de las 3 columnas es la misma. En cada fila, losdatos se ordenan de menor a mayor. Luego se calcula el rango medio por cada columna. Puesto que el p-valor es inferior a 0,05. hay diferencia estadísticamente significativa entre las medianas a unnivel de confianza del 95.0% Para determinar cuáles son las medianas significativamente diferentes entre si. seleccionar Gráfico de Caja y Bigotes de la lista de Opciones Gráficas y seleccione la opción muesca de mediana.

Figura C9.11. Test de Friedman.

NOTA

Siguiendo la secuencia de procedimientos que acabamos de realizar, vemos que tendríamos una forma alternativa de introducir los datos para el test de Kruskal-Wallis. La pantalla de datos contendría los valores de la variable en estudio en columnas distintas para cada nivel del factor.

Según lo que acabamos de ver en esta forma de introducir los datos, el ordenador tiene la misma pantalla de datos para los dos test y es el investigador el que sabe si son datos apareados (test de Friedman) o independientes (test de Kruskal-Wallis). El compartir procedimientos de ejecución puede llevar a alguna confusión a la hora de valorar las opciones gráficas *a posteriori.* En concreto, nos referimos al diagrama de cajas y bigotes y los intervalos que allí se calculan, pues parecen estar disponibles (véase StatAdvisor) para ambos test, cuando en realidad sólo se calculan para el test de Kruskal-Wallis.

APÉNDICE 

Métodos estadísticos con SPSS

Agustín Turrero y Pilar Zuluaga

En este Apéndice expondremos el manejo del paquete estadístico SPSS, mostrando los procedimientos adecuados para realizar los análisis presentados a lo largo del libro. Para ilustrar dichos procedimientos, revisaremos algunos de los ejemplos citados en el texto principal, así como los dos ejemplos biológicos presentados en el Apéndice C.

En algunos de los análisis estadísticos se incluirán técnicas y procedimientos alternativos a los presentados en el libro, sin pretender ser exhaustivos, pero con la finalidad de optimizar los recursos que ofrece este paquete estadístico.

Finalmente, se añadirán comentarios que faciliten la interpretación de los resultados obtenidos y que amplíen lo aprendido en los capítulos precedentes.

Introducción al SPSS

Este apartado no pretende ser un manual exhaustivo de SPSS, sino solamente una breve guía que permita ejecutar los procedimientos estadísticos vistos en el texto del libro con el paquete estadístico SPSS.

La primera pantalla ejecutable que aparece en SPSS, después de la carátula, es la que se muestra en la Figura D1. 1.

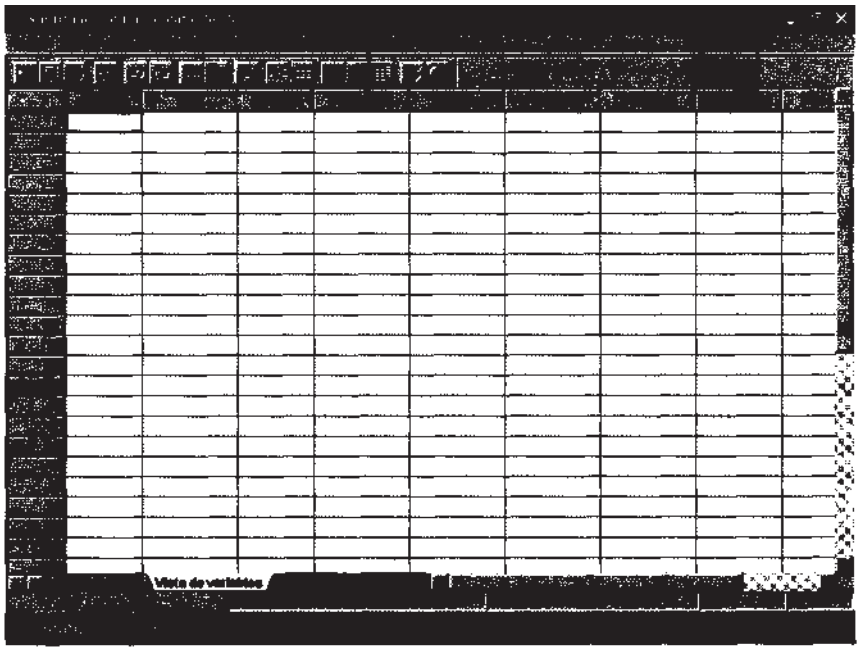


Figura D1.1. Pantalla inicial.

Distinguimos en ella tres elementos que nos permitirán comunicarnos con SPSS para realizar nuestros análisis: la barra de menú, la barra de herramientas y la pantalla de datos. Describimos a continuación cada uno de estos tres elementos.

D1.1. BARRA DE MENÚ

Aparece siempre, y está formada por 10 procedimientos (véase Fig. DI.2).

Figura D1.2. Barra de menú.

Según nos posicionamos en cada procedimiento y hacemos un clic aparece un menú emergente con las distintas posibilidades, algunas de ellas además acaban en una punta de flecha, que indica la existencia de un submenú, el cual vuelve a emerger al posicionamos encima con el ratón (véase Fig. D1.3).

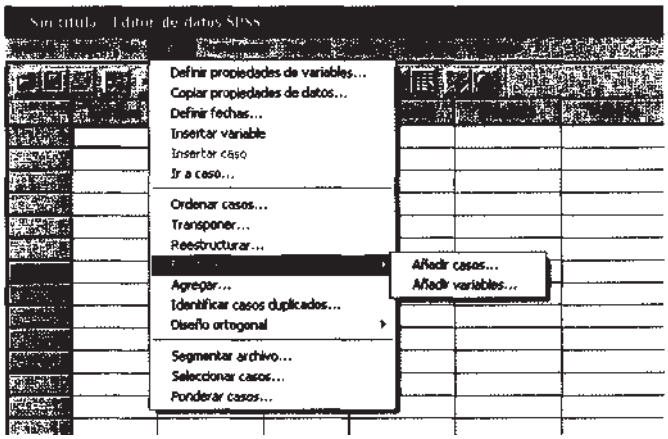


Figura D1.3. Menús emergentes de la barra de menú.

Comentaremos brevemente los 10 procedimientos:

- Archivo: Permite realizar tareas generales con los ficheros de datos como **Abrir, Cerrar, Guardar, Imprimir, Salir...,** algunos tienen submenús.
- **Edición:** Hace las mismas funciones que los procedimientos del mismo nombre en cualquier otra aplicación del entorno Windows, como **Copiar, Cortar, Pegar....**
- **Ver:** Sirve para controlar distintos elementos de las diversas barras.
- **Datos:** Permite realizar diferentes modificaciones al fichero de datos, ya sea **Añadir variables, Seleccionar casos** y, especialmente útil, **Ponderar casos** (véase Fig. D1.3).
- **Transformar:** Sirve esencialmente para transformar variables.
- **Analizar:** Este procedimiento contiene las distintas técnicas estadísticas que se pueden realizar con SPSS.
- Gráficos: Este procedimiento permite realizar diversos gráficos de interés estadístico.
- **Utilidades:** Permite ver información sobre los ficheros.
- **Ventana:** Hace las mismas funciones que los procedimientos del mismo nombre en cualquier otra aplicación del entorno Windows.
- ? : Es la ayuda sobre diversos temas relacionados con SPSS, contiene un «Tutorial».

D1.2. BARRA DE HERRAMIENTAS

Consta de distintos iconos, que realizan las mismas funciones que los procedimientos del mismo nombre en cualquier otra aplicación del entorno Windows. Para ejecutarlos bastará con situarnos en el icono correspondiente, y hacer clic para convertirlo en operativo (véase Fig. D1.4). De especial interés es el icono , que permite consultar y repetir los últimos procedimientos ejecutados.

Figura D1 .4. Barra de herramientas.

D1 .3. PANTALLA DE DATOS

La pantalla de datos consta de dos hojas: **Vista de datos** y **Vista de variables.** Se puede cambiar de una a otra pulsando las dos pestañas que con ese nombre aparecen abajo a la izquierda de la Figura D1.5. La Pantalla de datos (Hoja Vista de datos) es muy parecida a la hoja de datos de otros programas, por ejemplo, los datos del Ejemplo C2.1 del Apéndice C2, codificados numéricamente, se pueden ver en la Figura D1.5. Lo más habitual es que cada fila corresponda a los datos de un mismo individuo referentes a distintas características (variables) que aparecen en las columnas. Por lo tanto, un archivo de datos se puede considerar como una matriz donde, si nos fijamos en una columna, tenemos los valores de una variable para todos los individuos y, si nos fijamos en una fila, tenemos todas las variables para un individuo.

2	9	58	1	5	4	60	1	1	3,00
2	4	57	1	3	3	42	1	2	3,00
1	19	84	2	3	2	25	0	0	8,00
1	116	61	1	1	2	22	0	0	4,00
1	117	66	1	1	2	22	0	0	4,00
1	23	67	1	3	1	15	1	6	5,00
1	71	68	1	4	2	24	0	0	5,00
1	107	62	1	1	2	30	0	0	4,00
2	113	55	1	3	2	30	0	0	2,00
2	7	70	1	2	2	30	1	9	5,00
1	,100	71	1	1	1	20	0	0	6,00
2	12	58	1	2	2	30	1	2	3,00
1	24	56	1	3	2	28	1	1	2,00
2	8	53	1	4	4	34	0	0	2,00
2	81	68	2	3	3	50	0	0	5,00
1	94	57	1	4	2	40	0	0	3,00
1	84	70	2	3	1	18	1	1	5,00
1	15	47	1	6	3	60	0	0	1,00
1	25	53	1	3	3	45	1	9	2,00
1	98	72	1	1	1	20	0	0	6,00
1	48	69	2	2	1	15	1	3	5,00

Vista de datos

Figura D1 .5. Pantalla de datos. Hoja **Vista de datos.**

Los datos pueden leerse de ficheros ya existentes mediante el procedimiento **Archivo,** de la barra de menú, menú **Abrir** y submenú **Abrir datos,** o bien pueden crearse directamente sobre la plantilla que aparece en pantalla. Los ficheros que crea el SPSS tienen extensión sav.

En la parte inferior de la Figura D1.5 hay dos «pestañas»: **Vista de datos,** que es la que acabamos de describir y **Vista de variables.** Si nos posicionamos en la pestaña **Vista de variables,** veremos las características de cada variable o columna para los datos del Ejemplo C2.1 (véase Fig. D1.6).

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación
1	recidiva	Numérico	1	0	recidiva local	{1, no}...	Ninguno	6	Derecha
2	supervivenci	Numérico	3	0	supervivencia g	Ninguno	Ninguno	9	Derecha
3	edad	Numérico	2	0	edad en años	Ninguno	Ninguno	6	Derecha
4	sexo	Numérico	1	0	sexo	{1, varón}...	Ninguno	4	Derecha
5	localización	Numérico	1	0	localización de	{1, labio}...	Ninguno	9	Derecha
6	categoría.tu	Numérico	1	0	categoría del t	{1, <20 mm}...	Ninguno	7	Derecha
7	diámetro	Numérico	2	0	diámetro en m	Ninguno	Ninguno	7	Derecha
8	infiltración. g	Numérico	1	0	infiltración gan	{0, no}...	Ninguno	7	Derecha
9	número. gan	Numérico	1	0	nº ganglios infil	Ninguno	Ninguno	6	Derecha
10	edad.agrup	Numérico	8	2	edad	{1,00, <50}...	Ninguno	10	Derecha

Figura D1.6. Pantalla de datos. Hoja Vista de variables.

Para ver o modificar las características de una variable, bastará con posicionarse en la parle derecha de la característica que queremos modificar. Por ejemplo, si queremos ver qué etiquetas tienen los valores de la variable categoría del tumor, tendríamos que posicionarnos en la parte derecha de **Valores** (Fig. D1.6) de la fila 6, y emergería la ventana que aparece en la Figura D1.7, donde se podrían cambiar o añadir etiquetas.

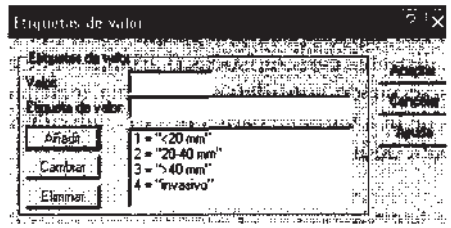


Figura D1.7. Valores de las etiquetas.

Una vez que se tienen en pantalla los datos con los que vamos a trabajar, ya se pueden ejecutar los procedimientos estadísticos que hay en **Analizar** o **Gráficos** de la barra de menú. Para ello, nos posicionarnos en el procedimiento a realizar y, haciendo clic con el botón izquierdo del ratón, nos aparecerá una ventana como la de la Figura D1.8, donde se encuentra la lista de variables de nuestra pantalla de datos. Si nos posicionarnos sobre el nombre de la variable objeto de nuestro estudio y pulsamos ■, el nombre de la variable elegida desaparece de la lista y pasa a *Variables.* Una seleccionada la variable a utilizar, si pulsamos **Aceptar** se realizarán las opciones que, por defecto tenga el procedimiento elegido.

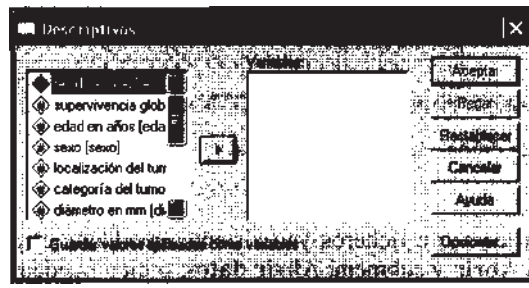
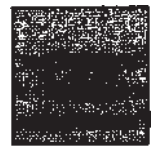


Figura D1.8. Elección de variable en estudio.

Los resultados de los procedimientos se guardan en una pantalla de resultados. Todos los resu tados que obtengamos en una sesión se pueden guardar en un fichero de extensión spo.



Estadística descriptiva

Utilizando los datos proporcionados por el Ejemplo C2.1 del Apéndice C2 vamos a mostrar el manejo del SPSS en lo que a técnicas de estadística descriptiva se refiere.

La forma más útil de introducir los datos del Ejemplo C2.1 para su posterior análisis con el paquete SPSS se muestra en el fichero de datos que aparece en la Figura D1.5 del Apéndice D1, creado en un documento de datos SPSS. En dicha figura se muestra la hoja **Vista de datos** de la pantalla de datos del SPSS y en ella se recogen los valores relativos a los primeros 21 pacientes. Las variables de naturaleza cuantitativa, discreta o continua, deben grabarse con sus valores respectivos. Las variables de naturaleza discreta cualitativa pueden grabarse con sus modalidades o etiquetas no numéricas, por ejemplo la **recidiva** como sí o no. En caso de codificar numéricamente dichas modalidades, por ejemplo la **recidiva** como 1 y 2, deberemos asignar etiquetas a cada número, si queremos que se muestren los nombres de las modalidades en los resultados que obtengamos. Esto puede hacerse en la otra hoja de la pantalla de datos, **Vista de variables,** en la que además se puede definir la escala de medida de cada variable. La Figura D2.1 muestra, respectivamente, esta última hoja y la ventana que contiene las etiquetas de la variable **localización del tumor.** Observando la última columna de la hoja de datos, Medida, se aprecia la diferente escala de medida de las variables.

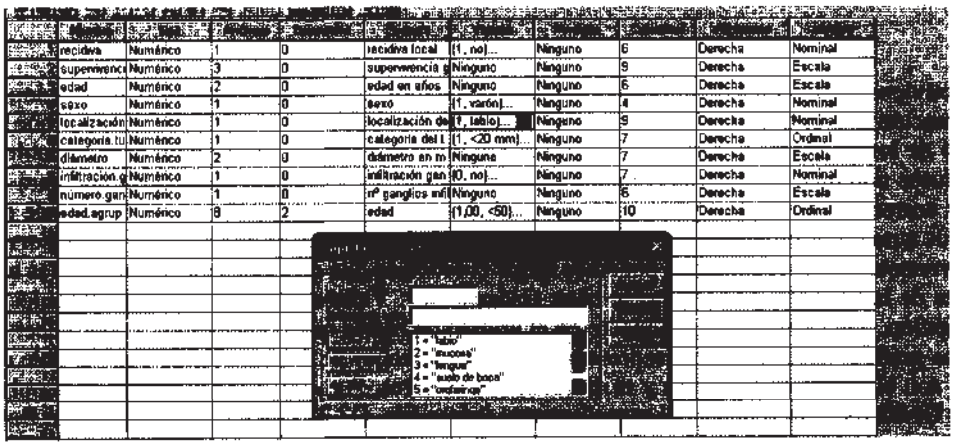


Figura D2.1. Fichero de datos correspondiente al Ejemplo C2.1, vista de variables, y etiquetas de la variable localización del tumor.

D2.1. DESCRIPCIÓN DE UNA VARIABLE CUALITATIVA

Para obtener la tabla de frecuencias de la variable cualitativa categoría del tumor, elegimos el procedimiento **Analizar** de la barra de menú, a continuación **Estadísticos descriptivos,** y finalmente **Frecuencias.** Una vez desplegada la ventana del procedimiento seleccionar, sobre el listado de variables, **categoría del tumor** y arrastrar dicha variable al cuadro *Variables,* seleccionar *Mostrar tablas de frecuencias.* La Figura D2.2 muestra la ventana de este procedimiento. Pulsando **Aceptar,** se obtienen los resultados que muestra la Tabla D2.1.

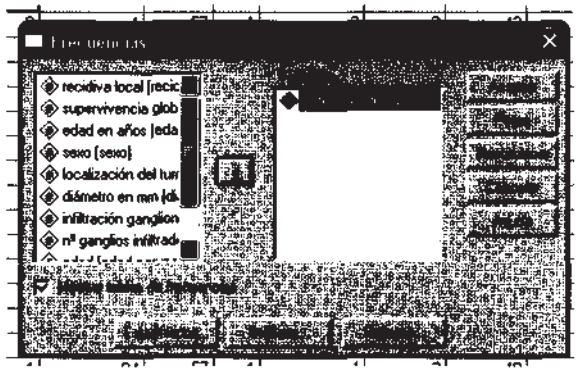


Figura D2.2. Ventana del procedimiento Analizar, Estadísticos descriptivos. Frecuencias.

Tabla D2.1. Distribución de frecuencias correspondiente **a** la categoría del tumor

		Categoría del tumor primario			
		Frecuencia	**Porcentaje**	**Porcentaje válido**	Porcentaje acumulado
Válidos	<20mm	31	35.2	35.2	35.2
	20-40 mm	37	42.0	42.0	77.3
	>40mm	7	8.0	8.0	85.2
	Invasivo	13	14.8	14.8	100.0
	Total	88	100.0	100.0	

Conviene recordar que la última columna de resultados, porcentaje acumulado, proporciona una información útil al tratarse de una variable cualitativa ordinal, por ejemplo, se observa que en el 77.3% de los pacientes la categoría del tumor era inferior o igual a los 40 mm. Si la variable descrita fuese nominal, esta última columna carecería de interés. Debemos puntualizar que para interpretar correctamente dichos porcentajes acumulados hemos asignado, al crear el fichero de datos, códigos numéricos ordenados según el propio orden de la variable. Si pulsamos el botón **Formato**, en la ventana del procedimiento, podemos elegir el orden en que deseamos que aparezcan las modalidades de la variable, por defecto se tiene *Ordenar por.* **Valores ascendentes,** mantenemos pues esta opción, que se aplicará igualmente en el gráfico de barras que comentamos a continuación.

Las gráficas adecuadas para representar las distribuciones de frecuencias de una variable cualitativa son el diagrama de barras y el diagrama de sectores. Para seleccionar una de ellas pulsar el botón **Gráficos** de la ventana del procedimiento (véase Fig. D2.2) y elegir *Tipo de gráfico:* **Barras, Sectores,** y *Valores del gráfico:* **Frecuencias, Porcentajes.** Los dos gráficos de porcentajes se muestran en las Figuras D2.3 y D2.4.

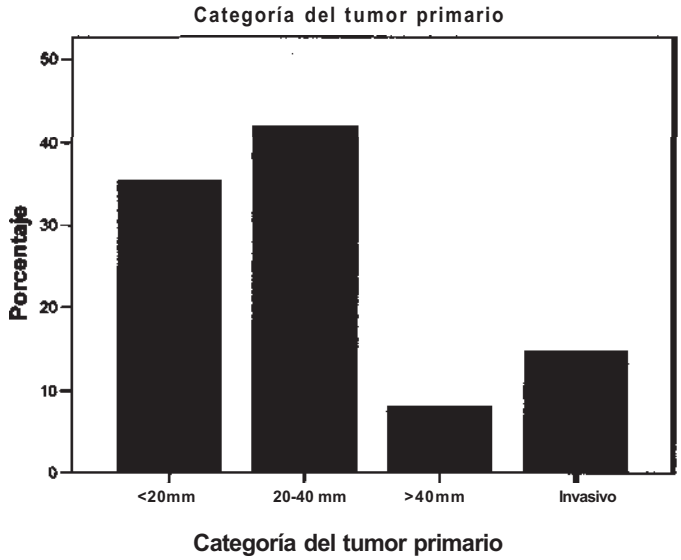


Figura D2.3. Gráfico de barras correspondiente a la categoría del tumor primario.

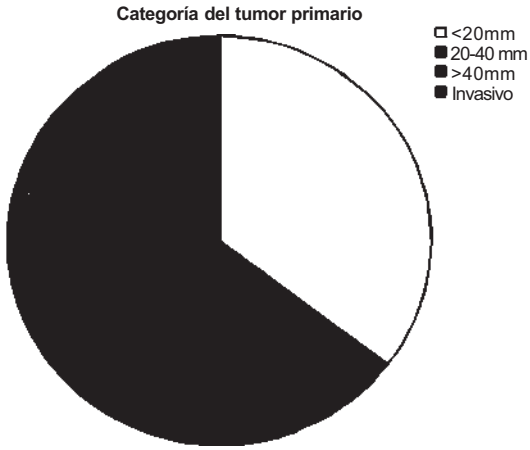


Figura D2.4. Gráfico de sectores correspondiente a la categoría del tumor primario.

Los títulos de ambos gráficos se han editado con el editor de gráficos del SPSS. Haciendo doble clic sobre el gráfico, aparecerá dicho editor; basta un solo clic sobre el gráfico con el botón derecho del ratón para desplegar una ventana, elegir la opción *Objeta gráfico de SPSS, Abrir.*

En el procedimiento **Gráficos** de la barra de menú también pueden seleccionarse los gráficos deseados.

D2.2. DESCRIPCIÓN DE UNA VARIABLE CUANTITATIVA DISCRETA

A diferencia del programa STATGRAPHICS Plus, con el SPSS puede obtenerse la descripción completa de la variable, frecuencias, gráficas y estadísticos, con un solo procedimiento. Vamos a

describir la variable cuantitativa discreta número de ganglios infiltrados. Para ello, elegimos el procedimiento **Analizar** de la barra de menú, a continuación **Estadísticos descriptivos,** y finalmente **Frecuencias.** Una vez desplegada la ventana del procedimiento, igual a la de la Figura D2.2, es necesario seleccionar **número de ganglios infiltrados,** seleccionar *Mostrar tablas de frecuencias,* pulsar el botón **Estadísticos** para seleccionar las medidas deseadas de tendencia central, de dispersión y valores percentiles, pulsar **Continuar,** pulsar el botón **Gráficos** y elegir *Tipo de gráfico:* **Histograma,** pulsar **Continuar,** de nuevo en la ventana del procedimiento pulsar **Aceptar.** Los resultados, numéricos y gráficos, se muestran en las Tablas D2.2 y D2.3, y en la Figura D2.5.

Tabla D2.2. Estadísticos correspondientes al número de ganglios infiltrados

Estadísticos		
Número de ganglios infiltrados		
N	Válidos	88
	Perdidos	0
Media		.94
Mediana		.00
Moda		0
Desviación típica		1.890
Varianza		3.571
Mínimo		0
Máximo		9
Percentiles	25	.00
	50	.00
	75	1.00

Algunos de los resultados mostrados en la Tabla D2.2 merecen un comentario más detallado. El hecho de que el valor 0 aparezca en 60 de los 88 valores registrados para esta variable explica el desplazamiento de todas las medidas centrales hacia dicho valor, esto cuestiona el apelativo de centrales de estas medidas, en rigor la única medida que obedece un criterio de posición central es la mediana, 0 ganglios infiltrados es el valor que aparece en el «medio» del conjunto de datos ordenados. La media aritmética, 0.94, se desplaza por el efecto de datos atípicos, hay 2 pacientes con 9 ganglios infiltrados. La moda, el valor más frecuente de la variable, no obedece a un criterio de posición o promedio como las dos anteriores, en su obtención sólo interviene la frecuencia absoluta de los diferentes valores de la variable por lo que, como ocurre en este caso, puede ser un valor extremo de dicha variable.

Al pulsar el botón **Estadísticos** y seleccionar *Valores percentiles,* podemos seleccionar cuartiles, o cualesquiera percentiles. La Tabla D2.2 muestra los 3 cuartiles de esta distribución que aparecen como percentiles 25, 50 y 75; el 2.° cuartil, o percentil 50, es la mediana de la distribución.

Al pulsar el botón **Estadísticos** (véase Fig. D2.2) se pueden seleccionar dos medidas de forma, Asimetría y Curtosis, que tienen utilidad cuando se manejan variables continuas. Al pulsar el botón **Gráficos** de la ventana de procedimiento se puede elegir **Gráficos de barras** que, al igual que **Gráficos de sectores,** es una opción más adecuada para variables cualitativas, ya que no permite escalar los valores de la variable. Para obtener la gráfica de la Figura D2.5 es necesario cambiar en el editor de gráficos del SPSS alguna de las opciones que, por defecto, usa dicho editor; una vez abierta la opción *Objeto gráfico de SPSS,* debe seguirse la secuencia *Edición, Seleccionar eje X, Escala, Rango:* Mínimo personalizado -1, Incremento mayor personalizado 1. Al pulsar **Aplicar, y Cerrar** el editor, el origen del eje X se desplaza a -1, y los valores de la variable aparecen de uno en uno.

Tabla D2.3. Distribución de frecuencias correspondiente al número de ganglios infiltrados

Número de ganglios infiltrados					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	0	60	68.2	68.2	68.2
	1	10	11.4	11.4	79.5
	2	5	5.7	5.7	85.2
	3	4	4.5	4.5	89.8
	4	4	4.5	4.5	94.3
	5	1	1.1	1.1	95.5
	6	2	2.3	2.3	97.7
	9	2	2.3	2.3	100.0
	Total	88	100.0	100.0	

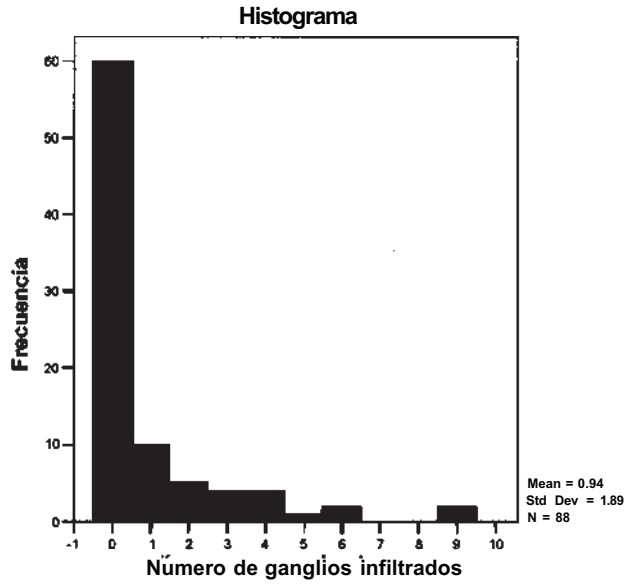


Figura D2.5. Histograma de frecuencias correspondiente al número de ganglios infiltrados.

D2.3. DESCRIPCIÓN DE UNA VARIABLE CUANTITATIVA CONTINUA

Para la descripción de la variable continua edad, elegimos el procedimiento **Analizar, Estadísticos descriptivos, Frecuencias.** Una vez desplegada la ventana del procedimiento seleccionar **edad agrupada,** hay que seleccionar *Mostrar tablas de frecuencias,* pulsar el botón **Estadísticos** para comprobar que no hay seleccionado ninguno, y hacer lo mismo con el botón **Gráficos.** Lo que pretendemos es obtener únicamente la tabla de frecuencias con los datos agrupados en intervalos de edad; por eso hemos seleccionado la variable **edad agrupada y** no **edad** del fichero de datos. La explicación es que el SPSS no permite, como opción numérica, el agrupamiento de los valores en intervalos. La variable **edad agrupada** se codificó con valores ordenados de 1 a 8, que corresponden a otros tantos lustros de edad (véase Fig. D2.6). Pulsando **Aceptar** se obtienen las distribuciones de frecuencias que muestra la Tabla D2.4.

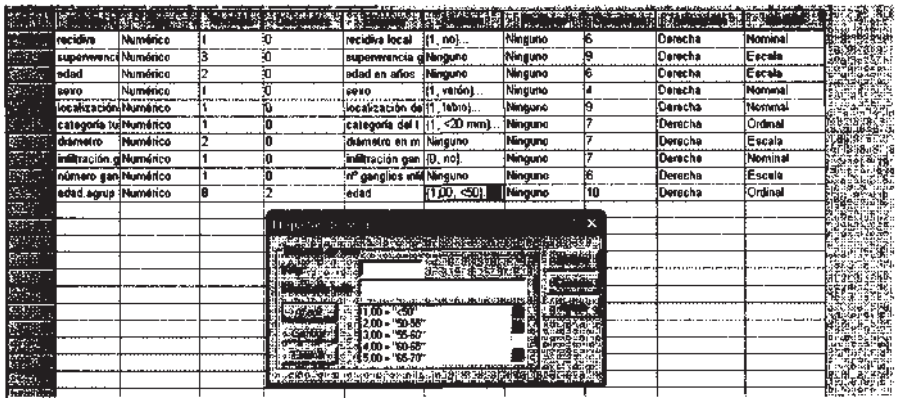


Figura D2.6. Etiquetas de la variable edad agrupada.

Tabla D2.4. Distribución de frecuencias correspondiente a la edad agrupada en intervalos de 5 años

Edad					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	<50	**14**	15.9	15.9	15.9
	50-55	17	19.3	19.3	35.2
	55-60	**16**	18.2	18.2	53.4
	60-65	17	19.3	19.3	72.7
	65-70	**16**	18.2	18.2	90.9
	70-75	**4**	4.5	4.5	95.5
	75-80	**2**	2.3	2.3	97.7
	>80	**2**	2.3	2.3	100.0
	Total	**88**	100.0	100.0	

Para continuar el análisis descriptivo de esta variable debemos volver al procedimiento anterior y seleccionar ahora la variable **edad,** pulsamos el botón **Estadísticos** y seleccionamos las medidas de centralización, dispersión y forma, deseadas, pulsamos **Continuar,** pulsamos el botón **Gráfico;** y elegimos *Tipo de gráfico:* **Histograma, con curva normal,** pulsamos **Continuar,** de nuevo en k. ventana del procedimiento pulsamos **Aceptar.** Los estadísticos obtenidos se muestran en la. Tabla D2.5, y el histograma en la Figura D2.7.

Los coeficientes de asimetría y curtosis son dos medidas de forma de la distribución. Una vez estandarizados, dividiendo cada uno por su error estándar, proporcionan dos estadísticos para el contraste de la normalidad de la variable. Si el valor de uno de estos coeficientes estandarizados queda fuera del intervalo (-1.96, 1.96), debería rechazarse la normalidad de la variable. En rigor estos test requieren muestras muy amplias, por lo que el contraste de la normalidad se iniciaría gráficamente, visualizando el histograma de frecuencias, y finalizaría con la utilización de un contraste de normalidad adecuado (véase apartado D3.4 del Apéndice D3).

Para obtener el histograma, tal y como aparece en la Figura D2.7, es necesario abrir el Editor de gráficos y seguir la secuencia *Edición, Seleccionar eje X.* Al activarse una ventana de **Propiedades** habrá que abrir *Opciones de histograma;* en *Tamaño de los intervalos,* elegir *Personalizado;* y cuando se activa el campo *Número de intervalos,* marcamos 8. Con esta especificación, elegimos únicamente el número de intervalos, todos de igual amplitud (en este caso 7.5 años), resultado de dividir la amplitud (90-30) entre 8. En general, el agrupamiento en clases de una variable continua

Tabla D2.5. Medidas de centralización, dispersión y forma correspondientes a la edad

Estadísticos		
Edad en años		
N	Válidos	88
	Perdidos	0
Media		59.25
Mediana		59.00
Moda		57
Desviación típica		9.674
Varianza		93.592
Asimetría		-0.035
Error típ- de asimetría		0.257
Curtosis		0.088
Error típ. de curtosis		0.508
Mínimo		33
Máximo		84
Percentiles	25	53.00
	50	59.00
	75	66.00

persigue el poder visualizar la forma, simétrica, asimétrica a la derecha, apuntada, etc., de su gráfica de frecuencias, por lo que es habitual llevar a cabo diferentes agrupamientos en el mismo análisis. Los valores 30 y 90 años aparecen, por defecto, al abrir la opción *Escala* de la ventana de **Propiedades** que acabamos de comentar. Aquí pueden modificarse estos extremos y ajustados al mínimo y máximo de los valores observados, 33 y 84 años, respectivamente. Para finalizar, la curva normal superpuesta sobre el histograma permite comprobar si los datos se ajustan «gráficamente» a la distribución normal (véase apartado D3.3 del Apéndice D3).

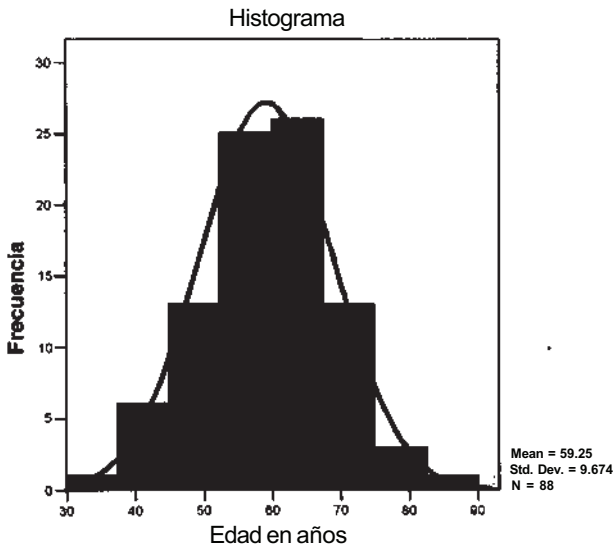


Figura D2.7. Histograma de frecuencias correspondiente a la edad.

Se describen dos gráficos adicionales, válidos para variables cuantitativas, el Diagrama de cajas y bigotes y el Diagrama de tallo y hojas (véanse apartados 1.2 y 1.6. del Capítulo 1). El primero de ellos, que se muestra en la Figura D2.8 en dos grupos separados para los pacientes con y sin recidiva local, permite mostrar medidas centrales y de localización así como dar una idea de la dispersión y forma de la distribución de frecuencias. La caja central tiene por extremos q_1 y q_3, el primer y tercer cuartiles; la línea horizontal dentro de la caja se corresponde con el valor de la mediana; las líneas horizontales, a ambos lados de la caja, llamadas bigotes, unen los valores adyacentes. Finalmente, en el exterior del bigote del diagrama relativo a los pacientes con recidiva, se marca el dato atípico 54 años. La Figura D2.9 muestra el Diagrama de tallo y hojas. Para obtener ambos gráficos, elegimos el procedimiento **Analizar, Estadísticos descriptivos, Explorar...** Una vez des-

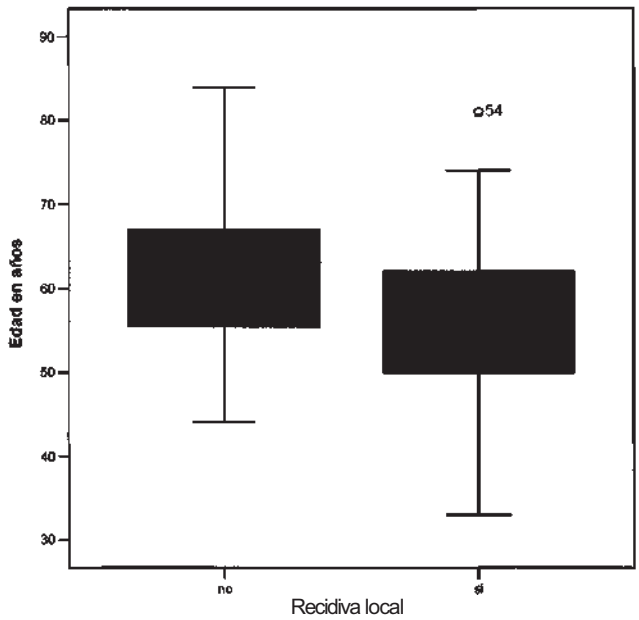


Figura D2.8. Diagrama de caja y bigotes correspondiente a la edad para los grupos con y sin recidiva local.

edad en años Diagrama de tallo y hojas

Frecuencia	Tallos y hojas
1.00	Extremos (=<33)
6.00	4 . 014444
5.00	4 . 55577
14.00	5 . 00111222333334
19.00	5 . 5555566777777888999
15.00	6 . 001111122334444
17.00	6 . 55556666777888889
7.00	7 . 0001234
1.00	7 . 7
3.00	8 . 014

Ancho de tallo: 10
Cada hoja: 1 caso(s)

Figura D2.9. Diagrama de tallo y hojas correspondiente a la edad para los grupos con y sin recidiva local.

plegada la ventana del procedimiento es necesario seleccionar la variable **edad** y arrastrarla al campo *Dependientes.* Si deseamos gráficos separados, según las modalidades de otra variable discreta (en nuestro caso recidiva), tenemos que arrastrar dicha variable al campo *Factores,* seleccionar *Mostrar gráficos,* pulsar el botón Gráficos y seleccionar en *Diagramas de caja,* **Niveles de los factores juntos,** y en *Descriptivos,* **Tallo** y **hojas,** pulsar **Continuar,** y de nuevo en la ventana del procedimiento, pulsar **Aceptar.** Con este procedimiento se pueden obtener gráficos simultáneos, incluidos histogramas, para diferentes variables.

D2.4. DESCRIPCIÓN CONJUNTA DE DOS VARIABLES

La descripción que se presenta a continuación se refiere a dos variables de cualquier naturaleza. Para obtener la tabla de frecuencias conjuntas de la categoría del tumor según la infiltración ganglionar, elegimos el procedimiento **Analizar, Estadísticos descriptivos, Tablas de contingencia.** Una vez emerge la ventana del procedimiento, hay que seleccionar **categoría del tumor** y arrastrarla al campo *Filas,* seleccionar **infiltración ganglionar** y arrastrarla al campo *Columnas,* pulsar el botón Casillas y seleccionar en *Frecuencias*[1], **Observadas,** y en *Porcentajes,* **Fila y Total,** pulsar **Continuar,** y de nuevo en la ventana del procedimiento, pulsar **Aceptar.** La Tabla D2.6 muestra los resultados obtenidos. Se proporcionan, para cada casilla, tres anotaciones, el número de pacientes y dos porcentajes: el primero es relativo puesto que se refiere al reparto de los pacientes de cada fila en las dos casillas, y el segundo es absoluto, ya que cuantifica el peso de cada casilla en relación al total de pacientes. En las últimas fila y columna se muestran los totales con y sin infiltración ganglionar y los totales de cada categoría del tumor, respectivamente.

La observación y comparación de los porcentajes relativos de las casillas, a lo largo de las filas, resulta de especial interés. Así, puede observarse cómo en los tumores < 20 mm y 20-40 mm es menos frecuente la aparición de infiltración ganglionar que en los tumores > 40 mm e invasivos. Como se ha comentado anteriormente, sólo un test estadístico conducirá a establecer, en este caso, la dependencia o no de los dos factores descritos en la tabla.

El gráfico de barras que muestra la Figura D2.10 permite visualizar lo que acabamos de comentar. Para obtener dicho gráfico basta seleccionar *Mostrar los gráficos de barras agrupadas* en la

Tabla D2.6. Tabla de contingencia correspondiente a la categoría del tumor y a la infiltración ganglionar

			Infiltración ganglionar		
			No	**Sí**	**Total**
Categoría del tumor primario	**< 20 mm**	Recuento	22	**9**	31
		% de categoría del tumor primario	71.0%	29.0%	100.0%
		% del total	25.0%	10.2%	35.2%
	20-40 mm	Recuento	28	**9**	37
		% de categoría del tumor primario	75.7%	24.3%	100.0%
		% del total	31.8%	10.2%	42.0%
	>40mm	Recuento	**3**	**4**	7
		% de categoría del tumor primario	42.9%	57.1%	100.0%
		% del total	**3.4%**	**4.5%**	8.0%
	Invasivo	Recuento	**7**	**6**	13
		% de categoría del tumor primario	53.8%	46.2%	100.0%
		% del total	**8.0%**	**6.8%**	14.8%
Total		Recuento	**60**	28	88
		% de categoría del tumor primario	68.2%	31.8%	100.0%
		% del total	68.2%	31.8%	100.0%

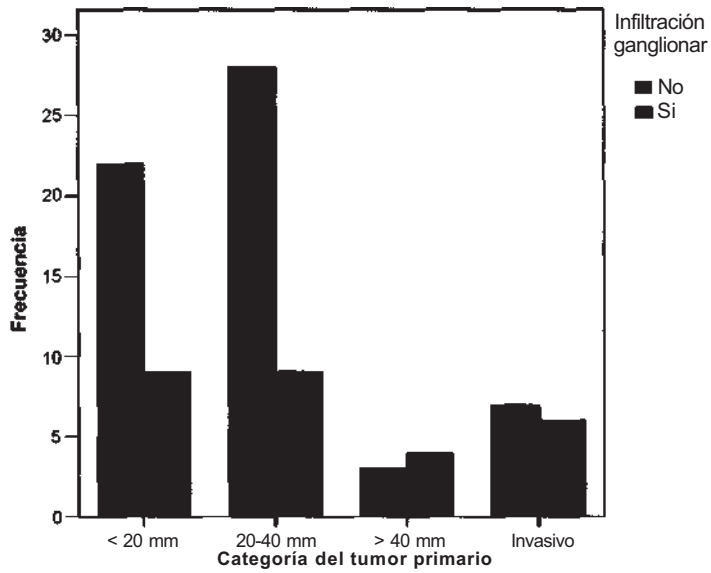


Figura D2.10. Gráfico de barras correspondiente a la categoría del tumor según la infiltración ganglionar.

ventana del procedimiento anterior. El procedimiento **Gráficos** de la barra de menú incluye en sus opciones este y otros gráficos de barras, permitiendo diferentes diseños, apilado, agrupado, así como diferentes medidas para las barras, frecuencias, porcentajes.

Como se comentó al principio de este apartado, el procedimiento mostrado sirve para la descripción conjunta de dos variables de cualquier naturaleza. Si una o las dos variables son cuantitativas continuas, es preciso categorizar dicha o dichas variables para que resulte útil la tabla de frecuencias. Por ejemplo, si queremos describir la localización del tumor según la edad, deberíamos elegir la variable edad agrupada, descrita en el apartado D2.3 de este Apéndice.

Distribuciones de probabilidad con SPSS

El SPSS no tiene un procedimiento específico para estudiar distintas distribuciones de probabilidad. Sin embargo, en ciertos procedimientos se pueden ver distintos aspectos de las distribuciones.

D3.1. FUNCIÓN DE DISTRIBUCIÓN

El SPSS permite evaluar el valor de la función de distribución para algunas distribuciones contenidas en dicho programa. En concreto, veremos las estudiadas en los Capítulos 4 y 5.

Binomial B(n, p)

Supongamos que queremos calcular la función de distribución de una variable X con distribución B(5, 0.3). El primer paso será crear, en la pantalla de datos (hoja Vista de datos), una variable que contenga todos los posibles valores de la variable, a la que llamaremos, por ejemplo, **abscisa.** Para calcular ahora los valores de la función de distribución en dichos puntos, seleccionamos el procedimiento **Transformar, Calcular,** y aparece una ventana como la de la Figura D3.1.

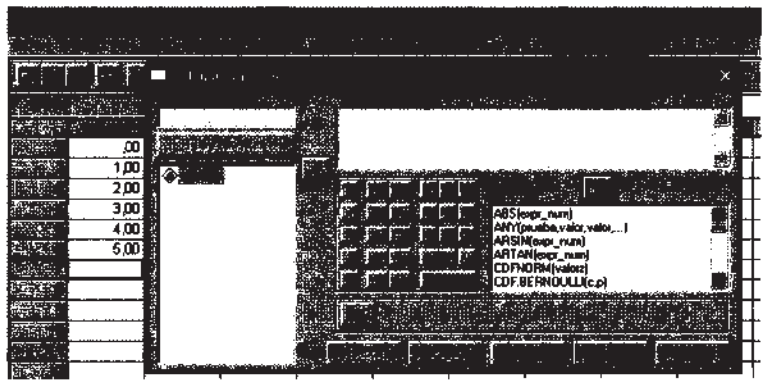


Figura D3.1. Datos y funciones disponibles.

En el campo *Variable de destino* (Fig. D3.1), debemos indicar el nombre que recibirá la variable donde se calcule la función de distribución, por ejemplo **distribución.** Para completar el campo *Expresión numérica* (Fig. D3.1), debemos seleccionar de entre las *Funciones* que tiene internamente el SPSS la que es **CDF.BINOM(c, n, p)** y pulsando el botón de flecha , aparecerá dicha función en el campo *Expresión numérica.*

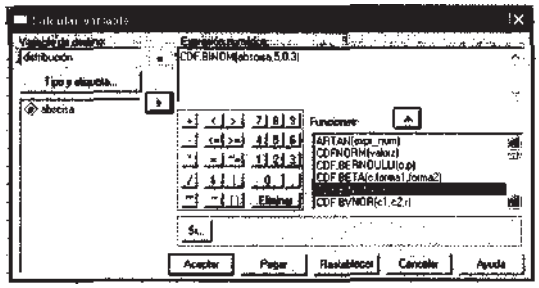
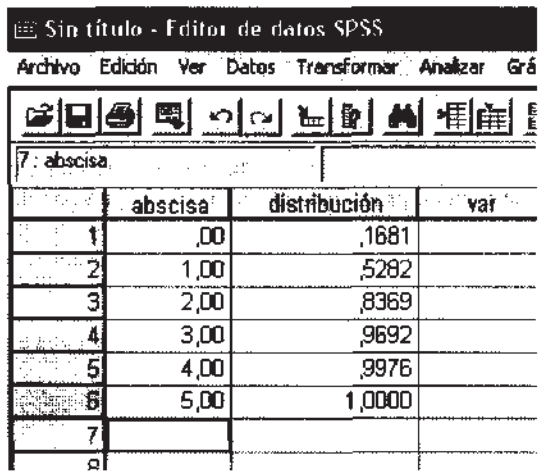


Figura D3.2. Elección de la función de distribución para la variable binomial.

Los tres parámetros de esta distribución han de rellenarse: *c* será nuestra variable **abscisa,** *n* será 5 y *p* valdrá 0.3 (véase Fig. D3.2). Si ahora pulsamos **Aceptar,** tendremos en la pantalla de dato (Hoja Vista de datos) los valores de la función de distribución (véase Fig. D3.3). Puede verse la coincidencia de estos valores con sus correspondientes de la Tabla I del Apéndice B. La ventaja de utilizar el paquete estadístico es que permite ampliar dicha Tabla, para valores tanto de *n* como de *p.* Debemos hacer notar que, por defecto, el SPSS da sólo 2 decimales en pantalla; si queremos ver 4, como en este caso, debemos marcar la pestaña inferior izquierda **Vista de variables** y en los **Decimales** de la variable **distribución** cambiar el 2, que viene por defecto, por un 4.



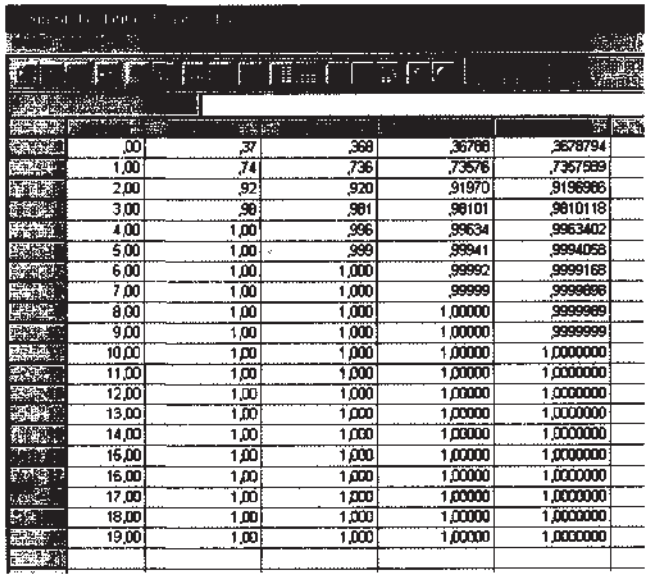
	abscisa	distribución	var
1	,00	,1681	
2	1,00	,5282	
3	2,00	,8369	
4	3,00	,9692	
5	4,00	,9976	
6	5,00	1,0000	
7			

Figura D3.3. Valores de la función de distribución para la variable B(5, 0.3).

Poisson $P(\lambda)$

Supongamos que queremos calcular la función de distribución de una variable X con distribución de Poisson de media 1, *P(l).* Al igual que en el caso anterior, el primer paso sería crear, en la pantalla de datos (Hoja Vista de datos), una columna que contenga todos los posibles valores de la

variable X; pero como el rango de valores de la distribución de Poisson va de 0 a ∞, pondremos, por ejemplo, sólo los primeros 20 valores. Llamaremos a esta variable **abscisa.** Para calcular ahora los valores de la función de distribución en dichos puntos, seleccionamos el procedimiento **Transformar, Calcular,** y aparece un cuadro en pantalla como el de la Figura D3.1. En el campo *Variable de destino,* debemos indicar el nombre que recibirá la variable donde se calcule la función de distribución, **distribución.** Para completar el campo *Expresión numérica,* debemos seleccionar de entre las *Funciones* la que tiene por nombre CDF.POISSON(c, **media),** y pulsando el botón de flecha , aparecerá dicha función en el campo *Expresión numérica.* Debemos rellenar sus campos: c será **abscisa** y *media* valdrá 1. Al pulsar ahora **Aceptar,** aparecen los valores en la columna **distribución** (véase Fig. D3.4).



	,00	,37	,368	,36788	,3678794
	1,00	,74	,736	,73576	,7357589
	2,00	,92	,920	,91970	,9196986
	3,00	,98	,981	,98101	,9810118
	4,00	1,00	,996	,99634	,9963402
	5,00	1,00	,999	,99941	,9994058
	6,00	1,00	1,000	,99992	,9999168
	7,00	1,00	1,000	,99999	,9999898
	8,00	1,00	1,000	1,00000	,9999989
	9,00	1,00	1,000	1,00000	,9999999
	10,00	1,00	1,000	1,00000	1,0000000
	11,00	1,00	1,000	1,00000	1,0000000
	12,00	1,00	1,000	1,00000	1,0000000
	13,00	1,00	1,000	1,00000	1,0000000
	14,00	1,00	1,000	1,00000	1,0000000
	15,00	1,00	1,000	1,00000	1,0000000
	16,00	1,00	1,000	1,00000	1,0000000
	17,00	1,00	1,000	1,00000	1,0000000
	18,00	1,00	1,000	1,00000	1,0000000
	19,00	1,00	1,000	1,00000	1,0000000

Figura D3.4. Valores de la función de distribución para la variable P(1).

En la Figura D3.4 se encuentran los valores de la función de distribución para una distribución de Poisson, P(l), considerando distinto número de decimales. En la columna **distribución** se han considerado 2 decimales. En **distribución3,** se consideran 3 decimales, coincidiendo estos valores con los dados en la Tabla II del Apéndice B. Análogamente, **distribución5** tiene 5 decimales y **distribución7** tiene 7 decimales. Observando las 4 últimas columnas se ve cómo el programa redondea las probabilidades. Por lo tanto, a partir de cierto valor de X (diferente en cada columna), la función de distribución ya vale 1 (volvemos a recordar que teóricamente el rango de valores de la Poisson va de 0 a ∞).

Normal N(μ, σ)

Supongamos que queremos calcular para una variable Z con distribución N(0,1), las siguientes probabilidades, P(-3 < Z< 3), P(-2 < Z < 2) y P(-1 < Z < 1).

El primer paso será crear, en la pantalla de datos, una variable que contenga los valores de la variable (en este caso, los extremos de los intervalos solicitados); le daremos el nombre de **abscisa** como hicimos en la Figura D3.1. Para calcular ahora los valores de la función de distribución en dichos puntos, seleccionamos el procedimiento **Transformar, Calcular,** y aparece un cuadro en pantalla como el de la Figura D3.1. En el campo *Variable de destino* indicamos el nombre de la

variable donde se evalúe la función de distribución, **distribución.** Para completar el campo *Expresión numérica,* debemos seleccionar de entre las *Funciones* la que es **CDFNORM(valorz)** *y* pulsando el botón de flecha [▶], aparecerá esta función en dicho campo. Debemos rellenar e campo *valorz* con **abscisa** y pulsar **Aceptar.** Los resultados los tenemos en la Figura D3.5, que coinciden con los de la Tabla III para la distribución normal, N(0, 1), del Apéndice B.

Ahora, para calcular las probabilidades que queríamos tendremos:

$$P(-3 < Z < 3) = P(Z < 3) - P(Z < -3) = 0.9987 - 0.0013 = 0.9974$$
$$P(-2 < Z < 2) = P(Z < 2) - P(< -2) = 0.9772 - 0.0228 = 0.9544$$
$$P(-1 < Z < 1) = P(Z < 1) - P(Z < -1) = 0.8413 - 0.1587 = 0.6826$$

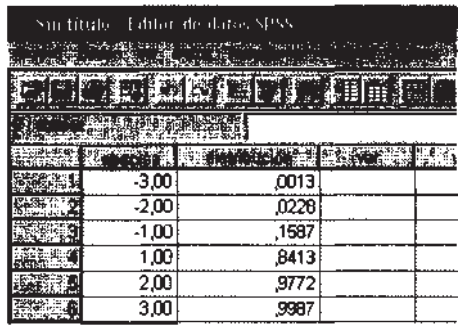


Figura D3.5. Valores de la función de distribución para la variable N(0, 1).

Si quisiéramos calcular probabilidades para una distribución normal de otros parámetros, no e necesario tipificar; basta utilizar la función **CDF.NORMAL(c, media, desv_típ).** Supongamos que: X es N(5, 3) y queremos calcular P(1 < X < 2); sólo tendríamos que rellenar c con **abscisa,** *media* con 5 y *desv-típ* con 3 (véase Fig. D3.6). Podemos ahora calcular:

$$P(1 < X < 2) = P(X < 2) - P(X < 1) = 0.1587 - 0.0912 = 0.0675$$

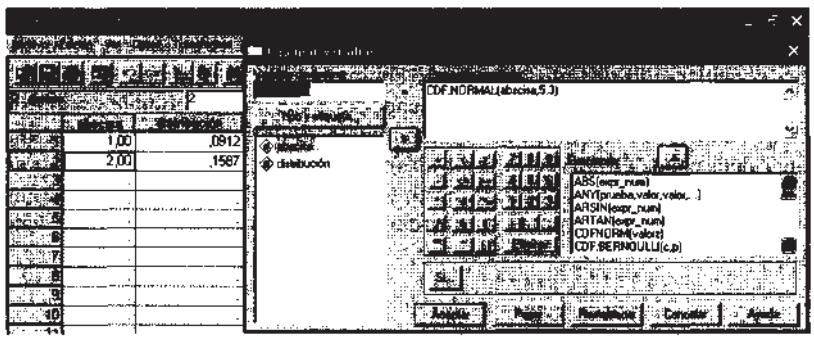


Figura D3.6. Valores de la función de distribución para la variable N(5, 3).

D3.2. INVERSO DE LA FUNCIÓN DE DISTRIBUCIÓN

Para la distribución normal entre otras (no para binomial ni Poisson), se puede realizar el procese inverso al anterior, es decir, para una probabilidad p, encontrar la abscisa c que deja a su izquierda

dicha probabilidad. En este caso, crearemos en la pantalla de datos una variable que contiene las probabilidades y que llamaremos, por ejemplo, probabilidad. Para calcular dichas abscisas, seleccionamos el procedimiento **Transformar, Calcular,** y aparece un cuadro en pantalla como el de la Figura D3.1. En el campo *Variable de destino,* indicamos el nombre de la variable donde se guardarán las abscisas, **abscisa.** Para completar el campo *Expresión numérica,* debemos seleccionar de entre las *Funciones* la que es **IDF.NORMAL(p, media, desv_típ)** y pulsando el botón de flecha ▀, aparecerá esta función en dicho campo. Si elegimos la variable N(7, 2) debemos rellenar el campo *p* con **probabilidad,** el campo *media* con 7 y el campo *des_típ* con 2 (véase Fig. D3.7).

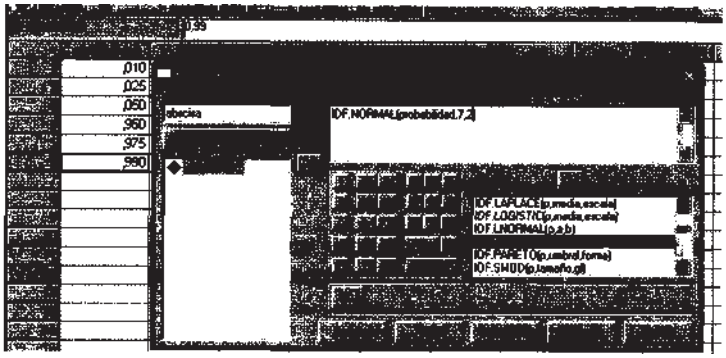


Figura D3.7. Proceso inverso de la función de distribución normal, N(7, 2).

Algunos resultados para una variable X con distribución N(7, 2) se dan en la Figura D3.8; por ejemplo, podemos ver que:

$$P(X < 2.35) = 0.01 \qquad P(X < 3.08) = 0.025$$

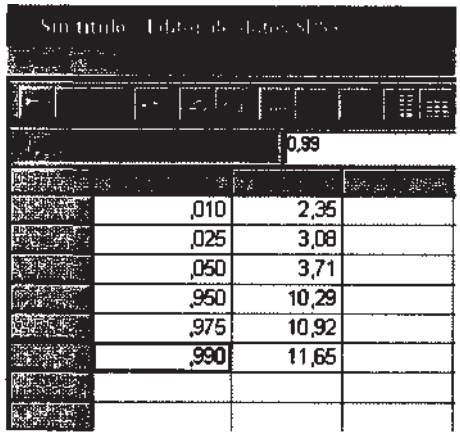


Figura D3.8. Resultados para la variable N(7, 2).

D3.3. AJUSTE GRÁFICO

El SPSS permite ver si unos datos se «ajustan» gráficamente sólo a una distribución normal de la misma media y desviación típica que los datos. Consideremos los datos del Ejemplo 1.1.1 del

Capítulo 1, y veamos si la edad se ajusta gráficamente a una normal. Elegimos el procedimiento **Gráficos, Histograma,** y una vez indicada la variable para la que se quiera obtener el histograma, marcamos *Mostrar curva normal* (véase Fig. D3.9).

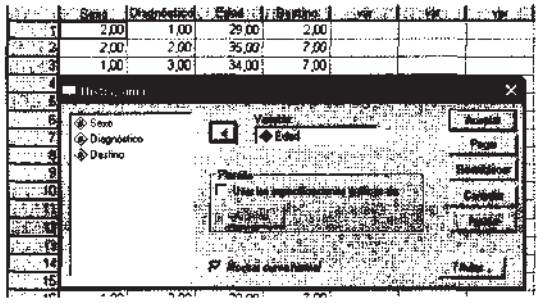


Figura D3.9. Instrucciones para histograma y curva normal.

En la Figura D3.10 se pueden ver los resultados gráficos: un histograma de los datos (ya comentado en el Apéndice D2) y la gráfica de la normal de la misma media y desviación típica.

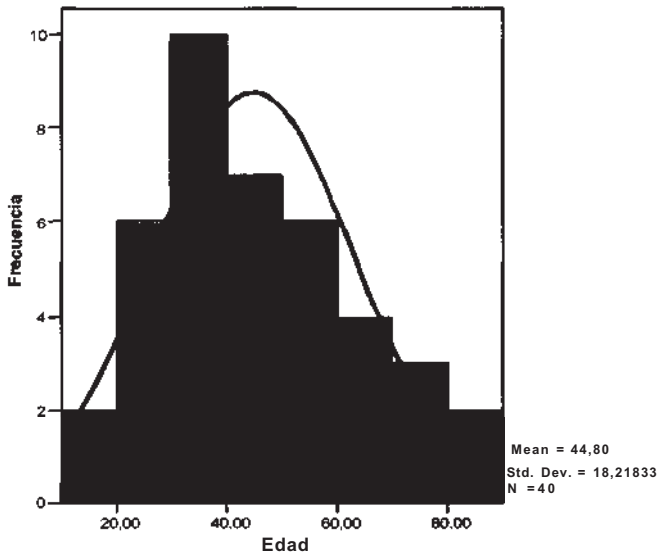


Figura D3.10. Histograma y curva normal.

D3.4. TEST DE BONDAD DE AJUSTE

En el Capítulo 13, en el apartado 13.1 se comenta una prueba gráfica para ver si los datos son normales, que denomina prueba de Lilliefors. No se ofrece ningún test estadístico para comprobar si los datos siguen una distribución dada (test de bondad de ajuste), pero sí se comenta cómo lo: paquetes estadísticos realizan al menos uno de ellos. El SPSS realiza el test de bondad de ajuste de Kolmogorov-Smirnov para las distribuciones normal, de Poisson, exponencial y uniforme. La prueba de Lilliefors, comentada en el Capítulo 13 para ver la normalidad de los datos, se basa en la comparación de la función de distribución muestral con la teórica, la misma idea que se utiliza en el test de Kolmogorov-Smirnov. Este test se encuentra en el procedimiento **Analizar, Pruebas no para-**

métricas, K-S una muestra. Si aplicamos este test para ver si la variable **supervivencia global** del Ejemplo C2.1. (Apéndice C2) sigue una distribución normal (véase Fig. D3.1l), obtenemos un valor de p de 0.019, con lo que se rechaza la normalidad de dicha variable.

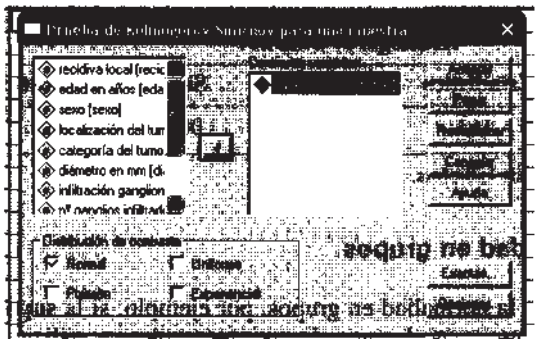


Figura D3.11. Test de ajuste de Kolmogorov-Smirnov.

En el caso de ajuste de unos datos a la distribución normal, el SPSS realiza otros dos test: la versión corregida de Lilliefors para el test de Kolmogorov-Smirnov, y el test de Sapiro-Wilk. Para realizar dichos test, elegimos el procedimiento **Analizar, Estadísticos descriptivos, Explorar;** emerge la ventana de la izquierda de la Figura D3.12: en el campo *Dependientes* ponemos **supervivencia global,** en el campo *Mostrar* marcamos **Gráficos,** y presionamos el botón **Gráficos;** emerge entonces la ventana de la derecha en la Figura D3.12, que debemos rellenar según se indica en ella.

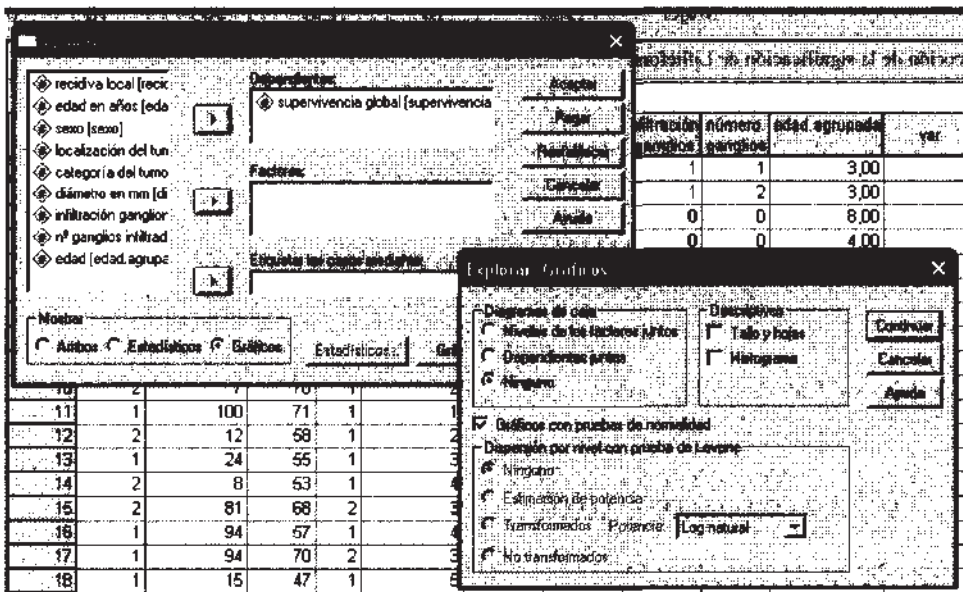


Figura D3.12. Opciones para realizar test de normalidad.

Los resultados de los test se muestran en la Tabla D3.1. Los valores de p para ambos test son 0.000, luego se rechaza la hipótesis de normalidad. Hay que señalar que si se quiere contrastar la normalidad con el test de Kolmogorov-Smirnov, de las dos versiones que calcula el SPSS es esta

última (debida a Lilliefors) la más adecuada (Fig. D3.12). El test de Sapiro-Wilk es adecuado cuando hay pocos datos.

Tabla D3.1. Resultados de los contrastes de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	**gl**	**Sig.**	**Estadístico**	**gl**	**Sig.**
Supervivencia global	.163	88	.000	.901	**88**	.000

[a] Corrección de la significación de Lilliefors.

Contraste de normalidad en grupos

Si quisiéramos contrastar la normaliad en grupos, por ejemplo, si la **supervivencia global** es normal en el grupo de hombres y en el grupo de mujeres, podemos utilizar el mismo procedimiento **Analizar, Estadísticos descriptivos, Explorar,** y cuando emerge la ventana de la Figura D3.12, rellenar el campo *Factores* con la variable **sexo.** Los resultados se pueden ver en la Tabla D3.2, donde vemos que se rechaza la normalidad de la variable **supervivencia global** en varones.

Tabla D3.2. Resultados de **los contrastes** de **normalidad en grupos**

		Kolmogorov-Smirnov*			Shapiro-Wilk		
	Sexo	**Estadístico**	**gl**	**Sig.**	**Estadístico**	**gl**	**Sig.**
Supervivencia global	Varón	.169	**72**	.000	.894	72	.000
	Mujer	.188	**16**	.136	.903	16	.088

[a] Corrección de la significación de Lilliefors.

Inferencia sobre los parámetros de una población

El SPSS permite hacer intervalos de confianza y contrastes de hipótesis bilaterales para la media de una distribución normal. Para el caso del parámetro p de una binomial, sólo permite hacer contrastes de hipótesis.

D4.1. MEDIA DE UNA DISTRIBUCIÓN NORMAL

En el caso de una sola población, el SPSS permite hacer inferencia sobre la media de una población normal cuando tenemos un fichero con todos los datos.

Consideremos el Ejemplo 1.1.1 del Capítulo 1, en concreto la variable **Edad,** que supondremos se distribuye normalmente (el ajuste gráfico se vio en el apartado D3.3 del Apéndice D3). En este estudio se consideran personas a las que se les quiere diagnosticar su situación mental. Supongamos que se quiere contrastar que la edad media es 40 años; para un nivel de significación $\alpha = 0.05$, el contraste se formula como: H_0: $\mu = 40$ frente a H_1: $\mu \neq 40$.

Elegimos el procedimiento **Analizar, Comparar medias, Prueba T para una muestra,** después indicamos en el campo *Contrastar variables* que **Edad** es la variable en estudio y en el campo *Valor de prueba* (Fig. D4.1), ponemos el valor de hipótesis nula, μ_0, en este caso 40; a continuación, pulsamos **Aceptar.** Los resultados se muestran en la Tabla D4.1, donde aparece el test bilateral y el intervalo de confianza al 95% para $\mu-\mu_0$, que calcula el programa por defecto.

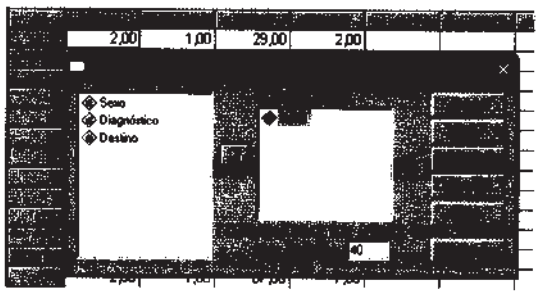


Figura D4.1. Contraste para la media de una normal.

Observando la Tabla D4.1, vemos que el valor del estadístico del contraste (véase Capítulo 6), t en la pantalla de resultados, es 1.666. El valor de p, Sig. (bilateral), se corresponde con $2P(T > 1.666) = 0.104$, siendo T una distribución t de Student con 39 grados de libertad (hay 40 datos). Por lo tanto, como $0.104 > 0.05$, no se rechaza la hipótesis nula, es decir, no hay evidencia estadísticamente significativa de que la edad media sea distinta de 40 años.

También calcula la diferencia entre 44.8 (media muestral de los datos) y 40 (valor de la hipótesis nula), Diferencia de medias, y un intervalo de confianza al 95% para μ - 40.

Tabla D4.1. Inferencia sobre la media de una normal

Prueba para una muestra						
Valor de prueba = 40						
					95% Intervalo de confianza para la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
Edad	1.666	39	.104	4.80000	-1.0265	10.6265

Si ahora queremos calcular un intervalo de confianza para μ, tendríamos que volver a la pantalla de la Figura D4.1 (para ello basta con pulsar el cuarto icono de la barra de herramientas) y cambiar *Valor de prueba* por 0, con lo que el intervalo de confianza que calcularía sería para μ - 0. Otra opción para obtener el intervalo para μ es sumar 40 al intervalo obtenido en la Tabla D4.1, es decir (-1.0265 + 40, 10.6265 + 40). Por cualquiera de los dos caminos obtenemos que el intervalo es (38.9735, 50.6265).

Hay que señalar que para cambiar el nivel de confianza se debe pulsar el botón **Opciones** en la Figura D4.1 y emerge una ventana como la de Figura D4.2, donde se cambia el nivel y después se pulsa **Continuar.** En la Figura D4.2 vemos que debajo del nivel de confianza está marcado, por defecto, *Excluir casos según análisis.* Esto indica que si pedimos simultáneamente intervalos para distintas variables, el número de casos que utiliza para cada intervalo es el número de datos que tiene en cada variable, mientras que si marcamos la otra opción, *Excluir casos según lista,* calcularía todos los intervalos sólo con los casos en los que estuviesen todos los datos de dichas variables.

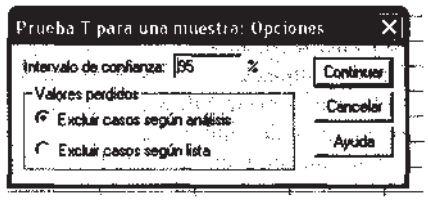


Figura D4.2. Posibles cambios para calcular intervalos de confianza.

Otra forma alternativa de calcular un intervalo de confianza para la media se encuentra seleccionando el procedimiento **Analizar, Estadísticos descriptivos, Explorar** y, por defecto, calcula un intervalo al 95% de confianza para la media de la variable que indiquemos.

D4.2. UNA PROPORCIÓN (MUESTRAS GRANDES)

En el Capítulo 8, se presenta un test para el parámetro p de la binomial cuando el tamaño muestral es grande, ya que en ese caso la distribución binomial se puede aproximar a la normal. En el apartado 13.9, del Capítulo 13, se da como regla de utilización valores de n y p tales que $p < 0.5$ $np > 5$ ó $p > 0.5$ y $n(1 - p) > 5$.

Datos en fichero

Supongamos que en el Ejemplo 1.1.1, del Capítulo 1, queremos ver si el porcentaje de mujeres que llegan para ser diagnosticadas es del 50%, es decir, queremos contrastar H_0: p = 0.5, frente a H_1: $p \neq 0.5$, siendo p la proporción de mujeres.

El SPSS realiza un contraste bilateral que es una versión corregida del estadístico visto en el Capítulo 8.

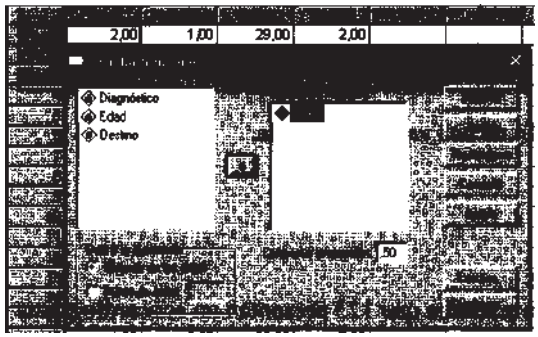


Figura D4.3. Contraste para el parámetro p de una binomial.

Elegimos el procedimiento **Analizar, Pruebas no paramétrícas, Binomial.** Si ahora pulsamos **Aceptar,** emerge una ventana (Fig. D4.3) en la que indicamos la variable para la que queremos hacer el contraste y en el campo *Contrastar proporción,* indicamos el valor de la hipótesis nula, en nuestro caso 0.5. La pantalla de resultados de este procedimiento se presenta en la Tabla D4.2. Vemos que la proporción muestral de mujeres (M) es 0.43 y que el valor de p (Sig.asintót.bilateral) es de 0.430, luego no se rechaza la hipótesis nula.

Tabla D4.2. Resultados para el contraste de p de la binomial

Prueba binomial						
		Categoría	N	**Proporción observada**	**Proporción de prueba**	Sig. **asintót. (bilateral)**
Sexo	Grupo 1	M	17	.43	.50	.430[a]
	Grupo 2	F	23	**.57**		
	Total		40	1.00		

" Basado en la aproximación Z.

En la Figura D4.3 vemos que este procedimiento permite dicotomizar una variable continua para el estudio de una proporción. Por ejemplo, si queremos ver si el 25% de las personas son menores de 30 años, pondríamos en el campo *Contrastar variables* **Edad;** en el campo *Definir la dicotomía,* marcaríamos *Punto de corte* (véase Fig. D4.4) y pondríamos el valor 30; por último, en el campo *Contrastar proporción,* pondríamos **0.25.**

Datos resumidos

Si nuestros datos están resumidos en sus frecuencias, también se puede aplicar el test anterior.

Resolvamos el Ejercicio 8.4.7, del Capítulo 8, donde se quiere estudiar si el 85% de los niños con dolor torácico tienen ecocardiograma normal. Para ello, se toma una muestra de 139 niños con

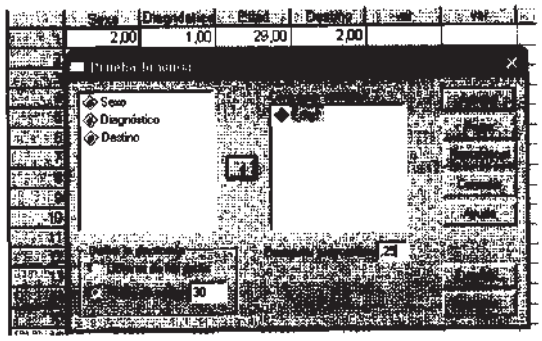


Figura D4.4. Dicotomización de una variable continua.

dolor torácico, se estudian y resulta que 123 presentan un ecocardiograma normal. El contraste setá H_0: p = 0.85, frente a H_0: p 0.85, siendo p la proporción de ecocardiograma normal entre los niños" con dolor torácico.

Primero creamos los datos en la Pantalla de datos como en la Figura D4.5, donde ecocardiograma normal se codifica como 1 y no normal como 2. Elegimos el procedimiento **Datos, Ponderar casos,** y en la ventana emergente indicamos la *Variable de frecuencia* que hemos llamado **Frecuencia.** Dicha variable es con la que vamos a ponderar, ya que contiene las frecuencias (véas ventana superior de Fig. D4.5). Después, ya podemos aplicar el test visto en el apartado anterior (Datos en fichero) (véase ventana inferior de Fig. D4.5).

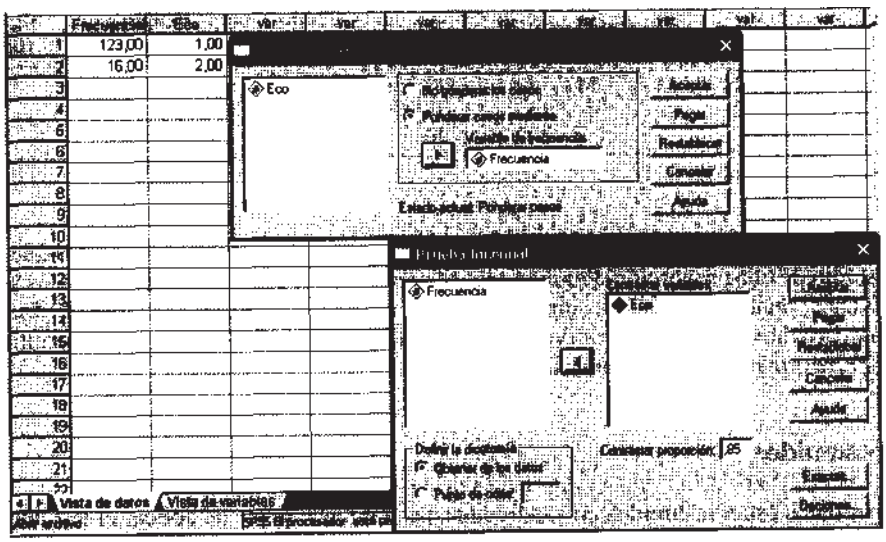


Figura D4.5. Contraste de una proporción con datos resumidos.

En la Tabla D4.3 se muestran los resultados para nuestro ejercicio. La proporción observada de niños con ecocardiograma normal es 0.88 (123/139) y el valor de p es 0.150, es decir, la diferencia no es estadísticamente significativa y, en consecuencia, no se puede rechazar que la proporción de niños con ecocardiograma normal sea 0.85.

Tabla D4.3. Resultados para el contraste de una **proporción**

Prueba binomial						
		Categoría	**N**	**Proporción observada**	**Proporción de prueba**	**Sig. asintót. (unilateral)**
Eco	Grupo 1	Normal	123	**.88**	**.85**	.150[a]
	Grupo 2	No normal	**16**	**.12**		
	Total		**139**	1.00		

* Basado en la aproximación Z.

Comparación de dos poblaciones

Cuando se aborda el problema de comparar dos poblaciones, la primera pregunta que debemos responder es si conocemos la distribución de la variable de interés en ambas poblaciones. Si dicha distribución sigue un modelo de los llamados paramétricos, la comparación de ambas poblaciones se reduce a la comparación de los parámetros que las caracterizan y los métodos de inferencia utilizados serán los conocidos como métodos paramétricos, mientras que si la distribución de los datos es desconocida deberían utilizarse métodos no paramétricos. Lo más frecuente es que la distribución de los datos sea, o pueda suponerse, normal. En este caso, el interés se centra en comparar las medias y/o las varianzas de ambas poblaciones.

Una segunda cuestión se refiere al tipo de muestreo utilizado. Así, si las dos muestras se obtuvieron de manera independiente, los métodos de inferencia estadística sobre los parámetros diferirán de los utilizados para el análisis de muestras apareadas o emparejadas.

En este Apéndice vamos a presentar los procedimientos del SPSS para la comparación de dos poblaciones normales independientes, de dos poblaciones normales apareadas y, finalmente, de dos proporciones independientes.

D5.1. COMPARACIÓN DE DOS MEDIAS Y DOS VARIANZAS DE POBLACIONES NORMALES INDEPENDIENTES

Considerando los datos contenidos en el Ejemplo C2.1 del Apéndice C2, vamos a comparar el diámetro del tumor en los dos grupos que define la variable recidiva local. Más en concreto, vamos a comparar estadísticamente las medias del diámetro del tumor de los pacientes con y sin recidiva local. Las dos muestras son claramente independientes. Supondremos que el diámetro del tumor se distribuye normalmente en ambas poblaciones. Fijamos el nivel de significación del contraste $\alpha = 0.05$.

Para llevar a cabo este análisis, elegimos el procedimiento **Analizar, Comparar inedias, Prueba T para muestras independientes.** Una vez abierta la ventana del procedimiento, es necesario seleccionar la variable **diámetro** y arrastrarla al campo *Contrastar variables,* seleccionar la variable **recidiva local** y arrastrarla al campo *Variable de agrupamiento;* debajo de este último campo se activa el botón *Definir grupos,* y pulsando sobre él se abre una ventana que muestra dos campos que debemos completar de la forma siguiente: *Grupo 1:* 1 y *Grupo* 2: 2; pulsamos **Continuar.** Los valores 1 y 2 son los códigos que figuran en la base de datos para nombrar la ausencia y presencia de recidiva local, respectivamente. Con la elección anterior, el procedimiento con-

siderará como primera muestra al grupo de pacientes sin recidiva local. En la inferencia acerca de la diferencia de medias, el minuendo será la media de esta primera muestra. Volviendo al procedimiento, si pulsamos el botón **Opciones,** podremos cambiar el nivel de confianza del intervalo para la diferencia de medias que, por defecto, es del 95%. Pulsando **Aceptar,** se obtienen los resultados que aparecen en la Tabla D5.1

Tabla D5.1. Comparación del diámetro del tumor entre pacientes con y sin recidiva local

Estadísticos de grupo

	Recidiva local	N	Media	Desviación típica	Error típico de la media
Diámetro en mm	**No**	51	24.08	9.051	1.267
	Sí	**37**	33.30	12.858	2.114

Prueba de muestras independientes

			Diámetro en mm	
			Se han asumido varianzas iguales	**No se han asumido varianzas iguales**
Prueba de Levene para la igualdad de varianzas	F		5.894	
	Sig.		17	
Prueba T para la igualdad de medias	**t**		-3.950	-3.740
	gl		86	60.870
	Sig. (bilateral)		.000	.000
	Diferencia de medias		-9.219	-9.219
	Error típico de la diferencia		2.334	2.465
	95% Intervalo de confianza para la diferencia	Inferior	-13.859	-14.147
		Superior	-4.579	-4.290

La primera parte de la tabla, Estadísticos de grupo, muestra los tamaños, las medias y las desviaciones típicas de los dos grupos, proporcionando una primera aproximación a la comparación que queremos llevar a cabo.

Una medida de especial importancia en la comparación de dos medias de poblaciones independientes es la varianza, pues de sus valores en los dos grupos, y más en concreto de su cociente, se derivará una u otra prueba estadística (véase el apartado 9.2 del Capítulo 9). Antes de comparar las medias debemos, pues, comparar las varianzas mediante un contraste que se formula de la forma siguiente: $\mathbf{H_0}$: $\sigma_1^2 = \sigma_2^2$ (varianzas iguales) frente a H_1: $\sigma_1^2 \neq \sigma_2^2$ (varianzas distintas). Para llevar a cabo este contraste existen diferentes estadísticos. Uno de ellos es el test de la F de Snedecor, al que acabamos de aludir, que se basa en el cociente de las varianzas muestrales. El SPSS no realiza este contraste de igualdad de varianzas, ejecutando en su lugar el test de Levene basado en la media y preferido por muchos autores, por ser poco sensible a la falta de normalidad de la variable. La segunda parte de la tabla, Prueba de muestras independientes, muestra los resultados de este último test. El valor del estadístico del contraste, F, es 5.894 y el valor p asociado, Sig., es 0.017, que es < 0.05 (nivel α establecido) por lo que se debe rechazar la hipótesis de que las varianzas poblacionales son iguales.

El contraste de igualdad de medias bilateral (con dos colas) se formula de la forma siguiente: H_0: $\mu_1 - \mu_2 = 0$ (diámetros medios iguales) frente a H_1: $\mu_1 - \mu_2 \neq 0$ (diámetros medios distintos). Los resultados de este contraste figuran a continuación del test de Levene en la Tabla D5.1. Debemos elegir la columna «No se han asumido varianzas iguales» pues éste fue el resultado del test de Levene. El

valor del estadístico del contraste, t, es -3.74 y el valor p asociado, Sig. (bilateral), es 0.000 (es decir, el valor de p es < 0.001) y, por tanto, menor que el nivel de significación establecido, por lo que se debe rechazar la hipótesis de que las medias poblacionales sean iguales, es decir, existen diferencias significativas entre los diámetros medios de ambos grupos. El análisis finaliza estimando esta diferencia mediante un intervalo de confianza al 95%. Dicho intervalo, para la diferencia de medias $\mu_1 - \mu_2$, es (-14.147, - 4.290); es decir, entre 4.29 y 14.147 mm se estima la diferencia de los diámistros medios del tumor de los pacientes con recidiva respecto de los pacientes sin recidiva.

Si deseamos llevar a cabo un contraste unilateral, por ejemplo H_0: $\mu_1 - \mu_2 = 0$ (diámetros medios iguales) frente a H_1: $\mu_1 - \mu_2 < 0$ (el diámetro medio del grupo con recidiva es mayor) el estadístico del contraste seguirá siendo -3.74 y el valor p asociado será la mitad del obtenido para el test bilateral, es decir, < 0.0005, por lo que se aceptaría la hipótesis alternativa H_1.

D5.2. COMPARACIÓN DE DOS MEDIAS PARA DATOS APAREADOS. POBLACIONES NORMALES.

En el Ejercicio 9.5.4 del Capítulo 9 se propone analizar el efecto de un programa de ejercicios regulares sobre la capacidad de trabajo en pacientes que habían sufrido previamente un infarto de miocardio. Para 11 sujetos, se mide el tiempo transcurrido hasta alcanzar una frecuencia de 160 latidos por minuto mientras caminan sobre una cinta sin fin, antes y después de 25 semanas de ejercicio controlado. Los datos aparecen en dicho ejercicio.

La hipótesis de los investigadores es que el tiempo transcurrido hasta alcanzar la frecuencia cardíaca prefijada será mayor después de las 25 semanas de ejercicio. Por consiguiente, la hipótesis estadística que se desea poner de manifiesto será H_1: $\mu_D - \mu_A > 0$, siendo μ_A y μ_D los tiempos medios, hasta alcanzar los 160 latidos por minuto, antes y después del período de ejercicios, respectivamente. Se trata pues de un contraste unilateral para la diferencia de medias de dos muestras apareadas o emparejadas (véase apartado 9.5 del Capítulo 9). Vamos a suponer que el tiempo se distribuye normalmente. Fijamos el nivel de significación del contraste $\alpha = 0.1$.

Para llevar a cabo este análisis elegimos el procedimiento **Analizar, Comparar medias, Prueba T para muestras relacionadas.** Una vez abierta la ventana del procedimiento, si en primer lugar seleccionamos la variable **tiempo después** ésta pasará a *Variable 1* en el campo *Selecciones actuales,* y si ahora seleccionamos la variable **tiempo antes** pasará a *Variable 2* del campo *Selecciones actuales;* pulsando sobre el botón de flecha  , la diferencia **tiempo después-tiempo antes** pasará al campo *Variables relacionadas.* Los resultados que obtengamos se referirán pues a esta diferencia de medias. Si pulsamos el botón **Opciones** podremos cambiar el nivel de confianza del intervalo para la diferencia de medias que, por defecto, es del 95%. Pulsando **Aceptar** se obtienen, entre otros, los resultados que aparecen en la Tabla D5.2.

Tabla D5.2. Comparación del tiempo antes y después del período de ejercicios

Prueba de muestras relacionadas

			Par 1
			Tiempo después-tiempo antes
Diferencias relacionadas	Media		5.12727
	Desviación típica		1.48195
	Error típico de la media		.44683
	95% Intervalo de confianza para la diferencia	Inferior	4.13168
		Superior	6.12286
t			11.475
gl			10
Sig. (bilateral)			.000

La Tabla D5.2 muestra, en sus tres primeras filas, diferentes estadísticos relativos a la diferencia de tiempos (después-antes), el valor medio, 5.12727, la desviación típica, 1.48195 y el error típico estimado de la media (véase el punto 6 de los Ejercicios 6.2 del Capítulo 6), 0.44683. En las tres últimas filas se muestran los resultados del contraste bilateral que coinciden, salvo el valor de p del test, con los del contraste unilateral solicitado. Así, el valor del estadístico del contraste, t, es 11.475; los grados de libertad de la t de Student en la que se basa el test, gl, son 10; y el valor de p asociado será la mitad del valor de p mostrado en la tabla con la denominación Sig. (bilateral), es decir, el valor de p del contraste solicitado será < 0.0005, y por tanto, < 0.1 (nivel α establecido), por lo que podemos afirmar que el tiempo transcurrido hasta alcanzar la frecuencia cardíaca prefijada será significativamente mayor después de las 25 semanas de ejercicios.

Finalmente, el intervalo de confianza al 95% para la diferencia de tiempos medios (después-antes) es (4.13168, 6.12286).

D5.3. COMPARACIÓN DE DOS PROPORCIONES INDEPENDIENTES. MUESTRAS GRANDES

Considerando los datos contenidos en el Ejemplo C2.1 del Apéndice C2, vamos a comparar la proporción de recidiva local en los dos grupos que define la variable infiltración ganglionar. Queremos llevar a cabo un contraste bilateral para evaluar si existen diferencias significativas entre ambas proporciones muestrales. Para ello, formulamos como hipótesis estadísticas las siguientes: $H_0 : p_1 - p_2 = 0$ (proporciones iguales) frente a $\mathbf{H_1: p_1 - p_2 \neq 0}$ (proporciones distintas), siendo p_1 y p_2 las proporciones teóricas de recidiva local en los grupos con y sin infiltración ganglionar, respectivamente. Fijamos el nivel de significación del contraste $\boldsymbol{\alpha = 0.1}$.

El contraste de proporciones descrito en el apartado 8.6 del Capítulo 8 se basa en la distribución aproximada de un estadístico muestral que requiere además muestras grandes. El SPSS no incluye el cálculo de este estadístico pero permite el cálculo de otros 4 estadísticos para muestras grandes y el estadístico exacto de Fisher, dentro de un procedimiento general, Tablas de contingencia, que en su aspecto descriptivo ya ha sido comentado en el Apartado D2.4 (Apéndice D2).

La comparación de dos proporciones es un caso particular del problema general de contrastar la homogeneidad de dos muestras de una variable cualitativa cuando ésta sólo presenta dos modalidades (véase Apéndice D8). Por ello, el procedimiento que debemos ejecutar será el análisis de una tabla de contingencia 2x2. Elegimos el procedimiento **Analizar, Estadísticos descriptivos, Tablas de contingencia.** Una vez emerge la ventana del procedimiento, es necesario seleccionar la variable **infiltración ganglionar** y arrastrarla al campo *Filas,* seleccionar la variable **recidiva local** y arrastrarla al campo *Columnas,* pulsar el botón **Casillas** y seleccionar en *Frecuencias,* **Observadas,** y en *Porcentajes,* **Fila,** pulsar **Continuar,** pulsar el botón **Estadísticos** y seleccionar **Chi-cuadrado** y, de nuevo en la ventana del procedimiento, pulsar **Aceptar.** Las Tablas D5.3 y D5.4 muestran la tabla de contingencia y los contrastes chi-cuadrado respectivamente.

Tabla D5.3. Tabla de contingencia correspondiente a la recidiva local según la infiltración ganglionar

			Recidiva local		
			No	Sí	Total
Infiltración ganglionar	No	Recuento	39	21	60
		% de infiltración ganglionar	65.0%	35.0%	100.0%
	Sí	Recuento	12	16	28
		% de infiltración ganglionar	42.9%	57.1%	100.0%
Total		Recuento	51	37	88
		% de infiltración ganglionar	58.0%	42.0%	100.0%

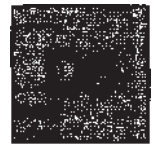
Tabla D5.4. Comparación de proporciones de recidiva local en pacientes con y sin infiltración ganglionar

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	3.841[b]	1	.050		
Corrección por continuidad[a]	2.986	1	.084		
Razón de verosimilitud	3.821	1	.051		
Estadístico exacto de Fisher				.065	.042
Asociación lineal por lineal	3.798	1	.051		
N de casos válidos	88				

[a] Calculado sólo para una tabla de 2 x 2.
[b] 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 11.77.

La Tabla D5.3 muestra la tabla de frecuencias conjunta. En cada casilla aparece, además de la frecuencia observada, el porcentaje que ésta representa sobre el total de la fila. Así, las proporciones muestrales que vamos a comparar son $\hat{p}_1 = (16/28) = 0.571$ y $\hat{p}_2$ - (21/60) = 0.35.

La Tabla D5.4 muestra los resultados de 5 estadísticos para la comparación de ambas proporciones. De modo general, para muestras grandes atenderemos al segundo de los estadísticos, Corrección por continuidad, que calcula el estadístico chi-cuadrado con la corrección de Yates, y para muestras pequeñas decidiremos a partir del estadístico exacto de Fisher. El valor del estadístico del contraste chi-cuadrado corregido es 2.986 y su valor p asociado es 0.084 (Sig. asintótica bilateral). El valor de p de la prueba exacta de Fisher es 0.065 (Sig. exacta bilateral). Comparando ambos valores de p con el nivel $\alpha = 0.1$ establecido, la decisión, con ambos estadísticos, es rechazar la hipótesis H_0, es decir, las diferencias observadas entre $\hat{p}_1$ y $\hat{p}_2$ son estadísticamente significativas.



Análisis de la varianza

Vamos a considerar el Ejemplo C6.1 del Apéndice C6 para ilustrar las técnicas de análisis de la varianza expuestas en el Capítulo 10. Consideraremos los modelos con uno y dos factores de clasificación, también conocidos como ANOVA de una y dos vías, respectivamente. Todos los modelos requieren, básicamente, dos hipótesis estructurales acerca de la variable respuesta, la normalidad e igualdad de varianzas de dicha variable en todas las poblaciones. Los contrastes para someter a prueba la hipótesis de normalidad fueron comentados en el Apartado D3.4 (Apéndice D3). Por ello, y al objeto de acortar la exposición, supondremos, en todos los casos, que la variable respuesta sigue una distribución normal. No obstante, si la respuesta se desvía de la normalidad, el procedimiento sigue siendo válido siempre que los tamaños muestrales no sean pequeños.

D6.1. ANÁLISIS DE LA VARIANZA DE UNA VÍA

Vamos a comparar el perímetro craneal, pc, de los recién nacidos atendiendo a la nacionalidad de los mismos; en concreto, queremos contrastar la hipótesis de que los perímetros medios de las tres poblaciones coinciden.

El contraste de igualdad de medias

El contraste estadístico se formula de la siguiente forma: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu$, H_1: $\mu_i \neq \mu$ para algún i = 1,2, 3.Vamos a suponer que los datos provienen de seleccionar tres muestras aleatorias de recién nacidos, una de cada nacionalidad. El diseño experimental corresponde pues, a un diseño de una vía, completamente aleatorio con efectos fijos (véase apartado 10.1 del Capítulo 10). Para llevar a cabo este contraste con el SPSS, debemos elegir el procedimiento **Analizar, Comparar medias, ANO VA de un factor.**.. Una vez emerge la ventana del procedimiento, es necesario seleccionar la variable pc y arrastrarla al campo *Dependientes,* seleccionar la variable **nacionalidad** y arrastrarla al campo *Factor,* pulsando **Aceptar,** obtenemos la Tabla de ANOVA que aparece en la Tabla D6.1.

El estadístico del contraste figura en la columna «F» y vale 5.292, y su valor p asociado es 0.007, por lo que se rechaza la hipótesis nula para cualquier nivel de significación $\alpha > 0.008$. Es decir, existen diferencias estadísticamente significativas entre los perímetros craneales de los grupos.

Tabla D6.1. Tabla de ANOVA para el perímetro craneal (pc) según nacionalidad

pc	Suma de cuadrados	gl	Media cuadrática	F	Sig.
ínter-grupos	1280.202	2	640.101	5.292	.007
Intra-grupos	8587.703	71	120.954		
Total	9867.905	73			

El contraste de igualdad de varianzas

Dentro del procedimiento que acabamos de describir, **Analizar, Comparar medias, ANOVA de un factor...,** tenemos la opción de contrastar la hipótesis previa de homogeneidad de las tres varianzas, $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$, $H_1 : \sigma_i^2 \neq \sigma^2$ para algún i = 1, 2, 3. Para ello, en la ventana del procedimiento, tendremos que pulsar **Opciones,** y en *Estadísticos* marcar **Prueba de homogeneidad de la varianza,** pulsar **Continuar y Aceptar.** El resultado es el contraste de Levene basado en la media. Existe otro procedimiento en el SPSS que permite llevar a cabo el mismo contraste de homogeneidad de varianzas pero calculando, además del test de Levene, otros tres estadísticos adicionales. Para ello, debemos elegir el procedimiento **Analizar, Estadísticos descriptivos, Explorar...** , completar los campos *Dependientes y Factores* arrastrando las variables **pc** y **nacionalidad,** respectivamente; en *Mostrar,* elegir **Gráficos,** pulsar el botón **Gráficos** y elegir cualquier gráfico de la lista, y en la opción *Dispersión por nivel con prueba de Levene,* elegir **Estimación de potencia** (véase Fig. D3.12 del Apéndice D3). Los resultados se muestran en la Tabla D6.2.

		Estadístico de Levene	gl1	gl2	Sig.
pc	Basándose en la media	4.173	2	71	.019
	Basándose en la mediana	1.644	2	71	.200
	Basándose en la mediana y con gl corregido	1.644	2	47.499	.204
	Basándose en la media recortada	3.816	2	71	.027

Se muestran los resultados del test de Levene y de otros tres test de Levene modificados. Él segundo de ellos, el que se basa en la mediana muestral resulta más eficaz que los otros, sobre todo cuando existe asimetría en la distribución. Vamos a basar la decisión de nuestro contraste en el valor de p de este test que, al ser 0.2, conduce a no rechazar la hipótesis de igualdad de las tres varianzas poblacionales.

La discusión acerca del test más adecuado a cada situación escapa del contenido de este libro. No obstante, conviene recordar que el test de Levene basado en la mediana es el que calcula el paquete STATGRAPHICS Plus, con el nombre de test de Levene, al tratar este problema. También es importante añadir que el contraste de igualdad de medias sigue siendo válido aunque las varianzas difieran mucho, siempre que los tamaños muestrales sean semejantes. Si se desea ampliar estos comentarios puede leerse el apartado **El contraste de igualdad de varianzas** del Apéndice C6.lL

El SPSS permite llevar a cabo el contraste de igualdad de medias en caso de varianzas diferentes y tamaños muestrales dispares. Pulsando el botón **Opciones,** hay que elegir en **Estadísticos, Brown-Forsythe y Welch,** pulsar **Continuar y Aceptar.**

Los contrastes de normalidad a los que aludíamos en la introducción de este Apéndice, y que se comentan en el apartado D3.4 del Apéndice D3, se llevan a cabo dentro de este último procedimiento que acabamos de comentar. Tras pulsar el botón **Gráficos,** bastará marcar **Gráficos con pruebas de normalidad** para obtener dichos contrastes.

Comparaciones múltiples

Una vez que encontramos diferencias estadísticamente significativas entre las medias de los perímetros craneales de las tres nacionalidades, el análisis debe proseguir para encontrar la/s media/s responsable/s de dicha significación. Para llevar a cabo las comparaciones de medias por parejas, en la ventana del procedimiento Analizar, Comparar medias, ANOVA de un factor..., pulsamos el botón Post hoc. Podemos elegir hasta 14 test *a posteriori* diferentes en la opción *Asumiendo varianzas iguales,* que es nuestro caso, y hasta 4 test *a posteriori* en la opción *No asumiendo varianzas iguales.* Vamos a elegir dentro de la primera opción Bonferroni y completamos la ventana modificando, si se desea, el *Nivel de significación* que, por defecto es 0.05; pulsamos Continuar y Aceptar. Los resultados obtenidos se muestran en la Tabla D6.3.

Tabla D6.3. Comparaciones múltiples. Método de Bonferroni

Variable dependiente: pc
Bonferroni

(I) Nacionalidad	(J) Nacional	Diferencia de medidas (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Española	Africana	-9.058(*)	3.133	.015	-16.74	-1.38
	Oriental	.089	3.059	1.000	-7.41	7.59
Africana	Española	9.058(*)	3.133	.015	1.38	16.74
	Oriental	9.148(*)	3.246	.019	1.19	17.11
Oriental	Española	-.089	3.059	1.000	-7.59	7.41
	Africana	-9.148(*)	3.246	.019	-17.11	-1.19

* La diferencia entre las medias es significativa al nivel .05.

La Tabla D6.3 presenta los resultados de los tres contrastes de hipótesis, $H_0 : \mu_i = \mu_j$, $H_1 : \mu_i \neq \mu_j$ para i, j = 1,2, 3, utilizando el método de Bonferroni (véase apartado 10.2 del Capítulo 10). Este método lleva a cabo todos los posibles contrastes entre pares de medias.

Los valores de p de los diferentes contrastes figuran en la columna «Sig». Sobre éstos, podemos concluir que existen diferencias significativas entre el perímetro medio de la nacionalidad 2, africana, y los de las nacionalidades 1 y 3, española y oriental, respectivamente, es decir, para $\alpha = 0.05$ la conclusión de las comparaciones múltiples es que $\mu_2 \neq \mu_1, \mu_3$. El símbolo * en la tabla destaca las diferencias significativas.

Finalmente, en las dos últimas columnas de la tabla se muestran los extremos de los intervalos de confianza, al 95%, para las diferencias de medias. Lógicamente, tienen mayor interés los que se refieren a los grupos que resultaron significativos. Por ejemplo, el intervalo de confianza para $\mu_{africana} - \mu_{española}$, al 95%, es (1.38,16.74).

Un método alternativo al de Bonferroni, menos conservador que éste (se requiere menor diferencia observada para declarar significativa dicha diferencia) es el método de Duncan (véase apartado 10.2 del Capítulo 10) que no requiere llevar a cabo todas las comparaciones. Para obtener los resultados de este método basta elegir **Duncan** en la opción *Asumiendo varianzas iguales* de las opciones **Post hoc,** que acabamos de comentar. Conviene puntualizar que si los tamaños muestrales son iguales y queremos llevar a cabo todas las comparaciones, podemos usar el método de Tukey, que proporciona resultados más precisos que el de Bonferroni. Finalmente, si deseamos construir un único intervalo de confianza para la diferencia de dos medias, debemos elegir el método **DMS.**

Podemos ilustrar los resultados con el Gráfico de las medias. Para ello, es necesario pulsar el botón **Opciones** en la ventana del procedimiento y marcar **Gráfico de las medias,** pulsar **Continuar y Aceptar.** La Figura D6.1 muestra dicha gráfica.

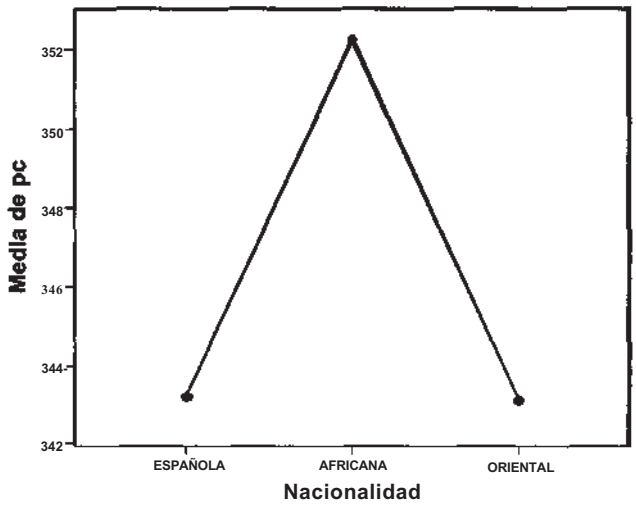


Figura D6.1. Gráfico de las medias del perímetro craneal según las tres nacionalidades.

D6.2. ANÁLISIS DE LA VARIANZA DE DOS VÍAS

Vamos a comparar ahora el perímetro craneal de los recién nacidos atendiendo al sexo y a la nacionalidad. Para ello, vamos a seleccionar, entre los nacidos de cada nacionalidad, dos muestras de tamaño 10, una de niños y otra de niñas; de esta forma dispondremos de 6 muestras de tamaño 10 cada una, que corresponden a las 6 combinaciones de niveles de ambos factores. El diseño experimental corresponde pues, a un diseño de dos vías, completamente aleatorio con efectos fijos (véase apartado 10.5 del Capítulo 10). El hecho de elegir todas las muestras de igual tamaño obedece a facilitar el método de análisis.

Para llevar a cabo este análisis con el SPSS, debemos elegir el procedimiento **Analizar, Modelo lineal general, Univariante.** Una vez emerge la ventana del procedimiento, tendremos que seleccionar la variable pc y arrastrarla al campo *Dependiente,* y las variables **nacionalidad** y sexo y arrastrarlas al campo *Factores fijos;* pulsando **Aceptar** se obtienen los resultados que muestra la Tabla D6.4. El modelo ajustado por defecto incluye un término de interacción entre el sexo y la nacionalidad.

La primera hipótesis que debemos contrastar es la de no interacción sexo-nacionalidad, que se formula como sigue: $H_0 : (\alpha\beta)_{ij} = 0$, $H_1 : (\alpha\beta)_{ij} \neq 0$ para algún i = 1, 2; j = 1, 2, 3 (véase apartado 10.5 del Capítulo 10). El valor del estadístico del contraste figura en la columna «F» y es 0.217 y su valor p, en la columna «Significación», es 0.805, lo que conduce a no rechazar la hipótesis nula, es decir, no hay evidencia estadística de interacción sexo-nacionalidad. Por consiguiente, podemos proseguir el análisis y llevar a cabo los contrastes para los efectos principales, que se formulan como sigue: (a) $H_0 : \mu_{1..} = \mu_{2..} = \mu_{sexo}$ y (b) $H_0 : \mu_{.1.} = \mu_{.2.} = \mu_{.3.} = \mu_{nacional}$. Los estadísticos de los contrastes (a) y (b) son, respectivamente, 22.256 y 7.033, y sus valores p asociados son $p < 0.0001$ y $p = 0.002$, por lo que se rechazan ambas hipótesis nulas, es decir, existen diferencias significativas entre los perímetros craneales medios de ambos sexos, y entre los de las tres nacionalidades.

El análisis proseguiría abordando las comparaciones múltiples para las tres nacionalidades. En el caso del sexo, por tener sólo 2 modalidades, el análisis finaliza con el contraste (a) anterior. Para ello, volveríamos a la ventana del procedimiento **Analizar, Modelo lineal general, Univariante,**

Tabla D6.4. Tabla de ANOVA para el perímetro craneal según nacionalidad y sexo

Pruebas de los **efectos ínter-sujetos**
Variable dependiente: perímetro craneal

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	3667.083(a)	5	733.417	7.351	.000
Intersección	7200270.417	1	7200270.417	72169.764	.000
Nacionalidad	1403.333	2	701.667	**7.033**	**.002**
Sexo	2220.417	1	2220.417	**33.256**	**.000**
Nacionalidad*Sexo	43.333	2	21.667	**.217**	**.805**
Error	5387.500	54	99.769		
Total	7209325.000	60			
Total corregida	9054.583	59			

(a) R cuadrado = .405 (R cuadrado corregida = .350).

y pulsaríamos el botón **Post hoc,** seleccionaríamos **nacionalidad** en *Factores* y la arrastraríamos al campo *Contrastes post hoc para;* en la opción *Asumiendo variamos iguales,* seleccionaríamos los test estadísticos deseados, Tukey y Duncan, pulsaríamos **Continuar** y **Aceptar** y obtendríamos una tabla similar a la Tabla D6.3, que nos permite interpretar los resultados obtenidos, comparaciones significativas, intervalos de confianza, de forma semejante a la expuesta en el apartado **Comparaciones múltiples** del apartado D6.1 de este Apéndice.

Podemos solicitar las medias muestrales de todas las casillas por sexos, por nacionalidades y la media global. Por ejemplo, para obtener las medias de todas las casillas debemos volver **a** la ventana del procedimiento **Analizar, Modelo lineal general, Univariante,** y pulsar el botón **Opciones;** completamos el campo *Mostrar medias para,* seleccionando **nacionalidad*sexo** en *Medias marginales estimadas;* pulsamos **Continuar y Aceptar.** La Tabla D6.5 muestra los resultados. Esta tabla es especialmente útil cuando la hipótesis de no interacción se rechaza, pues permite identificar la/s casilla/s responsable/s de la interacción.

Si deseamos visualizar esta información en un gráfico, debemos volver a la ventana del procedimiento, pulsar el botón **Gráficos;** seleccionar, en *Factores,* **nacionalidad** para el campo *Eje horizontal* y **sexo** para el campo *Líneas distintas;* pulsar **Añadir, Continuar y Aceptar.** La Figura D6.2 muestra esta gráfica. El paralelismo de los segmentos del gráfico resulta una confirmación geométrica de la no interacción de los dos factores estudiados: el decremento del perímetro craneal de niños a niñas es prácticamente el mismo en las tres nacionalidades.

Para finalizar conviene añadir que si la interacción resultara significativa, el estudio de los efectos principales exigiría analizar las posibles diferencias entre los perímetros de las tres nacionalidades para los recién nacidos niños, y estudiar lo mismo para las niñas. El procedimiento estadístico consistiría en dos ANOVA de una vía, uno para niños y otro para niñas.

Tabla D6.5. Tabla de medias para el perímetro craneal

Nacionalidad * Sexo
Variable dependiente: perímetro craneal

Nacionalidad	Sexo	Media	Error típico	Intervalo de confianza al 95%	
				Límite inferior	Límite superior
Española	Niño	350.000	3.159	343.667	356.333
	Niña	335.500	3.159	329.167	341.833
Africana	Niño	359.000	3.159	352.667	365.333
	Niña	347.500	3.159	341.167	353.833
Oriental	Niño	348.500	3.159	342.167	354.833
	Niña	338.000	3.159	331.667	344.333

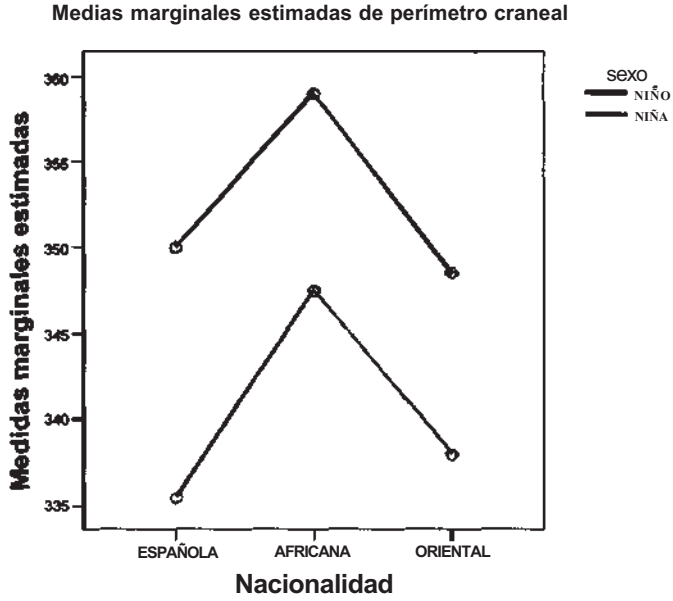


Figura D6.2. Gráfico de interacción sexo-nacionalidad.

D6.3. BLOQUES COMPLETOS ALEATORIZADOS

Si el diseño experimental se corresponde con un diseño de Bloques completos aleatorizados (véase apartado 10.4 del Capítulo 10), el análisis estadístico se lleva a cabo utilizando el modelo de ANOVA con dos vías, efectos fijos sin interacción, con una observación por casilla. Por ello, debemos elegir el procedimiento **Analizar, Modelo lineal general, Univariante,** y definir el modelo para no incluir la interacción bloque-factor. Para hacer esto, es necesario pulsar el botón **Modelo** y en *Especificar modelo* elegir **Personalizado,** arrastrar al campo *Modelo* sólo los efectos principales, es decir, el bloque y el factor de interés, que deben figurar en el cuadro *Factores y Covariables*. El contraste deseado, la comparación de los diferentes niveles del factor de interés, se lleva a cabo en la tabla de ANOVA, similar a la Tabla D6.4, a partir del valor de p correspondiente al estadístico del contraste del efecto principal relativo al factor de interés. En la tabla de ANOVA aparecerá también el contraste relativo a la variable de bloque, que en este diseño tiene un interés menor.

Regresión y correlación

D7.1. REGRESIÓN LINEAL SIMPLE

Para ilustrar este procedimiento vamos a utilizar el Ejemplo C6.1 del Apéndice C6. Deseamos estudiar la posible relación lineal entre la talla y el peso de los recién nacidos. Consideraremos la talla como variable dependiente o respuesta y el peso como variable independiente o predictora. Se trata de un estudio observacional, al escapar al control del investigador los valores de la variable predictora. Para visualizar la nube de puntos debemos elegir el procedimiento **Gráficos** de la barra de menú; en la ventana que emerge, seleccionar *Dispersión,* elegir la opción *Simple,* pulsar **Definir,** *Eje* ***Y:*** **talla,** *Eje* ***X:*** **peso,** y pulsar **Aceptar.** Obtendremos la nube de puntos de la Figura D7.1. El gráfico muestra la adecuación del modelo lineal y la tendencia creciente del mismo.

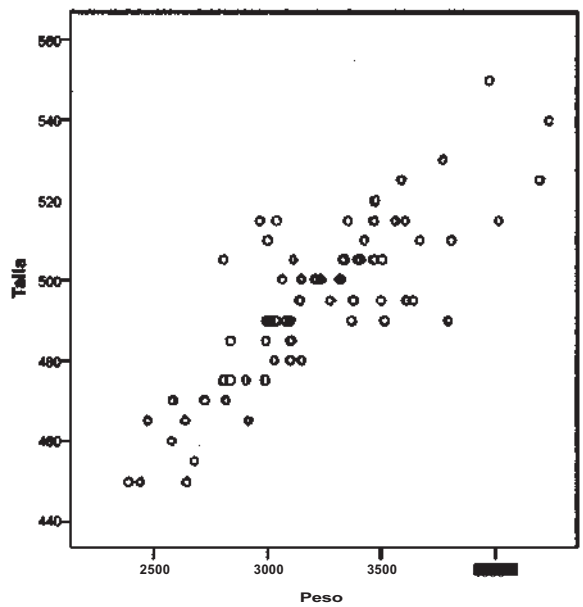


Figura D7.1. Nube de puntos de las tallas frente a los pesos.

Para obtener la recta de regresión mínimo cuadrática de la talla sobre el peso debemos elegir el procedimiento **Analizar, Regresión, Lineal;** una vez emerge la ventana del procedimiento, tendremos que seleccionar la variable **talla** y arrastrarla al campo *Dependiente,* seleccionar la variable **peso** y arrastrarla al campo *Independientes,* elegir *Método:* **Introducir,** pulsar el botón **Estadísticos** y elegir en *Coeficientes de regresión:* **Estimaciones** e **Intervalos de confianza,** marcar **Ajuste del modelo,** pulsar **Continuar,** pulsar el botón **Gráficos** y elegir *Y:* ***ZRESID, y** ***X:* ***ZPRED;** por último, habremos de elegir en *Gráficos de residuos tipificados:* **Gráfico de prob. Normal,** pulsar **Continuar y Aceptar.** Se obtienen, entre otros, los resultados que se muestran en las Tablas D7.1 y D7.2 y en las Figuras D7.2 y D7.3.

Tabla D7.1. Ajuste del modelo de regresión. Coeficiente de determinación y tabla de ANOVA

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típico de la estimación
1	.823[a]	.678	.674	11.605

[a] Variables predictoras: (Constante), peso.
[b] Variable dependiente: talla.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	20416.837	1	20416.837	151.600	.000[a]
	Residual	9696.677	72	134.676		
	Total	30113.514	73			

[a] Variables predictoras: (Constante), peso.
[b] Variable dependiente: talla.

La Tabla D7.1 muestra los resultados del ajuste del modelo de regresión (véase apartado 11.4 del Capítulo 11). En el Resumen del modelo encontramos el valor R cuadrado, que corresponde al coeficiente de determinación, que mide la bondad del ajuste de la recta de regresión a la nube de puntos: «el 67.8% de la variación total de la talla se explica por su relación lineal con el peso». En el ANOVA se muestra la descomposición de la variabilidad y se lleva a cabo el contraste de regresión, $H_0 : \beta = 0$, el estadístico del contraste aparece en la columna «F» y vale 151.6, y su valor p asociado, en la columna «Sig.», es $p < 0.0001$, por lo que se rechaza la hipótesis nula y se acepta la alternativa $\mathbf{H_1 : \beta \neq 0}$, es decir, existe una relación lineal significativa entre la talla y el peso.

La Tabla D7.2 muestra en la columna «B» las estimaciones de los dos parámetros de la recta de regresión, la ordenada en el origen α y la pendiente β. La ecuación de la recta ajustada es: **talla = 358.8 + 0.042*peso.**

Asimismo, se presentan los resultados de los dos contrastes bilaterales siguientes: H_0: $\boldsymbol{a} = 0$ y $H_0 : \beta = 0$. El primero de estos contrastes carece de interés en la mayoría de los casos ya que raramente el punto de corte de la recta de regresión con el eje de ordenadas (ordenada en el origen) será el punto (0, 0). Además, dicho punto de corte carece de significado casi siempre. En nuestro caso, la interpretación de $\boldsymbol{a}$ indica que sería la talla media que correspondería a un recién nacido con peso 0 gramos. El segundo contraste, el contraste de regresión, es una alternativa equivalente al contraste que acabamos de comentar en el ANOVA de la Tabla D7.1. El estadístico del test, que aparece en la columna «t» vale 12.313, y su valor p asociado, en la columna «Sig.», es $p < 0.0001$, por lo que se rechaza la hipótesis nula y podemos afirmar que existe una relación lineal significativa entre la talla y el peso. En las últimas columnas de la tabla se proporcionan los intervalos de confianza para α y β, al 95%.

Tabla D7.2. Ecuación **de regresión de la talla sobre el peso**

Coeficientes

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			Intervalo de confianza para B al 95%	
		B	Error típico	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	358.8	**11.061**		32.435	.000	336.702	380.800
	Peso	.042	.003	0.8234	12.313	.000	.035	.049

[a] Variable dependiente: talla.

Análisis de los residuos

El análisis de los residuos es una forma sencilla de contrastar, *a posteriori,* las hipótesis estructurales o supuestos del modelo de regresión lineal (normalidad, igualdad de varianzas de la talla para diferentes niveles del peso y linealidad de las tallas medias para dichos niveles). Estos supuestos resultan necesarios para validar las inferencias respecto a los parámetros. El análisis de los residuos, que siempre es aplicable, resulta especialmente útil cuando disponemos de pocos valores de la variable dependiente para algún/nos valor/es de la variable independiente. En este último caso, no se pueden utilizar los métodos analíticos vistos en los Apéndices D.3 y D.6 para contrastar la normalidad y la igualdad de varianzas, respectivamente.

Cada una de las tallas contenidas en la base de datos proporciona un residuo, que se define como la diferencia entre dicha talla observada y la talla estimada según la recta de regresión.

El gráfico de probabilidad normal de la Figura D7.2 muestra las funciones de distribución teórica y empírica de los residuos tipificados. En el eje de ordenadas se representa la función teórica bajo el supuesto de normalidad; y en el eje de abscisas, la función empírica. La situación de los puntos del gráfico sobre la diagonal del mismo confirma la hipótesis de normalidad.

Para contrastar la homogeneidad de varianzas observamos el gráfico de dispersión de la Figura D7.3, que muestra los residuos tipificados frente a las tallas estimadas tipificadas. Si trazamos una línea

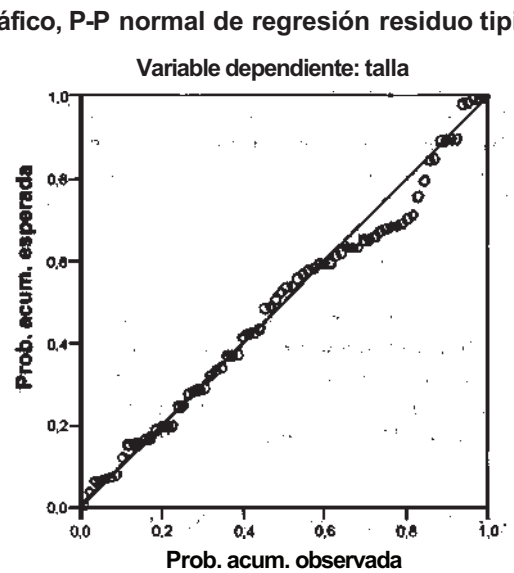


Figura D7.2. Gráfico de probabilidad normal para los residuos tipificados.

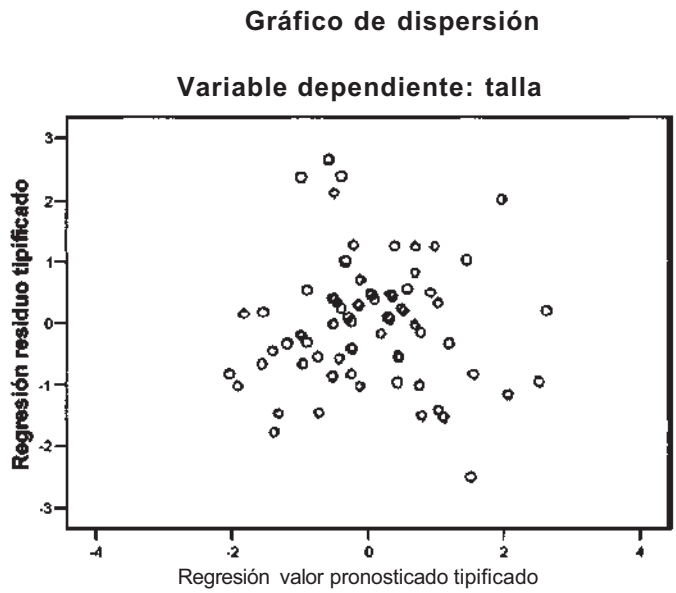


Figura D7.3. Gráfico de residuos tipificados frente a las estimaciones tipificadas.

horizontal a la altura del valor 0, la variación de los residuos sobre esta línea debería ser semejante para, las diferentes tallas, si las varianzas son iguales. En nuestro caso, ocurre esto si exceptuamos cuatro residuos atípicos, que están por encima del valor 2, y un residuo atípico por debajo del - 2 . Este mismo gráfico resulta muy útil para detectar posibles desviaciones de la hipótesis de linealidad. En nuestro caso, la hipótesis de linealidad ya fue confirmada con la nube de puntos de la Figura D7.1.

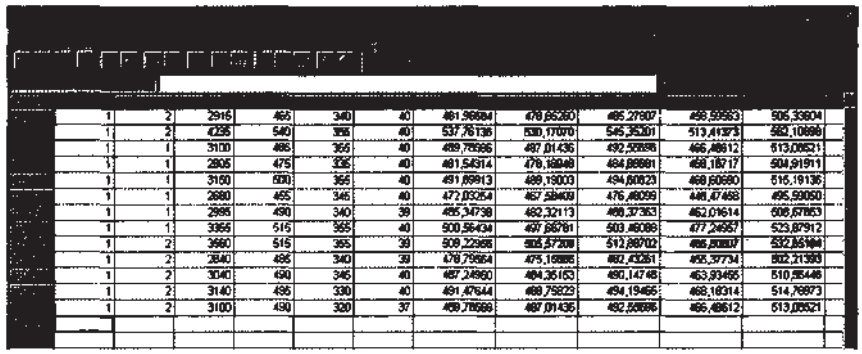
Predicciones

Si volvemos a la ventana del procedimiento **Analizar, Regresión, Lineal,** y pulsamos el botón **Guardar,** podemos pedir los valores predichos de la talla, por el modelo de regresión ajustado, para todos los recién nacidos, así como obtener los intervalos de predicción de la talla individual y de confianza de la talla media (véase apartado 11.5 del Capítulo 11). Para ello, una vez pulsado el botón **Guardar,** hemos de seleccionar en *Valores pronosticados:* **No tipificados,** y en *Intervalos de pronóstico:* **Media, Individuos,** pulsar **Continuar y Aceptar.** Los resultados se guardan en la pantalla de datos, pegados a continuación de éstos. La Figura D7.4 muestra los resultados relativos a los 13 primeros casos. En la columna PRE1 aparecen los valores predichos; en las columnas LMCI y UMCI, los extremos de los intervalos de confianza; y en las columnas LICI y UICI, los extremos de los intervalos de predicción. En la opción *Intervalos de pronóstico* anterior se puede modificar el nivel de confianza de los intervalos.

Si deseamos observar la gráfica de la recta de regresión ajustada, tendremos que elegir el procedimiento **Analizar, Regresión, Estimación Curvilínea.** Una vez emerge la ventana del procedimiento, será necesario seleccionar la variable **talla** y arrastrarla al campo *Dependientes,* seleccionar la variable **peso** y arrastrarla al campo *Independiente variable,* elegir *Modelo:* **Lineal,** marcar **Incluir constante en la ecuación y Representar los modelos;** pulsando **Aceptar,** obtendremos la gráfica deseada.

Comparación del modelo lineal con otros no lineales

En el procedimiento que acabamos de comentar, **Analizar, Regresión, Estimación Curvilínea,** podemos ajustar a nuestros datos hasta 10 modelos de regresión no lineales. Esta opción resulta



	1	2	2915	465	340	40	461,96694	478,65260	495,27907	498,59953	505,33904
	1	2	4295	540	395	40	537,76136	530,17070	545,35201	513,41373	582,10998
	1	1	3100	485	355	40	489,78586	487,01436	492,55696	466,48612	513,09521
	1	1	2805	475	335	40	481,54314	476,18048	484,88881	468,18717	504,91911
	1	1	3150	500	365	40	491,69913	488,19003	494,60023	468,60680	515,19136
	1	1	2680	455	345	40	472,05254	467,58409	476,46099	448,47468	495,59050
	1	1	2995	490	340	39	485,34738	482,32113	488,37363	462,01614	508,67863
	1	1	3355	515	365	40	500,56434	497,66761	503,46089	477,24957	523,87912
	1	2	3660	515	365	39	509,22986	505,57208	512,88702	486,80807	532,65164
	1	2	2840	495	340	39	476,79964	475,15886	482,43261	455,37734	502,21369
	1	2	3040	490	345	40	487,24960	484,35153	490,14748	463,93465	510,56448
	1	2	3140	495	330	40	491,47644	488,75823	494,19466	468,18314	514,76973
	1	2	3100	490	320	37	489,78586	487,01436	492,55696	466,48612	513,09521

Figura D7.4. Valores predichos e intervalos de confianza y predicción en la pantalla de datos.

interesante cuando el ajuste lineal parece inadecuado. Para ello, hemos de marcar los modelos elegidos en *Modelo* y marcar *Mostrar Tabla de ANOVA.* La comparación del ajuste de los diferentes modelos se puede hacer a partir de los valores del coeficiente de determinación de cada uno de ellos. Conviene puntualizar que la elección de un modelo de regresión debe tener en cuenta no sólo la bondad del ajuste numérico sino también la adecuación gráfica de los datos al mismo y, finalmente, su adecuación o explicación biológica.

D7.2. CORRELACIÓN LINEAL

Dado que las dos variables que estamos estudiando son aleatorias, podemos medir la asociación lineal entre ambas mediante el coeficiente de correlación lineal. El valor de este coeficiente se presentó en la Tabla D7.1, en el Resumen del modelo, y es 0.823, lo que indica que existe, entre la talla y el peso, una correlación moderada positiva (véase Fig. 11.14 del Capítulo 11). Podemos obtener simultáneamente los coeficientes de correlación lineal entre todas las variables continuas del fichero de datos. Para ello, debemos elegir el procedimiento **Analizar, Correlaciones, Bivariadas;** en la ventana del procedimiento, seleccionar las variables **peso, talla, pc** y eg, y arrastrarlas al campo *Variables;* elegir en *Coeficientes de correlación:* **Pearson,** y en *Prueba de significación*: **Bilateral**; seleccionar **Marcar las correlaciones significativas,** y pulsar **Aceptar.** Los resultados obtenidos se muestran en la Tabla D7.3.

Tabla D7.3. Matriz de correlaciones y significación de las mismas

Correlaciones

		Peso	Talla	pc	eg
Peso	Correlación de Pearson	**1**	.823**	.638**	.139
	Sig. (bilateral)		.000	.000	.237
	N	**74**	74	74	74
Talla	Correlación de Pearson	.823**	1	.603**	.343**
	Sig. (bilateral)	.000		.000	.003
	N	**74**	74	74	74
pc	Correlación de Pearson	.638**	.603**	1	.275*
	Sig. (bilateral)	.000	.000		.018
	N	**74**	74	74	74
eg	Correlación de Pearson	.139	.343**	.275*	1
	Sig. (bilateral)	.237	.003	.018	
	N	**74**	74	74	74

** La correlación es significativa al nivel 0.01 (bilateral).
* La correlación es significativa al nivel 0.05 (bilateral).

En cada casilla de la Tabla D7.3 figuran tres anotaciones: 1) el coeficiente de correlación entre las variables que definen la casilla, 2) el valor de p del contraste de incorrelación o independencia lineal, $H_0 : \rho = 0$, $H_1 : \rho \neq 0$, y 3) el tamaño muestral. Los coeficientes estadísticamente significativos se destacan con el símbolo * y su interpretación figura en el pie de la tabla.

El contraste de incorrelación es equivalente al contraste de regresión, $H_0 : \beta = 0$, $H_1 : \beta \neq 0$, por lo que su resultado, en el caso de la talla y el peso, ya ha sido comentado en las Tablas D7.1 y D7.2.

D7.3. REGRESIÓN LINEAL MÚLTIPLE

Vamos a continuar con los datos del Ejemplo C6.1. Deseamos estudiar la posible relación lineal entre la talla y el peso, el perímetro craneal y la edad gestacional, contemplando estas tres últimas como regresores o variables predictoras. La talla será la variable respuesta. Para obtener la ecuación de regresión debemos elegir el procedimiento **Analizar, Regresión, Lineal.** Una vez emerge la ventana del procedimiento, hemos de seleccionar la variable **talla** y arrastrarla al campo *Dependiente,* seleccionar las variables **peso, pe y eg,** y arrastrarlas al campo *Independientes,* elegir *Método:* **Pasos suc.,** pulsar el botón **Estadísticos** y elegir en *Coeficientes de regresión:* **Estimaciones;** marcar **Ajuste del modelo y Correlaciones parcial y semiparcial,** pulsar **Continuar y Aceptar.** Las Tablas D7.4, D7.5 y D7.6 muestran la parte sustancial de los resultados.

Tabla D7.4. Ajuste del modelo de regresión múltiple. Coeficiente de determinación

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típico de la estimación
1	.823[a]	.678	.674	11.605
2	.855[b]	.731	.724	10.675

[a] Variables predictoras: (Constante), peso.
[b] Variables predictoras: (Constante), peso, eg.

Para construir el modelo de regresión se ha elegido el método «por pasos». El proceso de selección de variables termina en el paso 2 con la inclusión de las variables peso y edad gestacional. La Tabla D7.4 muestra los coeficientes de determinación corregidos de los dos modelos ajustados, cuando sólo se incluye el peso $R^2 = 0.674$ y cuando se incluyen el peso y la edad gestacional $R^2 = 0.724$, lo que significa que el modelo de regresión obtenido en el paso 2 explica el 72.4% de la variación total de la talla.

Tabla D7.5. Coeficientes del modelo de regresión múltiple

Coeficientes[a]

		Coeficientes no estandarizados		Coeficientes estandarizados			Correlaciones	
Modelo		**B**	**Error típico**	**Beta**	**t**	**Sig.**	**Orden cero**	**Parcial**
1	(Constante)	358.751	11.061		32.435	.000		
	Peso	.042	.003	0.823	12.313	.000	.823	.823
2	(Constante)	222.571	37.672		5.908	.000		
	Peso	.041	.003	.791	12.733	.000	.823	.834
	eg	3.615	.963	.233	3.754	.000	.343	.407

[a] Variable dependiente: talla.

Las Tablas D7.5 y D7.6 contienen los detalles del ajuste de los dos modelos de regresión. La Tabla D7.5 muestra los coeficientes del modelo en la columna «B», y los contrastes acerca de la significación de los coeficientes aparecen en las columnas «t», estadístico del contraste, y «Sig.», valor de p asociado. A partir de dichos valores de p, podemos afirmar que las dos variables predictoras del modelo 2, peso y eg, tienen coeficientes de regresión significativos. Por consiguiente, la ecuación del modelo 2, que es el modelo final, es: **talla = 222.571 + 0.041*peso + 3.615*eg**.

Los coeficientes de correlación lineal simple y parcial entre cada variable predictora y la respuesta figuran en las dos últimas columnas de la Tabla D7.5. El valor del coeficiente de correlación parcial determina la entrada de una variable en el modelo. Esto se aprecia claramente en la Tabla D7.6: si observamos su columna «Correlación parcial», de las dos variables que no entraron en el paso 1, pc y eg, la edad gestacional es la que presenta mayor asociación lineal con la talla, 0.407, una vez eliminado el efecto lineal del peso. Siguiendo con el paso 1, en las columnas «t» y «Sig.» figura, de nuevo, el contraste sobre la significación de la variable eg. Observando la última fila de la tabla, y en particular su columna «Sig», encontramos la razón por la que la variable pc no entra en el modelo (p = 0.461). En la última columna se muestran los valores de la tolerancia de las variables excluidas en cada modelo. Un valor próximo a 0 indicará una fuerte asociación lineal entre la variable excluida y la/s que figuran dentro del modelo, hecho este conocido como el problema de la multicolinealidad, cuyo tratamiento excede al contenido de este libro.

Tabla D7.6. Variables excluidas en el proceso de selección «por pasos»

Variables excluidas[c]

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
						Tolerancia
1	pe	.131[a]	1.518	.133	.177	.593
	eg	.233[a]	3.754	.000	.407	.981
2	pe	.061[b]	.741	.461	.088	.557

[a] Variables predictoras en el modelo: (Constante), peso.
[b] Variables predictoras en el modelo: (Constante), peso, eg.
[c] Variable dependiente: talla.

Contrastes para datos cualitativos

El SPSS evalúa el test ji-cuadrado (también llamado de la chi-cuadrado) de independencia y homogeneidad, visto en el Capítulo 12, en el mismo procedimiento que el de descripción de tablas de contingencia, ya vistas en el apartado D2.4 del Apéndice D2. Deberá ser el investigador el que plantee si su problema es de homogeneidad o de independencia, pues el paquete no lo distingue.

El SPSS permite analizar estos problemas cuando los datos vienen dados en dos formas: fichero de datos y datos resumidos en una tabla de contingencia.

D8.1. FICHEROS DE DATOS

Consideremos los datos del Ejemplo C2.1 del Apéndice C2. Supongamos que queremos ver si están relacionadas la variables recidiva del tumor y sexo. Las hipótesis a contrastar son H_0: las dos variables son independientes, frente a H_1: las dos variables están relacionadas.

Elegimos el procedimiento **Analizar, Estadísticos descriptivos, Tablas de contingencia,** e indicamos qué variable queremos que esté en las filas (p. ej., sexo) y cuál en las columnas (p. ej., **recidiva).** Al pulsar el botón **Estadísticos** aparece una ventana con distintos nombres de test, y debemos marcar *Chi-cuadrado.* Si pulsamos el botón **Casillas,** podremos pedir porcentajes y frecuencias esperadas (véase Fig. D8.1). Al pulsar **Aceptar,** se obtienen los resultados de la

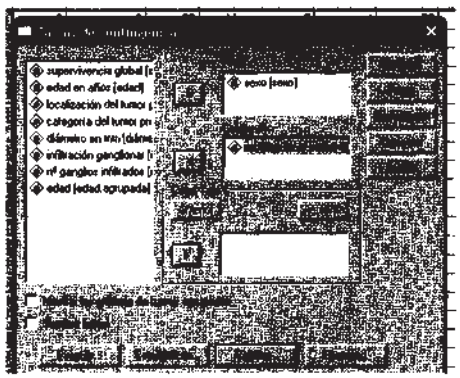


Figura D8.1. Opciones de una tabla de frecuencias.

Tabla D8.1, donde vemos que el estadístico es 0.166 y su valor de p (Sig. asintótica bilateral) es 0.684, luego no se rechaza la hipótesis de independencia.

La distribución del estadístico chi-cuadrado es asintótica. Es por ello que en el caso de las tablas 2x2, algunos autores creen conveniente la corrección de dicho estadístico. El nombre que recibe dicha corrección es de Yates y es calculada por defecto en el SPSS, a la vez que el test de la chi-cuadrado (véase Tabla D8.1), con el nombre de corrección por continuidad. La utilización o no de dicha corrección es tema de controversia actual entre los estadísticos.

En el Capítulo 12, al final del apartado prueba de independencia se comenta que un test alternativo para muestras pequeñas, en tablas 2 x 2, es el test de Fisher, que se ha calculado por defecto en este caso.

Tabla D8.1. Resultados del test de la chi-cuadrado, muestras pequeñas

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	**.166**[b]	**1**	**.684**		
Corrección por continuidad[a]	**.016**	**1**	.899		
Estadístico exacto de Fisher				.784	.784
N de casos válidos	**88**				

[a] Calculado sólo para una tabla de 2 x 2.
[b] 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6.73.

D8.2. DATOS RESUMIDOS EN TABLA DE CONTINGENCIA

Supongamos que los datos que tenemos de entrada son la propia tabla de contingencia. Consideremos el Ejemplo 12.1.1 del Capítulo 12, donde se consideran 549 personas que son vacunadas y 534 que no son vacunadas. Después, de los no vacunados 70 contraen hepatitis y de los vacunados, 11. Se quiere ver la eficacia de la vacuna frente a la hepatitis. Las hipótesis a contrastar son H_0: las muestras son homogéneas (el porcentaje de personas con hepatitis es igual entre los vacunados que entre los no vacunados), frente a H_1: las muestras no son homogéneas.

El primer paso será introducir los datos de las 1083 (549 + 534) personas que entran en el estudio, sin tener que teclear uno a uno los 1083 casos. Utilizaremos el procedimiento de ponderar casos. Para ello, tenemos que escribir en una columna las frecuencias que llamaremos **Frecuencia** y en otras dos columnas las variables **Vacuna** (SÍ = 1, NO = 2) y **Hepatitis** (SÍ = 1, NO = 2), como en la Figura D8.2. Elegimos el procedimiento **Datos, Ponderar casos;** aparece una ventana como la de la Figura D8.2 donde indicaremos cuál es la variable que contiene las frecuencias. Una vez que tenemos leídos los datos, podemos formar la tabla y el test como en el apartado anterior: elegi-

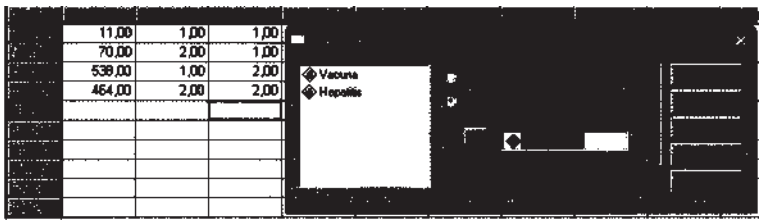


Figura D8.2. Entrada de datos para tablas de contingencia

mos el procedimiento **Analizar, Estadísticos descriptivos, Tablas de contingencia,** indicaremos qué variable queremos que esté en las filas (p. ej., **Vacuna)** y cuál en las columnas (p. ej., **Hepatitis).** Al pulsar el botón **Estadísticos,** aparece una ventana con distintos nombres de test, y debemos marcar *Chi-cuadrado.*

Al ejecutar el procedimiento se obtienen el valor del estadístico que es 48.242 y su valor de p es 0.000 (véase Tabla D8.2), con lo que se rechazaría la hipótesis nula de homogeneidad, es decir, concluimos que hay diferencia significativa en el porcentaje de hepatitis entre los vacunados y no vacunados.

Tabla D8.2. Resultados del test de la chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	48.242[b]	1	.000
Corrección por continuidad[a]	46.650	1	**.000**
N de casos válidos	1083		

[a] Calculado sólo para una tabla de 2 x 2.
[b] 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuenci mínima esperada es 39.94.

NOTA

Hay que señalar que el test de homogeneidad de muestras que acabamos de ver, en el caso de una tabla 2×2, coincide con el test bilateral de comparación de 2 proporciones independientes presentado en el apartado D5.3 del Apéndice D5.

Contrastes no paramétricos

Cuando no se puede suponer la normalidad, no se deben aplicar los test vistos en los Apéndices D5 y D6. En este Apéndice veremos las alternativas no paramétricas de dichos test, que sirven para comparar una variable continua entre 2 y k poblaciones. Veremos el caso de muestras tanto independientes como relacionadas. Estos test se corresponden con los descritos en el Capítulo 13.

D9.1. DOS POBLACIONES

Muestras independientes

En apartado 13.4 del Capítulo 13 se explica el test denominado «suma de los rangos de Wilcoxon», que sería la alternativa no paramétrica de los test que, basados en la t de Student, sirven para comparar dos poblaciones independientes. En dicho apartado se comenta que otra alternativa es el test conocido como U de Mann-Whitney, pero que dichos test son equivalentes. El SPSS calcula dos estadísticos que llama W de Wilcoxon y U de Mann-Whitney, pero al ser equivalentes da un único valor de p. Además, a la hora de calcular dicho valor de p aplica una aproximación a la distribución normal, la cual sólo es válida para muestras grandes. Veamos cómo proceder en un ejemplo concreto.

Consideremos que queremos estudiar, para los pacientes con cáncer del Ejemplo C2.1 del Apéndice C2, si la supervivencia global es igual entre hombres y mujeres. Primero debemos contrastar la normalidad de la supervivencia global en el grupo de hombres y en el de mujeres. Como ya se vio en el apartado D3.4 (Apéndice D3), esta hipótesis no puede admitirse. Una vez constatada la ausencia de normalidad, hacemos un contraste no paramétrico para comparar las medianas de la supervivencia de hombres y mujeres. Elegimos el procedimiento **Analizar, Pruebas no paramétricas, 2 muestras independientes;** emerge la ventana de la izquierda de la Figura D9.1, donde en el campo *Contrastar variables* indicamos **supervivencia global,** y en el campo *Variable de agrupación* indicamos sexo. Debemos ahora pulsar el botón **Definir grupos** y emerge la ventana de la derecha de la Figura D9.1, que debemos rellenar de la siguiente forma: en los campos *Grupol y Grupo 2* pondremos los valores con los que están codificados varón (con 1) y mujer (con 2), respectivamente.

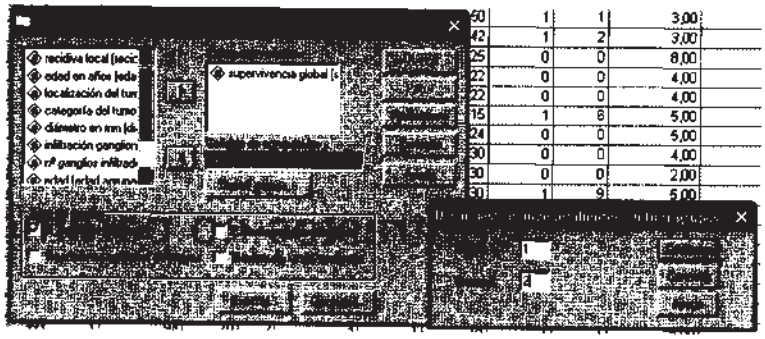


Figura D9.1. Instrucciones para el test de Wilcoxon-Mann-Whitney.

Al pulsar **Continuar,** está marcado por defecto el test U de Mann-Whitney, así que al pulsar **Aceptar** se obtienen las Tablas D9.1 y D9.2. Como el valor de p es 0.733, concluimos que no existe diferencia significativa entre las medianas del tiempo de supervivencia de hombres y mujeres.

Tabla D9.1. Descripción de rangos

	Sexo	N	Rango promedio	Suma de rangos
Supervivencia global	Varón	72	44.06	3172.50
	Mujer	16	46.47	743.50
	Total	88		

Tabla D9.2. Test de Wilcoxon-Mann-Whitney

Estadísticos de contraste*

	U de Mann-Whitney	W de Wilcoxon	Z	Sig. asintót. (bilateral)
Supervivencia global	544.500	3172.500	-.341	.733

* Variable de agrupación: sexo.

Muestras apareadas

En el apartado 13.3 del Capítulo 13, se dan dos test para el caso de muestras relacionadas. El test que se denomina «contraste de los signos para la mediana de las diferencias» lo realiza el SPSS con el nombre «Signos». El test que se denomina «Contraste de los rangos de signos de Wilcoxon: datos emparejados», lo realiza el SPSS con el nombre «Wilcoxon», pero en una versión aproximada a la normal, sólo válida para muestras grandes.

Resolvamos el Ejercicio 13.3.3 del Capítulo 13, donde se quiere estudiar si la tensión arterial disminuye después de intervenir para eliminar una obstrucción renal en enfermos con función renal deteriorada. Para ello, se toma la tensión antes y después de la operación a 10 pacientes. Dado que tenemos sólo 10 datos, debemos utilizar el test de signos, pues el de Wilcoxon es una aproximación. La hipótesis nula del contraste es que la mediana de la diferencia de tensiones (antes-después) es 0 frente a la alternativa que postula que dicha mediana es > 0.

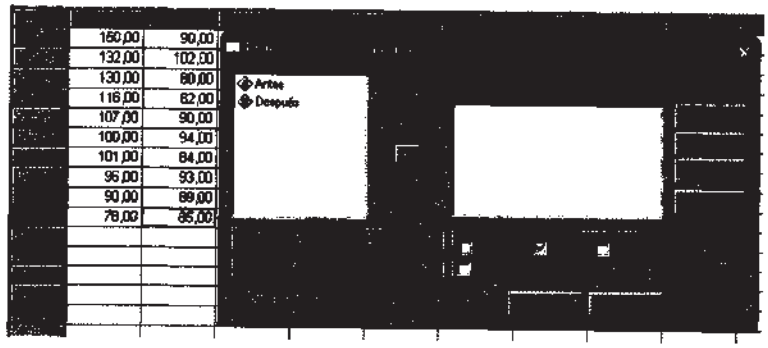


Figura D9.2. Datos e instrucciones para muestras apareadas.

Los datos deben introducirse según la Figura D9.2; después elegiremos el procedimiento **Analizar, Pruebas no paramétrícas, 2 muestras relacionadas,** y emerge una ventana como la de la Figura D9.2, que debemos rellenar como allí se indica. Al pulsar **Aceptar,** se obtienen las Tablas D9.3 y D9.4. En la Tabla D9.3 se presenta información sobre las frecuencias de los signos de la variable diferencia.

Tabla D9.3. Descripción de la variable diferencia

		N
Después-Antes	Diferencias negativas[a]	9
	Diferencias positivas[b]	1
	Empates[c]	0
	Total	10

[a] Después < Antes
[b] Después > Antes
[c] Después = Antes

Tabla D9.4. Test de los signos

Estadísticos de contraste[b]

	Después-Antes
Sig. exacta (bilateral)	.021[a]

[a] Se ha usado la distribución binomial.
[b] Prueba de los signos.

Si observamos la Tabla D9.4, vemos que el valor de p es 0.021, resultado correspondiente al test bilateral. Como nuestro test es unilateral, deberíamos dividir esta probabilidad por dos, es decir p = 0.0105, por lo que se aceptaría la eficacia de la operación para a = 0.05.

El valor de p se puede obtener de la Tabla II del Apéndice B, eligiendo X = B(10, 0.5) y calculando $P(X \leq 1) = 0.0107$.

D9.2. K POBLACIONES

Muestras independientes

En el apartado 13.5 del Capítulo 13, se explica el test de Kruskal-Wallis como alternativa al ANOVA de una vía. Dicho test lo realiza el SPSS.

Consideremos el Ejemplo C2.1 del Apéndice C2, y supongamos que queremos ver si la supervivencia global varía con la localización del tumor. El primer paso sería contrastar la normalidad de la variable **supervivencia global** en cada una de las localizaciones (como se vio en el apartado D3.4, del Apéndice D4), pero aquí hay localizaciones con sólo 8 datos, con lo que no se podría comprobar la normalidad con un test. Apliquemos un contraste no paramétrico.

Elegimos el procedimiento **Analizar, Pruebas no paramétricas, k muestras independientes,** y emerge la ventana de la izquierda de la Figura D9.3, que debemos completar: en el campo *Contrastar variables* ponemos **supervivencia global,** en el campo *Variable de agrupación* ponemos **localización** y al pulsar el botón **Definir Rango,** emerge la ventana de la derecha, en la que debemos indicar el valor *Máximo* (5) y *Mínimo* (1) de localización. Como por defecto está marcado el test de Kruskal-Wallis, al pulsar **Aceptar** se obtienen los resultados de las Tablas D9.5 y D9.6.

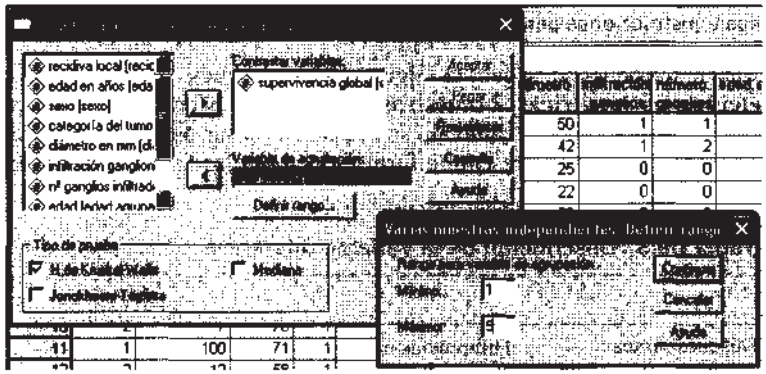


Figura D9.3. Instrucciones para el test de Kruskal-Wallis.

Observando la Tabla D9.6, vemos que el valor de p es 0.002, con lo que se acepta que existen diferencias significativas entre las medianas del tiempo de supervivencia de sendas localizaciones, para cualquier $\boldsymbol{a} > 0.002$. Aunque el SPSS no hace test *a posteriori,* si nos fijamos en la columna Rango promedio de la Tabla D9.5, podremos ver qué localizaciones difieren más.

Tabla D9.5. Descripción de rangos por grupos

	Localización del tumor primario	N	Rango promedio
Supervivencia global	Labio	12	54.92
	Mucosa	8	18.75
	Lengua	36	53.35
	Suelo de boca	24	38.58
	Orofaringe	88	32.56
	Total		

Tabla D9.6. Test de Kruskal-Wallis

Estadísticos de contraste[a,b]

	Chi-cuadrado	gl	Sig. asintót.
Supervivencia global	17.485	4	**.002**

[a] Prueba de Kruskal-Wallis.
[b] Variable de agrupación: localización del tumor primario.

Muestras apareadas

En el apartado 13.6 del Capítulo 13, se explica el test de Friedman como alternativa no paramétrica cuando tenemos k muestras relacionadas. Dicho test sigue asintóticamente una distribución chi-cuadrado y esta versión es la que realiza el SPSS.

Resolvamos el Ejercicio 13.6.2 del Capítulo 13. Se toman 12 truchas sometidas a dosis subletales de metilo de mercurio; se quiere ver si existen diferencias significativas entre las medianas de las concentraciones de mercurio en tres localizaciones del cuerpo. Para ello se estudia dicha concentración en tres localizaciones (encéfalo, musculatura y ojo) de cada trucha.

Los datos deben introducirse en la pantalla de datos según la Figura D9.4. Elegimos el procedimiento **Analizar, Pruebas no paramétricas, k muestras relacionadas,** y emerge una ventana que debemos completar como se indica en la Figura D9.4. Por defecto está marcado el test de Friedman, así que al pulsar **Aceptar,** se obtiene la pantalla de resultados de la Tabla D9.7.

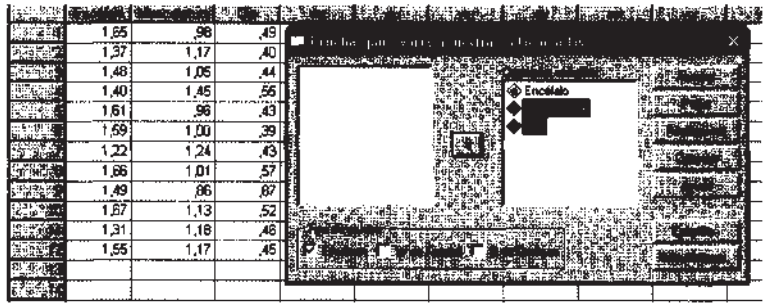


Figura D9.4. Entrada de datos e instrucciones para el test de Friedman.

Observamos en la Tabla D9.7 que el valor de p es 0.000, de lo que concluimos que existen diferencias significativas entre las concentraciones medianas de mercurio en al menos una de las localizaciones. A pesar de que el SPSS no realizar ningún contraste *a posteriori,* vemos que el Rango promedio para ojos es inferior al de las otras dos localizaciones.

Tabla D9.7. Resultados del test de Friedman

Rangos

	Encéfalo	**Musculatura**	**Ojo**
Rango promedio	2.83	2.08	1.08

Estadísticos de contraste[a]

N	**Chi-cuadrado**	**gl**	**Sig. asintót.**
12	18.500	2	.000

[a] Prueba de Friedman.

REFERENCIAS

1. Beyer, William, ed.: *Handbook of Tables for Probability and Statistics,* 2.ª ed., CRC Press, Boca Ratón, Fla., 1968.
2. Bradley, James V.: *Distribution Free Statistical Tests,* Prentice-Hall, Englewood Cliffs, N. J., 1968.
3. Cohover, W. J.: *Practical Non-parametric Statistics,* Wiley, New York, 1971.
4. Daniel, W.: *Applied Nonparametric Statistics,* PWS-Kent, Boston, 1990.
5. Hoaglin, D., F. Mosteller, and J. Tukey: *Understanding Robust and Exploratory Data Analysis,* Wiley, New York, 1983.
6. Iman, R.: «Graphs for Use with the Lilliefors Test for Normal and Exponencial Distributions», *The American Statistician,* vol. 36, núm. 2, 1982.
7. Koopmans, Lambert: *An Introduction to Contemporary Statistics,* PWS-Kent, Boston, 1985.
8. Larson, Harold: *Introduction to Probability Theory and Statistical Inference,* Wiley, New York, 1982.
9. Leontnor, M., and T. Bishop: *Experimental Design and Analysis,* Valley Book Company, Blacksburg, Va., 1986.
10. Lentner, M., J. Arnold, and K. Hinkelmann: «How to Use the Ratio of MS (Blocks) and MS (Error) Correctly», *American Statistician,* VOL. 43, núm. 2, 1989, págs. 100-108.
11. Milton, J. S., and J. Arnold: *Introduction to Probability and Statistics: Principies and Applications for Engineering and the Computing Sciences,* McGraw-Hill, New York, 1990.
12. Myers, R.: *Classical andModern Regression with Applications,* PWS-Kent, Boston, 1990.
13. Myers, R., and J. S. Milton: *A First Course in the Theory of Linear Statistical Models,* PWS-Kent, Boston, 1991.
14. *SAS User's Guide,* Version 5, SAS Institute Inc., Raleigh, N. C, 1985.
15. Snedecor, G. W., and William Cochran: *Statistical Methods,* 6.ª ed., Iowa State University Press, Ames, 1967.
16. Tukey, John: *Exploratory Data Analysis,* Addison-Wesley, Reading, Mass., 1977.

Respuestas a problemas impares sueltos

EJERCICIOS 1.1

1. *a)*

Categoría	Frecuencia	Frecuencia relativa	Porcentaje
RM	3	3/24 = 0.1250	12.5
EM	7	7/24 = 0.2917	29.17
FE	14	14/24 = 0.5833	58.33

3. **Diagnóstico**

Sexo	RM	EM	FE	Distribución por sexo
M	2 0.0833 8.335%	3 0.125 12.5%	4 0.1667 16.67%	9 0.375 37.5%
F	1 0.0417 4.17%	4 0.1667 16.67%	10 0.4167 41.67%	15 0.625 62.5%
Distribución por diagnóstico	3 0.125 12.5%	7 0.2917 29.17%	14 0.5833 58.33%	24

5. a) **Dolor de cabeza**

Sexo	V	T	C	Distribución por sexo
M	17	16	14	47
	0.0960	0.0904	0.0791	0.2655
	9.60%	9.04%	7.9%	26.55%
F	66	30	34	130
	0.3729	0.1695	0.1921	0.7345
	37.29%	16.95%	19.21%	73.45%
Distribución por diagnóstico	83	46	48	177
	0.4689	0.2599	0.2712	
	46.89%	25.99%	27.12%	

Las frecuencias por celda están divididas por 177.

b)

Sexo	V	T	C		
M	17	16	14	‖	
	0.3617	0.3404	0.2979	‖	47
	36.17%	34.04%	29.79%	‖	
F	66	30	34	‖	130
	0.5077	0.2308	0.2615	‖	
	50.77%	23.08%	26.15%	‖	

Las frecuencias por celda para los varones están divididas por 47; las de las mujeres por 130.

7. a) la especie 2; el 46.15 % de estas especies participó en la actividad, frente al 38.469b de otras especies.
   b) 42.31%
   c) sí; los porcentajes del apartado *α)* parecen ser bastante diferentes, lo que indica que la especie 2 parece participar en el cortejo substancialmente más que la especie 1

EJERCICIOS 1.2

1. a*)*

```
4|05
5|123
6|122678
7|011189
8|124
9|018
10|0
11|1
```

b) Los datos sugieren forma de campana.

c) sí; casi todos los datos están por debajo del rango «normal» de 1.0 a 2.9

3. *a)*
```
12|12
12|5678
13|0123344
13|5678
14|01
14|6
```
b) no
c) la media está en algún sitio del «tallo» 13; quizás alrededor de 13.3; no, todas estas lecturas indican una protección por encima del nivel de protección estándar de 1 mg/mL

5. *a)*

Acústico	Visual
081	07|01123
08|6	07|578
09|	08|01
09|9	08|9
10|0123	09|0
11|34	09|579
11|57	10|012334
12|0	10|
12|6	

b) no; ambos diagramas parecen ser simétricos
c) sí; la latencia para el estímulo visual parece tener forma de U
d) visual; hay más puntos alejados del centro

7. a)

Geriátrico ll	Geriátrico 2
1 |88	
2|035899	
3|0114555569	3|6
4|0112227	4|89
5|011269	5|45677
6|1245	612334579
7|126	7|128
8|79	8|01339
	9|2

La forma de ambas parece similar, no así la posición. Los pacientes del «geriátrico» 2 tienden en promedio a ser más viejos que los del 1.

9. c)

Estéril	No estéril
0|	0|23
0|99	0|6679
1|0001	1|11344
1|578	115689
2|	2|0
2|5688	2|
3|004	
3|5	

«No estéril» tiene forma aproximada de campana. «Estéril» parece estar bastante dispersa. «No estéril» parece tener un menor centro de posición.

Estéril	No estéril
	0|23
99	0|6679
1000	1|11344
875	1|5689
	2|0
8865	21
400	3|
5	3|

Los rábanos parecen crecer mejor en suelo estéril. Sin embargo, su crecimiento parece ser más consistente en el suelo no estéril.

EJERCICIOS 1.3

1. **a)** 12 130; 737 ***b)*** 11393 c) 6 *d)* 1898.833 *e)* 1899 *f)* 736.5

g)

Clase	Límites	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	736.5 a 2635.5	8	0.178	8	0.178
2	2635.5 a 4534.5	9	0.200	17	0.378
3	4534.5 a 6433.5	9	0.200	26	0.578
4	6433.5 a 8332.5	12	0.267	38	0.845
5	8332.5 a 10231.5	5	0.111	43	0.956
6	10231.5 a 12130.5	2	0.044	45	1.000

3. **a)**

Clase	Límites	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	1499.5 a 2320.5	9	0.375	9	0.375
2	2320.5 a 3141.5	1	0.042	10	0.417
3	3141.5 a 3962.5	10	0.417	20	0.834
4	3962.5 a 4783.5	0	0.000	20	0.834
5	4783.5 a 5604.5	4	0.167	24	1.000

5. 0.616

7. ***a)***
0|578
1|01223455677888999
2|00123333444556778
3|0011245678
4|05
5|0

b) sí; la forma de la distribución indica asimetría a derecha

c)

Clase	Límites	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	0.45 a 1.25	7	0.14	7	0.14
2	1.25 a 2.05	15	0.30	27	0.44
3	2.05 a 2.85	15	0.30	37	0.74
4	2.85 a 3.65	8	0.16	45	0.90
5	3.65 a 4.45	3	0.06	48	0.96
6	4.45 a 5.25	2	0.04	50	1.00

9. a) no; el algoritmo del texto utiliza 6 clases para los datos de los sitios incendiados, mientras que SAS utiliza 7
   b) 0.4
   c) 0.4; no; sí, un futuro dato de 0.400 podría caer en este límite
   d) incendiados: 0.1085 a 0.5125; no incendiados: 0.0095 a 0.3755; no

11.

Experimental	Control
	|08|
	|08|8
	|09|0
	|09|68
	|10|012
	|10|5699
	|11|012222
	|11|556789
	|12|012334
	|12|8
0	|13|
98887765	|13|
44322	|14|
9877655	|14|
4211110	|15|
65	|15|

Hay una diferencia mayor en la posición. El diagrama de tallos y hojas doble adosado da una mejor perspectiva de la distribución que los histogramas por separado.

EJERCICIOS 1.4

1. conjunto I: $\bar{x} = 2.1$, $\tilde{x} = 2$
   conjunto II: $\bar{x} = 3.5$, $\tilde{x} = 3.5$
3. a) petirrojos: $\bar{x} = 100.48$, $\tilde{x} = 75.0$
   palomas: $\bar{x} = 224.66$, $\tilde{x} = 170$
   b) $\bar{x} = 184.02$, $\tilde{x} = 168.35$
   La mediana se ve menos afectada por la presencia de un dato atípico.
5. $\bar{x} = 0.72$, $\tilde{x} = 0.71$
7. $\bar{x} = 1.1$

EJERCICIOS 1.5

1. conjunto I: rango = 4, $s = 1.5$, $s^2 = 2.11$
   conjunto II: rango = 6, $s = 2.2$, $s^2 = 4.94$
3. *a)* petirrojos: $s = 62.54$, $s^2 = 3911.446$ *b)* $s = 119.62$; no
   palomas: $s = 234.52$, $s^2 = 55\,000.499$ $s^2 = 14\,308.124$
   c) petirrojos: 90.25; palomas: 220.6 *d)* 216.35; sí
   e) petirrojos: 255.0; palomas 1186.1 *f)* 367.8; no
5. $s = 0.98$, $s^2 = 0.954$, rango = 4.5, iqr = 1.3

7. a) 0|1
0|9
1|0234
1|55789
2|001234
2|5

b) i o I: $\bar{x} = 20.0$, $\tilde{x} = 19.0$
sitio II: $\bar{x} = 16.6$, $\tilde{x} = 17.5$

c) sitio I: $s = 10.1$, $s^2 = 102.13$
sitio II: $s = 6.1$, $s^2 = 37.44$
$\tilde{x} = 0.71$, $s^2 = 0.1956$, iqr = 0.22

9. c) Cualquier número que sea mayor que el más grande de los valores dados o más pequeño que el menor de todos ellos cambiará el rango pero tendrá efectos mínimos sobre la mediana.
b) Cualquier valor cercano a la media.

EJERCICIOS 1.6

1. a) 0|2
0|67889
1|00111122334
1|556
2|0

b) $\bar{x} = 11$, $q_1 = 9$, $q_3 = 13$, iqr = 4, $f_1 = 3$, $f_3 = 19$.
$F_1 = -3$, $F_3 = 25$, $a_1 = 6$, $a_3 = 16$
c) Los valores 2 y 20 son ligeramente atípicos.

3. $\tilde{x} = 2.72$, $q_1 = 1.32$, $q_3 = 5.25$, iqr = 3.93, $f_1 = 4.57$, $f_3 = 11.145$; no hay datos atípicos
5. *a)* A: $\tilde{x} = 130$, $q_1 = 119.5$, $q_3 = 139.0$, iqr = 19.5, $f_1 = 90.25$, $f_3 = 168.25$, $F_1 = 61.0$, $F_3 = 197.5$
atípicos: 12.5; 170 200
$a_1 = 105$, $a_3 = 156$
R: $\tilde{x} = 253$, $q_1 = 238$, $q_3 = 271$, iqr = 33, $f_1 = 188.5$, $f_3 = 320.5$, $F_1 = 139.0$, $F_3 = 370.0$
atípico: 561
$a_1 = 221$, $a_3 = 283$
b) $\bar{x} = 128.07$, $s^2 = 1282.681$ *c)* $\bar{x} = 133.95$, $s^2 = 503.83$
d) 516 es un dato atípico; no
7. a) 1: 7.7; 2: 5.5 *b)* grupo 2 c) grupo 1
d) grupo 1; no, puede ser un error al introducir los datos o una mala toma en el experimento
e) sí

EJERCICIOS 1.7

1. *a)*

Clave	Punto medio	Frecuencia acumulada
1	7	1
2	12	3
3	17	8
4	22	11

b) $\bar{x} \cong 16.5$, $s^2 \cong 22.27$, $s \cong 4.7$, $\tilde{x} \cong 17.5$

3. a) **Varones** **Mujeres**

Clave	Punto medio	Frecuencia acumulada	Clave	Punto medio	Frecuencia acumulada
1	7	1	1	7	0
2	12	5	2	12	2
3	17	12	3	17	12
4	22	35	4	22	19
5	27	51	5	27	22
6	32	58	6	32	27
7	37	68	7	37	29

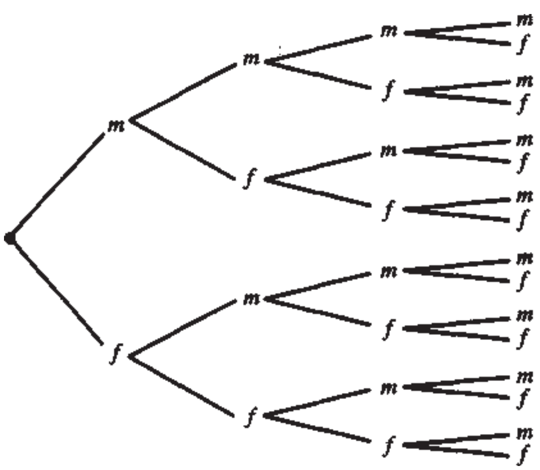
b) varones: $\bar{x} \cong 25.1$, $s^2 = 52.26$, $s \cong 7.2$, $\tilde{x} \cong 24.4$
mujeres: $\bar{x} = 22.9$, $s^2 \cong 51.91$, $s \cong 7.2$, $\tilde{x} \cong 21.6$

EJERCICIOS 2.1

1. personal o frecuencia relativa. 3. frecuencia relativa; $\frac{5}{150}$ 5. clásico; $\frac{1}{4}$
7. frecuencia relativa; $\frac{8}{100}$ 9. clásico; $\frac{1}{5}$
11. a) 0.800
b) El 5 % de las veces hay reacción en los primeros 5.9 minutos y el 95% no la hay; por tanto $p \cong 0.95$
c) 0.15
13. alto; debido a las graves consecuencias del pinchazo, uno en veinte posibilidades parece algo elevado.

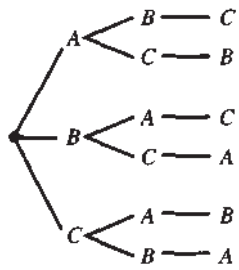
EJERCICIOS 2.2

1.



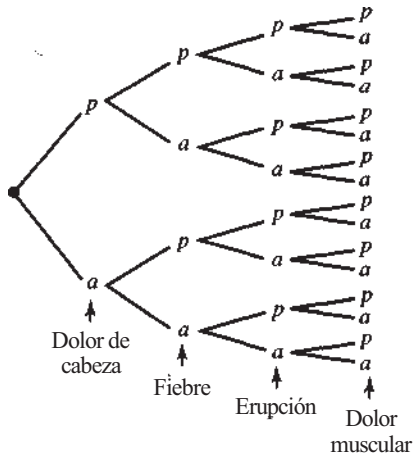
$\frac{1}{2}$; $\frac{6}{16}$; $\frac{4}{16}$; 0

3. a)



b) $\frac{1}{3}$

5. *a)*



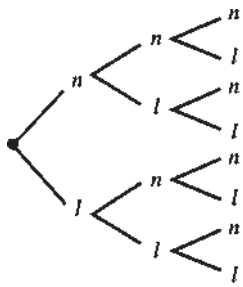
b) pppp, pppa, papp, papa, appp, appa, aapp, aapa

c) pppp, ppap, appp, apap

d) pppp, appp

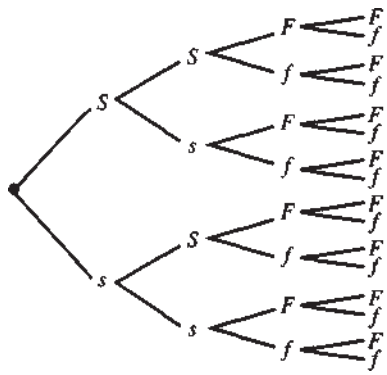
e) ppaa, paap, paaa, apaa, aaap, aaaa

7. *a)*



b) Las trayectorias son igualmente probables si en cada test el recuento celular norma) es igualmente probable que el bajo; no

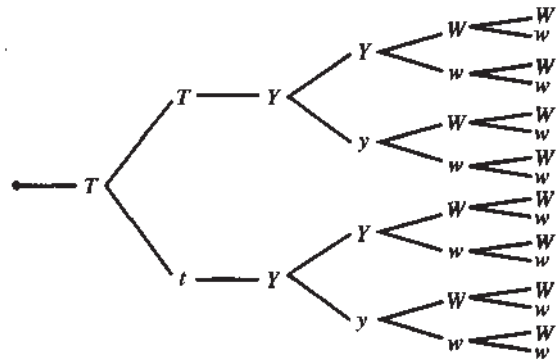
9. *a)*



b) **(i)** $\frac{1}{16}$ **(ii)** $\frac{9}{16}$ **(iii)** $\frac{4}{16}$ **(iv)** $\frac{12}{16}$

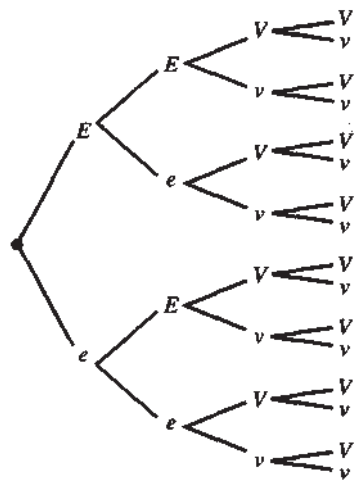
11. *a)* alta, amarilla, redonda

b)



c) cada una es alta y amarilla; algunas son arrugadas y otras redondas
d) 1 e) $\frac{4}{16}$ ***f) 0***

13. a)



b) 4 c) 12 d) $\frac{12}{16}, \frac{1}{6}, \frac{15}{16}$

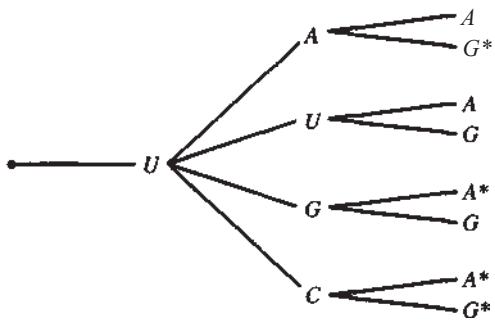
EJERCICIOS 2.3

1. permutación
3. combinación
5. normal-alta-alta-normal-baja; nnhhl; permutación
7. tutu = marcado-no marcado-marcado-no marcado
tuut = marcado-no marcado-no marcado-marcado
9. combinación; pppwwwwp; pwpwpwwp; permutación

EJERCICIOS 2.4

1. $20 \cdot 15 = 300$ *3.* $5 \cdot 24 \cdot 23 \cdot 22 \cdot 21 = 1\,275\,120$
5. α) 1-4-2 = 8 b) 1-2-2 = 4 c) $\frac{4}{64}$

d)



7. 1 1 - 4 = 4
9. 64 • 64 • 64 = 262 144; 64 • 63 • 62 = 249 984; 262 144 - 249 984 = 12 160 (número de palabras con repetición); P[repetición] = $\frac{1212\ 160}{262\ 144}$

   61 • 60 • 59 = 215 940 = número de palabras sin repetición y que califican un aminoácido;

   P[no repetición y codifican un aminoácido] = $\frac{215\ 940}{262\ 144}$
11. *a)* 7! = 5040
    b) 2 • 1 • 5! = 240 formas de colocar AyB en posiciones uno y dos; *P[A* y *B* en posiciones uno y dos] = $\frac{240}{5040}$ = 0.0476
    c) sí; la probabilidad de que esto ocurra por casualidad es pequeña (0.0476)
13. *a)* 16 *b)* 5
15. *a)* *n,n,-l,n- 2, ..., n - r + 1* *c)* ${}_{10}P_4 = 10 \cdot 9 \cdot 8 \cdot 7$ *d)* ${}_nP_n = n!$
    e) ${}_{10}P_3 = 10 \cdot 9 \cdot 8 = 720$

EJERCICIOS 2.5

1. $\frac{3!}{2!\ 1!} = 3$ 3. $\frac{14!}{4!\ 10!} = 1001$
5. $\frac{8!}{4!\ 2!\ 2!} = 420$ (los períodos de descanso se alternan con los de actividad y la observación después del período de descanso)

   $\frac{7!}{4!\ 2!\ 1!} = 105$ formas de tener períodos de larga actividad consecutivos; $\frac{105}{420}$
7. $\frac{10!}{3!\ 2!\ 4!\ 1!} = 12\ 600$; $\frac{9!}{3!\ 2!\ 4!} = 1260$ formas de tener parada y final; 12 600 - 1260 = = 11 340 formas de tener parada en otro sitio; P[parada no final] = $\frac{11\ 340}{12\ 600}$

EJERCICIOS 2.6

1. a) 15 *b)* 56 *c)* 1 *d)* 1

3.: $\dfrac{\binom{7}{7}\binom{8}{0}}{\binom{15}{7}} = \dfrac{1}{6435}$; $\dfrac{\binom{7}{5}\binom{8}{2}}{\binom{15}{7}} = \dfrac{588}{6435}$ 5. $\dfrac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}} = \dfrac{36}{120}$

7. $\dfrac{\binom{5}{0}\binom{95}{15}}{\binom{100}{15}} = 0.4357$, $\dfrac{\binom{5}{0}\binom{95}{14}}{\binom{100}{15}} = 0.4034$

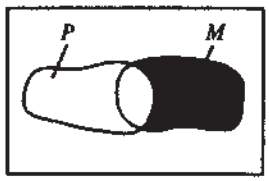
P[al menos 1] = 0.4357 + 0.4034 = 0.8391

9. $\binom{50}{15} \cong 2\,250\,000\,000\,000$

EJERCICIOS 3.1

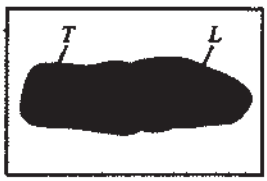
1. a) tiene leucemia y el recuento de leucocitos alto
   b) tiene leucemia o el recuento de leucocitos alto
   c) tiene leucemia pero el recuento de leucocitos no es alto
   d) no tiene leucemia y el recuento de leucocitos no es alto
3. *a)* El niño ha recibido ambas vacunas

b)

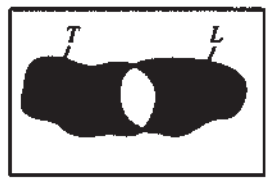


c)

d)



e)



5. b, c, d, f, g, h 7. 0.65 9. 0.10
11. c) 0.27 *b)* 0.67 *c)* 0.18 *d)* 0.82
13. $S \circ \varnothing = S$. Por tanto, $P[S \circ \varnothing] = P[S]$. Ya que S y $\varnothing$ son mutuamente excluyentes, $P[S] = P[S \circ \varnothing] = P[S] + P[\varnothing]$ por el Axioma 3. Por el Axioma 1, P[S] = 1. Entonces

$$1 = 1 + P[\varnothing]$$

Despejando $P[\varnothing]$ se obtiene $P[\varnothing] = 0$.

15. *a)* $B = A \circ (B$ pero no $A)$. Ya que A y B pero no A son mutuamente excluyentes, $P[B] = P[A] + P[B$ pero no $A]$ por el Axioma 3. Por el Axioma 2

$$P[B \text{ pero no A}] \geq 0$$

Sumando $P[A]$ a cada uno de los miembros de la desigualdad anterior se obtiene

$$P[A] + P[B \text{ pero no A}] \geq P[A]$$

Sustituyendo

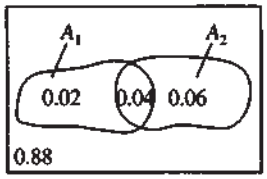
$$P[B] \geq P[A]$$

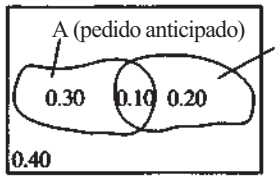
b) Observe que, puesto que C es un suceso, C está contenido en S. Por el apartado *a* $\boldsymbol{P[C] \leq P[S]}$. Ya que por el Axioma 1 $P[S] = 1$, se tiene que $\boldsymbol{P[C] \leq 1}$.

EJERCICIOS 3.2

1. *a)* 0.12 *b)* 0.02 *c)* 0.06 *d)* 0.88

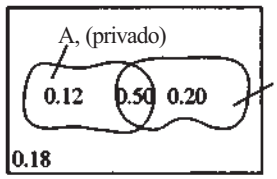
3.





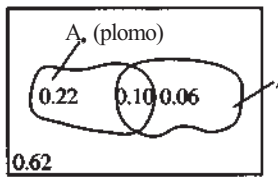
a) 0.60 *b)* 0.20
c) 0.40(5000)(100) = 200 000

5.



0.82; 0.20

7.



$\boldsymbol{P[A_1 \text{ o } A_2] = P[A_1] + P[A_2] - P[\text{ambos}]}$
$\boldsymbol{0.38 = 0.32 + P[A_2] - 0.10}$
$\boldsymbol{P[A_2] = 0.16}$

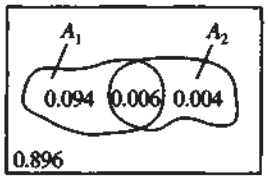
0.16; 0.22

EJERCICIOS 3.3

1. a) $\frac{6}{16}$ *b)* $\frac{2}{8}$ *c)* $\frac{1}{2}$ *d)* $\frac{1}{2}$
3. a) A_1: 65 ó más
A_2 : insuficiencia cardíaca leve
$\boldsymbol{P[A_1 \text{ o } A_2] = P[A_1] + P[A_2] - P[A_1 \text{ y } A_2]}$

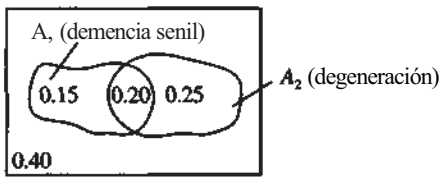
$$0.104 = 0.10 + 0.01 - P[A_1 \text{ y } A_2]$$
$$P[A_1 \text{ y } A_2] = 0.006$$

b)



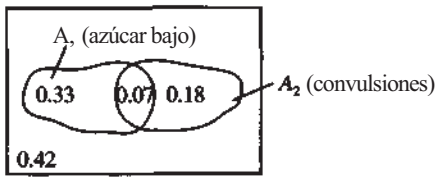
c) $\frac{0.006}{0.094} = 0.0638$ *d)* $\frac{0.004}{0.900} = 0.0044$

5.



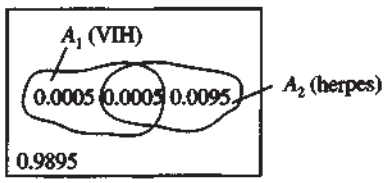
$\frac{0.20}{0.45} = 0.4444$; $\frac{0.25}{0.65} = 0.3846$

7.



0.42; $\frac{0.07}{0.40} = 0.175$

9.



$0.0095 + 0.9895 = 0.999$; $\frac{0.9895}{0.99} = 0.9995$

EJERCICIOS 3.4

1. $\alpha = P[\text{test} + | \text{realidad} -] = \frac{12}{142}$ $\beta = P[\text{test} - | \text{realidad} +] = \frac{4}{58}$
3. $\alpha \cong \frac{7}{402}$; $\beta \cong \frac{19}{98}$
5. $\frac{130}{142}$; alto porque la especificidad es la probabilidad de obtener un resultado del test correcto entre los sujetos realmente negativos.
7.

		Estado real		
		-	+	
Test	+	13	92	105
resultado	-	62	8	70
		75	100	175

$\alpha \cong \frac{3}{75}$

9. *a)* $\alpha \cong \frac{1000}{99\,969} \cong 0.01$; especificidad = $1 - \alpha \cong 0.99$
   b) $\beta \cong \frac{1}{31}$; sensitividad = $1 - \beta = \frac{30}{31}$
11. $\frac{130}{134}$
13. $\alpha \cong \frac{83}{115}$; $\frac{116}{199}$; no, el valor predictivo positivo está próximo a 0.58
15. *a)*

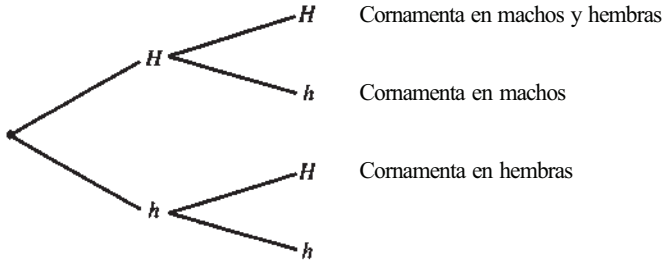
		TB presente		
		sí	no	
consumo	sí	11	296	307
drogas	no	35	1650	1685
		46	1946	1992

   b) $\frac{\frac{11}{307}}{\frac{35}{1685}} = 1.725$

EJERCICIOS 3.5

1. A_1, B_1: no independientes, no mutuamente excluyentes
   A_2, B_2: no independientes, no mutuamente excluyentes
   A_3, B_3: independientes, no mutuamente excluyentes
   A_4, B_4: no independientes, mutuamente excluyentes
   A_5, B_5: no independientes, no mutuamente excluyentes
   A_6, B_6: no independientes, no mutuamente excluyentes
   A_7, B_7: independientes, no mutuamente excluyentes
3. P[BOD alta] • P[acidez elevada] = 0.35(0.10) = 0.035
   Puesto que este producto es distinto que 0.4, la probabilidad de que ambos sucesos ocurran, no son independientes.

   F[acidez elevada | BOD alta] = $\frac{0.04}{0.35} = \frac{4}{35}$
5. P[ambas unidades correctas] = $(0.99)^2 = 0.9801$
7. a) 0.0112 (pppp) 0.0048 (appp) *b)* 0.1 *c)* 0.16 *d)* 0.756
   0.0448 (pppa) 0.0192 (appa)
   0.1008 (ppap) 0.0432 (apap)
   0.4032 (ppaa) 0.1728 (apaa)
   0.0028 (papp) 0.0012 (aapp)
   0.0112 (papa) 0.0048 (aapa)
   0.0252 (paap) 0.0108 (aaap)
   0.1008 (paaa) 0.0432 (aaaa)
9. a)



   b) *P*[macho y cornamenta] = $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$; P[hembra y cornamenta] = $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$

c) P[cornamenta] = P[macho y cornamenta] + P[hembra y cornamenta] = $\frac{4}{8}$;
P[cornamenta | macho] = $\frac{3}{4} \neq$ P[cornamenta]

11. *a)* $P[AB^-] = P[AB]P[RH"] = (0.04)(0.15) = 0.006$
b) $P[AB^-] = (0.04)(0.15) = 0.006$
c) sí; la probabilidad de ser del grupo sanguíneo AB^- es la misma, independientemente de la raza
d) no; para blancos $P[A^-] = (0.40)(0.15) = 0.06$ pero para negros $P[A^-] = (0.27)(0.15) = 0.0405$

13. P[al menos un falso positivo] = 1 - P[ningún falso positivo]
$= 1 - (0.95)^{10}$
$= 0.4013$

EJERCICIOS 3.6

1. P[- -] = 0.15; P[+ +] = 0.37
P[la madre sea - - y el padre sea + +] = 0.15 (0.37) = 0.0555
P[eritroblastosis] = 0.0360 + 0.0555 = 0.0915

3. P[varón sobreviva] = P[sobreviva | varón] P[varón]
$= (0.35)(0.52)$
$= 0.182$

5. P[padezca diabetes y lo ignore] = P[lo ignore | padezca diabetes]P[padezca diabetes]
$= (0.50)(0.02)$
$= 0.01$

7. P[error y observable] = P[observable | error]P[error]
$= 0.35(0.4)$
$= 0.14$

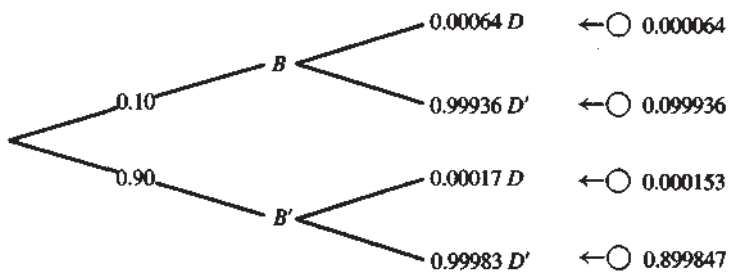
9. *a)* P[sí] = P[sí y A] + P[sí y β]
= P[sí|A]P[A] + P[sí|B]P[B]
Despejando P[sí|B] se obtiene

$$P[\text{sí} \mid B] = \frac{P[\text{sí}] - P[\text{sí} \mid A]P[A]}{P[B]}$$

b) $$P[\text{fraudulenta}] = P[\text{sí} \mid B] = \frac{0.60 - 0.5(0.5)}{0.5} = 0.70$$

EJERCICIOS 3.7

1. a) B = negra
D = muere



b) $P[B|D] = \dfrac{P[D \text{ y } B]}{P[D]} = \dfrac{0.000064}{0.000064 + 0.000153} \cong 0.295$

c) $P[B|D] = \dfrac{P[D|B]P[B]}{P[D|B]P[B] + P[D|B']P[B']}$

$$= \frac{0.00064(0.10)}{0.00064(0.10) + 0.00017(0.90)} = 0.295$$

3. $P[B \text{ responsable } | \text{remisión}] =$

$$\frac{P[\text{remisión}|B]P[B]}{P[\text{remisión}|B]P[B] + P[\text{remisión}|A]P[A] + P[\text{remisión}|C]P[C]}$$

$$= \frac{\frac{3}{4}\left(\frac{1}{3}\right)}{\frac{3}{4}\left(\frac{1}{3}\right) + \frac{1}{2}\left(\frac{1}{3}\right) + \frac{6}{10}\left(\frac{1}{3}\right)}$$

$$= \frac{\frac{1}{4}}{\left(\frac{1}{4}\right) + \left(\frac{1}{6}\right) + \left(\frac{2}{10}\right)} \cong 0.4054$$

5. *a)* $P[\text{realidad} + | \text{test} +] =$

$$= \frac{P[\text{realidad} + | \text{test} +]P[\text{realidad} +]}{P[\text{test} + | \text{realidad} +]P[\text{realidad} +] + P[\text{test} + | \text{realidad} -]P[\text{realidad} -]}$$

$$= \frac{0.95(0.10)}{0.95(0.10) + 0(0.90)}$$

$$= 1$$

b) $\dfrac{0.95(0.10)}{0.95(0.10) + 0.50(0.90)} = 0.1743$

c) disminuye de 1 a 0.1743 (diferencia = 0.8257)

d) $\dfrac{1(0.10)}{1(0.10) + 0.05(0.90)} \cong 0.6896$

e) $\dfrac{0.5(0.10)}{0.5(0.10) + 0.5(0.90)} \cong 0.5263$

f) disminuye de 0.6896 a 0.5263 (diferencia = 0.1633)

EJERCICIOS 4.1

1. discreta 3. continua 5. discreta 7. continua 9. continua
11. discreta 13. continua

EJERCICIOS 4.2

1. a) 0.09; P[x = 5] *b)* **0.41; $P[X \leq 2]$ =** *P[a* lo sumo 2 personas solicitan un tratan miento innecesario en un día aleatoriamente seleccionado] c) 0.11 *d)* 0.19

3. *a)*

x	0	1	2	3	4
$f(x)$	$(0.1)^4$	$4(0.1)^3(0.9)^1$	$6(0.1)^2(0.9)^2$	$4(0.1)(0.9)^3$	0.9^4

o

x	0	1	2	3	4
$f(x)$	0.0001	0.0036	0.0486	0.2916	0.6561

b) 0.0037; $P[X \leq 1]$ = P[a lo sumo un paciente obtendrá alivio]

c) sí; P[ninguno obtendrá alivio] $= f(0) = 0.0001$; sería muy raro que este suceso ocurriese

5. $E[X] = \mu = 0.2$

7. *a)* 8.75 *b)* 67.25 *c)* 13.32 *d)* 0.36

e)
$$E[(X - \mu)^2] = E[(X - 0.2)^2]$$
$$= (-2.2)^2(0.1) + (-1.2)^2(0.2) + (-0.2)^2(0.3) + (0.8)^2(0.2) + (1.8)^2(0.2)$$
$$= 1.56$$

f)
$$E[(X - \mu)^2] = E[(X - 3)^2]$$
$$= (-2)^2(0.1) + (-1)^2(0.15) + (0)^2(0.5) + (1)^2(0.15) + (2)^2(0.1)$$
$$= 1.1$$

9. *b)* $E[X^2] - (E[X])^2 = 8.75 - (2.75)^2 = 1.1875$

c) $67.25 - (8.15)^2 = 0.8275$

11. *a)*

x	0	1	2	3
$f(x)$	$(0.05)^3$	$3(0.95)(0.05)^2$	$3(0.95)^2(0.05)$	$(0.95)^3$

o

x	0	1	2	3
$f(x)$	0.000125	0.007125	0.135375	0.857375

b) $E[X]$ = número medio de pacientes desensibilizados = 2.85

c) $\mu = E[X] = 2.85$

d) $E[X^2] = 8.265$

e) $\sigma^2 = \text{Var } X = 0.1425$; $\sigma = \sqrt{0.1425} \cong 0.3775$

13. *a)*

0 1 2

$$P[|X - 1| \leq 2] = P[|X - 1| \leq 4\sigma] \geq 1 - \frac{1}{4^2} = \frac{15}{16}$$

La probabilidad de que X tome valores por encima de 2 ó por debajo de 0 es, a lo sumo, de $\frac{1}{16}$. Como la pluviosidad no puede ser negativa, $P[X > 2]$ es, como mucho, $\frac{1}{16}$.

b) Ya que 2 está 7.5σ alejado de la media, $P[X \geq 2] \leq \frac{1}{(7.5)^2} = \frac{4}{225}$

EJERCICIOS 4.3

1. a)

x	6	7	8	9	10
$f(x)$	0.05	0.15	0.75	0.90	1.00

b) 0.75 *c)* $P[X > 7] = 1 - P[X \leq 7] = 0.85$
d) $P[7 \leq X \leq 9] = P[X \leq 9] - P[X \leq 6] = 0.90 - 0.05 = 0.85$

3. *a)*

x	0	1	2	3
$f(x)$	0.15	0.40	0.90	1.00

b) $P[X \leq 1] = 0.40$ *c)* $P[X \geq 2] = 1 - P[X \leq 1] = 0.60$

5. *a)* 0.6, 0.9, 0.1, 0.6
b)

x	0	1	2	3	4	5	6
$f(x)$	0.1	0.1	0.1	0.3	0.2	0.1	0.1

c) 3.1 *d)* 2.89 *e)* $\sqrt{2.89}$; casos de SIDA.

EJERCICIOS 4.4

1. *a)* binomial; $n = 5$; $p = 0.006$
b) binomial; $n = 5$; $p = 0.001$
c) binomial; $n = 10$; $p = 0.9$
d) no binomial; p cambia debido a que se extraen muestras sin reemplazamiento de un grupo pequeño de objetos
e) aproximadamente binomial; $n = 8$; $p \cong 0.05$
f) no binomial; no se ha fijado el número de pruebas

3. *b)* $f(x) = \dfrac{10!}{x!\,(10-x)!}(0.01)^x(0.99)^{10-x} \quad x = 0, 1, 2, \ldots, 10$

c) $E[X] = 10(0.01) = 0.1$ *d)* no; las pruebas no son independientes
$\text{Var } X = 10(0.01)(0.99) = 0.099$
$\sigma = \sqrt{0.099}$

5. *a)* $\dfrac{4!}{x!\,(4-x)!}(0.2)^x(0.8)^{4-x} \quad x = 0, 1, 2, 3, 4$ *b)* $f(0) = \dfrac{4!}{0!\ 4!}(0.2)^0(0.8)^4 = 0.4096$

c) $f(1) = \dfrac{4!}{1!\ 3!}(0.2)^1(0.8)^3 = 0.4096$, o $P[X \leq 1] = 0.4096 + 0.4096 = 0.8192$; $n = 4$ no está en la tabla

7. *a)* 0.6331 *b)* 0.3823 *c)* 0.2508 *d)* 0.3669 *e)* 0.0548 *f)* 0.7779 *g)* 0.6054 *h)* 0.6665 *i)* 0.3546

9. $E[X] = 20(0.6) = 12$; $P[X \leq 9] = 0.1275$

EJERCICIOS 4.5

1. *a)* 10 *b* $f(x) = \dfrac{e^{-10}10^x}{x!}$ C = 0,1,2,... *c)* 0.029 *d)* 0.933 *e)* 0.782 *f)* 0.125

3. $\lambda = 2;\ s = 3;\ \lambda s = 6;\ P[X \leq 4] = 0.285;\ E[X] = 6$
   $P[X \geq 12] = 1 - P[X \leq 11] = 1 - 0.98 = 0.02$
5. $\lambda = 5;\ s = \frac{16}{20} = 0.8;\ \lambda s = 4;\ P[X = 0] = 0.018$
   $P[X \geq 9] = 1 - P[X \leq 8] = 1 - 0.979 = 0.021$
7. $\lambda = 1;\ s = 0.5;\ \lambda s = 0.5;\ P[X = 0] = 0.607$
   $P[X \geq 1] = 1 - 0.607 = 0.393$
9.

x	3	4	5	6	7	8	9	10
Binomial	0.8670	0.9568	0.9887	0.9976	0.9996	0.9999	1.000	1.000
Poisson	0.857	0.947	0.983	0.995	0.999	1.000	1.000	1.000

11. $n = 2\,000\,000;\ p = \dfrac{5}{1\,000\,000};\ np = \lambda s = 10$

    $P[X \geq 1] = 1 - P[X = 0] \cong 1.000$
13. $E[X] = (150)(0.10) = 15;\ P[X = 10] = P[X \leq 10] - P[X \leq 9] \cong 0.118 - 0.07 = 0.048$

EJERCICIOS 5.1

1. *c)* P[hayan de prescribirse como mucho 0.2 cm^3] $= P[Z \leq 0.2]$
   d) P[hayan de prescribirse como mucho 0.1 cm^3] $= P[Z \leq 0.1]$
   e) P[hayan de prescribirse entre 0.1 y 0.2 cm^3] $= P[0.1 \leq Z \leq 0.2]$
   f) Réstese la respuesta a la parte *d* de la respuesta a la parte *c*.
   g) μ debe estar próxima a 0.2.
   h) $f(0.2) = \frac{200}{9} \cdot \frac{2}{10} = \frac{40}{9};\ P[X \leq 0.2] = \frac{1}{2}\left(\frac{2}{10}\right)\frac{40}{9} = \frac{4}{9}$
   i) $f(0.1) = \frac{200}{9} \cdot \frac{1}{10} = \frac{20}{9};\ P[X \leq 0.1] = \frac{1}{2}\left(\frac{1}{10}\right)\frac{20}{9} = \frac{1}{9}$
   j) $\frac{4}{9} - \frac{1}{9} = \frac{3}{9}$
3. *b)* $\frac{9}{16}$ *c)* 0 (X es continua)
5. *c)* $P[X > 3] = 2 \cdot \frac{1}{5} = \frac{2}{5}$ *d)* 2.S minutos

EJERCICIOS 5.2

1. a) I y II
   b) *P[el* marcapasos dure entre 2 y 4 años] $= P[2 \leq X \leq 4]$
   c) P[el marcapasos dure al menos 8 años] $= P[X \geq 8];\ 1 - F(8)$
   d) F(4)
3. a) I, III y III *b)* 1 - F(18); III, IV y V *c)* $F(36) - F(27)$; IV

EJERCICIOS 5.3

1. a) $f(x) = \dfrac{1}{2\sqrt{2\pi}} e^{-1/2(x-5/2)^2} \quad -\infty < x < \infty$ *V)* 5 ± 2 (3 y 7)
3. *a)* 0.0643 *b)* 0.9147 *c)* 0.9147 *d)* 0.9222 *e)* 0.0239 *f)* 0.8451
   g) 0 *h)* −1.645 *i)* 0.67 *j)* 1.28 *k)* −0.84 *l)* 1.96 *m)* 2.575

5. *a)* $z = \dfrac{100 - 153}{25} = -2.12;\ P[Z \leq -2.12] = 0.0170$

b) $z = \dfrac{180 - 153}{25} = 1.08;\ P[Z \geq 1.08] = 0.1401$

c) $z = -2.12$ y $z = \dfrac{175 - 153}{25} = 0.88;\ P[-2.12 \leq Z \leq 0.88] = 0.7936$

d) $z = \dfrac{128 - 153}{25} = -1$ y $z = \dfrac{178 - 153}{25} = 1;\ P[-1 \leq Z \leq 1] = 0.6827$

e) $\dfrac{x_0 - 153}{25} = -1.28; x_0 = 153 + 25(-1.28) = 121$

f) $\dfrac{x_0 - 153}{25} = -1.555;\ x_0 = 153 + 25(1.555) = 191.875$

7. *a)* $P[X \leq 120] = P\left[Z \leq \dfrac{120 - 106}{8}\right] = P[Z \leq 1.75] = 0.9599$

b) $P[90 \leq X \leq 120] = P[-2 \leq Z \leq 1.75] = 0.9599 - 0.0228 = 0.9371$ o 93.71 %

c) $P[106 \leq X \leq 110] = P[0 \leq Z \leq 0.5] = 0.6915 - 0.5 = 0.1915$

d) $P[X \geq 121] = P[Z \geq 1.875] = 1 - 0.9699 = 0.0301$

e) $\dfrac{x_0 - 106}{8} = -0.67;\ x_0 = 106 - 8(0.67) = 100.64$

9. *a)* $f(x) = \dfrac{1}{0.2\sqrt{2\pi}} e^{-1/2(x - 1.5/0.2)^2} \qquad -\infty < x < \infty$

$P[-1.1 \leq X \leq 1.9] = P[-2 < Z < 2] = 0.9544$

b) $P[X < 0.9] = P\left[Z < \dfrac{0.9 - 1.5}{0.2}\right] = P[Z < -3] = 0.0013$

c) $P[X > 2] = P[Z > 2.5] = 0.0062$; sí, este es un hecho inusual

EJERCICIO 5.4

1. $\mu = 20.3$ y $\sigma = 1.4$; 18.9 a 21.7 está dentro de una desviación típica de μ; esto ocurre con probabilidad aproximada de 0.68; no es usual
3. $\mu = 0.25$ y $\sigma = 0.11$; 0.03 a 0.47 está dentro de dos desviaciones típicas de μ; esto ocurre con probabilidad aproximada de 0.95
5. *a)* varón menor de 21: 140 a 180
varón de 21 a 29 : 140 a 260
varón de 30 o más : 160 a 280
c) 20 de edad y 125 es inusualmente bajo.
20 de edad y 200 es inusualmente elevado,

EJERCICIOS 6.2

1. $\bar{x} = 408.3$ 3. más corto

7. a) $\text{Var } \bar{X} = \dfrac{\sigma^2}{n} = \dfrac{0.4829}{10} = 0.04829$ **c)** error estándar $= \sqrt{0.04829} = 0.21975$

9. $E[\bar{X}] = \mu = 13$; $\text{Var } \bar{X} = \dfrac{\sigma^2}{n} = \dfrac{9}{16}$; error estándar $= \dfrac{3}{4}$

11. *a)*

x	0	1
$f(x_i)$	1/2	1/2

$E[X_i] = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$

$E[X_i^2] = 0^2\left(\frac{1}{2}\right) + 1^2\left(\frac{1}{2}\right) = \frac{1}{2}$

$\text{Var } X_i = E[X_i^2] - (E[X_i])^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

a) $\sum X_i =$ número de caras y $\dfrac{\sum X_i}{n} =$ proporción de caras

c) $\bar{X}$ está, aproximadamente, normalmente distribuida con media $\frac{1}{2}$ y varianza $\left(\frac{1}{4}\right)$ **30**

d) normal estándar

EJERCICIOS 6.3

1. a)
```
0|4
0|7877
1|1400011000
1|7655
2|2
```
Tiene forma aproximada de campana.

b) $\bar{x} = 11.533$; $s = 4.210$; $s^2 = 17.7258$

3. *a)* 1.753 *b)* 2.131 *c)* −1.753 *d)* -2.131 *e)* 0.01 *f)* 0.10 *g)* 0.90 *h)* 2.131 *i)* 2.947

5. $\bar{x} \pm 1.96(s/\sqrt{n})$ o $9.5 \pm 1.96\ (0.5/\sqrt{1000})$; 9.5 ± 0.031; (9.469, 9.531)

7. $\bar{x} \pm 1.699\left(\dfrac{s}{\sqrt{n}}\right)$ o $41.2 \pm 1.699\left(\dfrac{2.1}{\sqrt{30}}\right)$; 41.2 ± 0.65; (40.55, 41.85); más largo; no; todo el intervalo está por encima del 35 %

9. $\bar{x} \pm 3.355\left(\dfrac{s}{\sqrt{n}}\right)$ o $5.66 \pm 3.355\left(\dfrac{0.49}{\sqrt{9}}\right)$; (5.11, 6.21)

11. $\bar{x} \pm 2.131\left(\dfrac{s}{\sqrt{n}}\right)$ o $11.533 \pm 1.729\left(\dfrac{4.210}{\sqrt{20}}\right)$; (11.146; 11.920)

EJERCICIOS 6.4

1. *a)* H_0: $\mu \geq 0.08$, H_1: $\mu < 0.08$

b) Concluiremos que el porcentaje medio de metal en basuras caseras ha disminuido cuando en realidad no lo ha hecho.

c) No detectaremos el hecho de que el porcentaje medio ha disminuido cuando en realidad lo ha hecho.

3. a) H_0: $\mu \geq 9$, H_1: $\mu < 9$

b) Tipo I: Se cree que el nivel de DDT es más bajo de lo que lo es en la actualidad. Por lo tanto, los controles se cumplieron tan cuidadosamente como se exigía.
Tipo II: No se puede probar que el nivel de DDT se ha reducido cuando esto es lo que ha ocurrido en la realidad. Los controles con los que se trabaja podrían ser aún más rigurosos.

5. Tipo I: Se obtiene un «falso positivo». El paciente cree que tiene el virus del SIDA cuando en realidad no es cierto.
Tipo II: El virus está presente pero no es detectado. El resultado del test es un «falso negativo».

7. H_1: $\mu < 8$, H_0: $\mu \geq 8$
Tipo I: Se piensa que la lluvia ácida impide el crecimiento de los pimpollos pero esto no es verdad.

EJERCICIOS 6.5

1. $P = P[T_A > 3]$; $0.001 < P < 0.005$; se rechaza H_0
3. $P = P[T_{15} < -1.5] + [T_{15} > 1.5]$; $0.10 < P < 0.20$; no se debería rechazar H_0
5. *a)* H_0: $\mu \leq 0.035\,\%$, H_1: $\mu > 0.035\,\%$
   b) $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{0.09 - 0.035}{0.25/\sqrt{144}} = 2.64$; $P = P[T_{143} > 2.64]$; usando la fila ∞, $0.001 < P < 0.005$

7. $H_1 = \mu > 25$; $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{27 - 25}{18/\sqrt{80}} = 0.99$; $P = P[T_{79} > 0.99]$; $0.10 < P < 0.25$
Los datos no apoyan lo que afirma el investigador. Si se hace la afirmación, la probabilidad de error está entre 0.10 y 0.25.

9. sí; sí; no
En los Ejercicios 1 y *2, P* es menor que 0.05. En el Ejercicio 3, *P* es mayor que 0.05.

11. *a)* H_0: $\mu = 2.5$, H_1: $\mu \neq 2.5$
    b) $\bar{x} = 2.66$; $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{2.66 - 2.50}{0.2/\sqrt{10}} = 2.53$; $s = 0.20$
    $P = P[T_9 < -2.53] + P[T_9 > 2.53]$; $0.05 < P < 0.10$
    Rechazamos H_0 al nivel $\alpha = 0.10$ porque $P < 0.10$. Es posible un error Tipo I.

13. H_0: $\mu \geq 1.3$, H_1: $\mu < 1.3$; $\bar{x} = 0.8$; $s = 0.8$; $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{0.8 - 1.3}{0.8/\sqrt{20}} = -2.80$
$P = P[T_{19} < -2.80]$; $0.005 < P < 0.01$; puesto que $P < 0.01$, H_0 puede ser rechazada al nivel $\alpha = 0.01$.

EJERCICIOS 6.6

1. $n = \dfrac{z^2 s^2}{d^2} = \dfrac{(2.576)^2(0.066)}{(0.1)^2} = 43.796 \cong 44$

3. $s = 1.43$; $n = \dfrac{(1.96)^2(1.43)^2}{(0.5)^2} \cong 32$

5. $s = 2.3$; $\Delta = \frac{2}{2.3} = 0.9$; $\alpha = 0.05$; $\beta = 0.05$; $n = 15$ (contraste de una cola)

EJERCICIOS 7.1

1. a) 0.01 *b)* 0.25 c) 0.05 d) 19.0 e) 2.09 *f)* $\chi_1^2 = 3.33;\ \chi_2^2 = 16.9$

3. *a)* $\bar{x} = 3840.2$ *b)* $s^2 = 1261.69$

   c) $\bar{x} \pm 1.86\left(\frac{s}{\sqrt{n}}\right)$ o $3840.2 \pm 1.86(35.5/\sqrt{9})$; (3818.2, 3862.2)

   d) $L_1 = \frac{8(1261.69)}{15.5} = 651.19$

   $L_2 = \frac{8(1261.69)}{2.73} = 3697.26$; $(651.19, 3697.26)$; $(\sqrt{651.19}, \sqrt{3697.26})$

5. *a)* $\chi^2_{0.05} \cong \frac{1}{2}[1.645 + \sqrt{2(79) - 1}]^2 = 100.46$

   $\chi^2_{0.90} \cong \frac{1}{2}[-1.28 + \sqrt{2(79) - 1}]^2 = 63.28$

   b) $\chi^2_{0.025} \cong \frac{1}{2}[-1.96 + \sqrt{2(99) - 1}]^2 = 72.91$

   $\chi^2_{0.975} \cong \frac{1}{2}[1.96 + \sqrt{2(99) - 1}]^2 = 127.93$

   c) $\chi^2_{0.005} \cong \frac{1}{2}[-2.575 + \sqrt{2(74) - 1}]^2 = 45.59$

   $\chi^2_{0.995} \cong \frac{1}{2}[2.575 + \sqrt{2(74) - 1}]^2 = 108.04$

7. **a)** $\bar{x} \pm 1.671\left(\frac{s}{\sqrt{n}}\right)$ o $3400 \pm 1.671\left(\frac{100}{\sqrt{61}}\right)$; (3378.6, 3421.4)

   b) $\chi^2_{0.05} = \frac{1}{2}[-1.645 + \sqrt{2(60) - 1}]^2 = 42.91$; $\chi^2_{0.95} = \frac{1}{2}[1.645 + \sqrt{2(60 - 1}]^2 = 78.80$

   $L_1 = \frac{60(100)^2}{78.80} = 7614.21$; $L_2 = \frac{60(100)^2}{42.91} = 13\,982.75$

   $\sqrt{L_1} = 87.3$; $\sqrt{L_2} = 118.2$

EJERCICIOS 7.2

1. *a)* $\bar{x} = 12.11,\ s = 1.48,\ s^2 = 2.186$

   b) $\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{12.11 - 13}{\frac{148}{\sqrt{12}}} = -2.08$; $P = P[T_{11} < 2.08]$; sí, $0.025 < P < 0.05$

   c) $\frac{(n-1)s^2}{\sigma_0^2} = \frac{11(2.186)}{2.25} = 10.69$; $P = P[X^2_{11} < 10.69]$; no, $0.50 < P < 0.75$

3. *a)* $\frac{\bar{x} - \mu_0}{\frac{\sigma_0^2}{\sqrt{25}}} = \frac{6.3 - 6.5}{\frac{1.7}{\sqrt{25}}} = -0.588$; $P = P[T_{24} < -0.588]$; no, $0.25 < P < 0.40$ y, por tanto, $P > 0.05$

b) $\dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{24(1.7)^2}{(1.2)^2} = 48.17$; $P = P[X_{24}^2 > 48.17]$; sí, $P < 0.005$ y, por tanto, $P < 0.05$

EJERCICIOS 8.1

1. $\hat{p} = \frac{32}{36} = 0.889$ **3.** $\hat{p} = \frac{11}{15} = 0.733$
7. a) $E[X_i] = 0(1-p) + 1(p) = p$; $E[X_i^2] = 0^2(1-p) + 1^2(p) = p$;
   $\text{Var } X_i = E[X_i^2] - (E[X_i])^2 = p - p^2 = p(1-p)$
   d) normal estándar

EJERCICIOS 8.2

1. $0.027 \pm 1.96\sqrt{\dfrac{0.027(0.973)}{150}}$ o (0.001, 0.053); menor
3. a) $\hat{p} = \dfrac{125}{200} = 0.625$ **b)** $0.625 \pm 1.645\sqrt{\dfrac{0.625(0.375)}{200}}$ o (0.569, 0.681)
   c) no; los valores 60-68 % están contenidos en el intervalo
5. $\hat{p} = 0.3$; $0.3 \pm 1.645\sqrt{\dfrac{0.3(0.7)}{1000}}$ o (0.276, 0.324); 0.276 (2 billones) = 552 millones, 0.324 (2 billones) = 648 millones
7. $0.84 \pm 1.96\sqrt{\dfrac{0.84(0.16)}{191}}$ o (0.788, 0.892)

EJERCICIOS 8.3

1. $0.08 \pm 2.33\sqrt{\dfrac{0.08(0.92)}{13\,573}}$ o (0.075, 0.085) **3.** $n \cong \dfrac{(1.88)^2}{4(0.03)^2} \cong 982$
5. $\hat{p} = 0.84$; $n \cong \dfrac{(1.88)^2(0.84)(0.16)}{(0.03)^2} \cong 528$
7. *a)* $g'(p) = 1 - 2p$
   b) $1 - 2p = 0$
   $2p = 1$
   $p = \frac{1}{2}$
   c) $g''(p) = -2$; el signo negativo de la segunda derivada implica concavidad hacia abajo y, por lo tanto, un máximo

EJERCICIOS 8.4

1. *a)* $H_0: p \leq \frac{1}{2}$, $H_1: p > \frac{1}{2}$

b) $\hat{p} = \frac{270}{500} = 0.54; \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.54 - 0.50}{\sqrt{\frac{0.50(0.50)}{500}}} = 1.79$

$P = P[Z > 1.79] = 0.036$; se rechaza H_0 puesto que P es pequeño; sujeto a error de Tipo I; el proyecto se pararía innecesariamente.

3. $\hat{p} = \frac{14}{50} = 0.28; z = \frac{0.28 - 0.70}{\sqrt{\frac{0.7(0.3)}{50}}} = -6.48; P < 0.0002$

5. Necesitamos que z sea al menos 1.96 $(z_{0.025})$. Por tanto, debe cumplirse

$$\frac{\hat{p} - 0.7}{\sqrt{\frac{0.7(0.3)}{50}}} \geq 1.96$$

Despeje para ver que debe ser al menos 0.827. Esto significa que $\frac{x}{50}$ deberá ser al menos 0.827 ó x al menos 50 (0.827) = 42.

7. **a)** $H_1: p > 0.85$ *b)* $\hat{p} = \frac{123}{139} = 0.885$; quizás, porque $\hat{p} > 0.85$

c) $z = \frac{0.885 - 0.850}{\sqrt{\frac{0.850(0.150)}{139}}} = 1.16; P = P[Z > 1.16] = 0.123$; no se puede rechazar H_0 al nivel $\alpha = 0.10$ porque $P > 0.10$

d) El tamaño de la muestra es algo pequeño.

EJERCICIOS 8.5

1. $\hat{p}_1 = \frac{29}{742} = 0.039; \hat{p}_2 = \frac{13}{733} = 0.018; \hat{p}_1 - \hat{p}_2 = 0.021$
3. $\hat{p}_1 - \hat{p}_2 = 0.40 - 0.19 = 0.21$
5. $0.316 \pm 1.645\sqrt{\frac{0.895(0.105)}{19} + \frac{0.579(0.421)}{19}}$ o 0.316 ± 0.219; $(0.097, 0.535)$; sí, porque el intervalo contiene únicamente valores positivos.
7. $\hat{p}_1 - \hat{p}_2 = 0.89 - 0.504 = 0.386; 0.386 \pm 1.645\sqrt{\frac{0.89(0.11)}{500} + \frac{0.504(0.496)}{500}}$ o 0.386 ± 0.043; $(0.343, 0.429)$
9. $n = \frac{(1.96)^2[0.039(0.961) + 0.018(0.982)]}{(0.02)^2} \cong 530;$

$n = (1.645)^2 \frac{[0.039(0.961) + 0.018(0.982)]}{(0.02)^2} \cong 374$

11. $n = \frac{(1.96)^2}{2(0.02)^2} \cong 4802$

Tener disponible una estimación anterior permite que utilicemos muestras más pequeñas en el estudio real.

EJERCICIOS 8.6

1. Tipo II: La vitamina C pudo ser realmente eficaz, pero no se detecta, así que un tratamiento eficaz ha podido desecharse; no; no se ha probado que es ineficaz; simplemente no hemos demostrado que sea eficaz.

3. H_0: $p_1 \leq p_2$, H_1: $p_1 > p_2$; $\hat{p}_1 = \frac{38}{101} = 0.376$; $\hat{p}_2 = \frac{6}{31} = 0.194$; $\hat{p}_1 - \hat{p}_2 = 0.182$;

$$\hat{p} = \frac{38 + 6}{132} = 0.333; \; z = \frac{0.182}{\sqrt{0.333(0.667)\left(\frac{1}{101} + \frac{1}{31}\right)}} = 1.88$$

$P = P[Z > 1.88] = 0.0301$; se rechaza H_0 al nivel $\alpha = 0.05$ porque $P < 0.05$

5. H_0: $p_1 - p_2$, H_1: $p_1 - p_2$; $\hat{p}_1 = \frac{162}{2055} = 0.079$; $\hat{p}_2 = \frac{14}{266} = 0.053$;

$\hat{p}_1 - \hat{p}_2 = 0.026$; $\hat{p} = \frac{162 + 14}{2055 + 266} = 0.076$;

$z = \frac{0.026}{\sqrt{0.076(0.924)\left(\frac{1}{2055} + \frac{1}{266}\right)}} = 1.51$; $P = P[Z > 1.51] = 0.0655$; se rechaza para $\alpha = 0.10$ pero no con $\alpha = 0.05$

EJERCICIOS 9.1

1. $\bar{x}_1 = 3.65$; $\bar{x}_2 = 3.5$; $\widehat{\mu_1 - \mu_2} = 0.15$
3. $E[\bar{X}_1 - \bar{X}_2] = 15 - 10 = 5$; $\text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{16}{20} + \frac{18}{25} = 1.52$; $\sigma = \sqrt{1.52} = 1.23$; Z
5. $\widehat{\mu_1 - \mu_2} = 5.4 - 6.8 = -1.4$

EJERCICIOS 9.2

1. $s_1^2/s_2^2 = (6.34)^2/(3.20)^2 \cong 3.93$; hay evidencia de que $\sigma_1^2 \neq \sigma_2^2$
3. $s_2^2/s_1^2 = (11.02)^2/(8.68)^2 = 1.61$; imposible concluir que $\sigma_1^2 \neq \sigma_2^2$
5. *a)* $b = 2.57$ $a = 1/2.76 \cong 0.36$ ***b)*** $b = 2.75$ $a = 1/2.91 \cong 0.34$
   c) $b = 2.85$ $a = 1/3.29 \cong 0.30$
   $s^2 = 0.092$; $s_w^2 = 0.291$; $s_w^2/s_m^2 = 3.16$; $P = P[F_{7,9} > 3.16]$; $0.05 < P < 0.10$
7. Ya que $P > 0.05$, no puede rechazarse H_0 con $\alpha = 0.05$.
9. ***a)*** sí; sí *b)* sí; no ***c)*** $F_{20,10}$; $s_1^2/s_2^2 = 3.00$ ***d)*** $F_{20,10}$; $s_1^2/s_2^2 = 2.00$
   e) No; si el valor F de la tabla es mayor que 2, entonces la regla práctica conduce al rechazo incluso aunque el contraste *F* no lo haga. Si el valor *F* es 2 o menos, entonces la regla práctica y el contraste *F* presentan el mismo nivel de rechazo.

EJERCICIOS 9.3

1. a) $s_p^2 = \frac{9(42) + 13(37)}{10 + 14 - 2} = 39.05$ **b)** $s_p^2 = \frac{28 + 30}{2} = 29$

c) $s_p^2 = \dfrac{9(20) + 49(40)}{10 + 50 - 2} = 36.896$

3. varones: $s^2 = 0.092$, $\bar{x} = 3.65$
mujeres: $s^2 = 0.291$, $\bar{x} = 3.50$
$\dfrac{0.291}{0.092} = 3.16 > 2$; por la regla práctica, no es apropiado estimarlas de forma conjunta.

5. $s_1^2/s_2^2 = (10.1)^2/(10)^2 = 1.02 < 2$; es apropiada la estimación conjunta.

$$s_p^2 = \frac{(10.1)^2 + (10)^2}{2} = 101.005$$

$(\bar{x}_1 - \bar{x}_2) \pm 1.746\sqrt{101.005\left(\frac{1}{9} + \frac{1}{9}\right)}$ o 1 ± 8.27; $(-7.27, 9.27)$; no puede concluirse que hay diferencias en el tiempo medio de supervivencia porque el intervalo contiene al 0.

7. $s_1^2/s_2^2 = (9.9)^2/(9.5)^2 = 1.09 < 2$; es apropiada la estimación conjunta.

$$s_p^2 = \frac{17(9.9)^2 + 4(9.5)^2}{18 + 5 - 2} = 96.53$$

H_0: $\mu \leq \mu_2$, H_1: $\mu_1 > \mu_2$; $t = \dfrac{(42.7 - 27.7)}{\sqrt{96.53\left(\frac{1}{18} + \frac{1}{5}\right)}} = 3.02$; $P = P[T_{21} > 3.02]$;
$0.001 < P < 0.005$; se rechaza H_0.

9. $s_2^2/s_1^2 = (2)^2/(1.9)^2 = 1.11 < 2$; es apropiada la estimación conjunta.

$$s_p^2 = \frac{50(1.9)^2 + 40(2)^2}{90} = 3.78$$

H_0: $\mu_1 \geq \mu_2$, H_1: $\mu_1 < \mu_2$; $t = \dfrac{59.1 - 65.2}{\sqrt{3.78\left(\frac{1}{51} + \frac{1}{41}\right)}} = -14.96$;
$P = P[T_{90} < -14.96]$; $P < 0.0005$; se rechaza H_0.

11. $s_1^2/s_2^2 = \frac{4}{3.5} = 1.14 < 2$; es apropiada la estimación conjunta; $s_p^2 = \dfrac{32(4) + 13(3.5)}{45} = 3.856$

H_0: $\mu_1 \geq \mu_2$, H_1: $\mu_1 < \mu_2$; $t = \dfrac{11.3 - 12.6}{\sqrt{3.856\left(\frac{1}{33} + \frac{1}{14}\right)}} = -2.08$;
$P = P[T_{45} < -2.08]$; $0.01 < P < 0.025$; se rechaza H_0.

EJERCICIOS 9.4

1. $t = 1.83$; 67; $0.025 < P < 0.05$; aunque el valor P se ve afectado, la conclusión es la misma.

3. $s_1^2/s_2^2 = (6.5)^2/(3.6)^2 = 3.26 > 2$; no es apropiada la estimación conjunta
H_0: $\mu_1 \geq \mu_2$, H_1: $\mu_1 < \mu_2$; $t = 15.53'$ $g.l. = 33.95$; $P < 0.005$

5. conjunta; $s_p^2 = \dfrac{3460(8.68)^2 + 2237(11.02)^2}{5697} = 93.44$

H_0: $\mu_1 \leq \mu_2$, H_1: $\mu_1 > \mu_2$; $t = 852.76$; $P < 0.0005$; se rechaza H_0.

7. no conjunta; H_0: $\mu_1 \leq \mu_2$, H_1: $\mu_1 > \mu_2$; $t = 8.73$; $g.l. = 24.03$; $P < 0.005$
9. no conjunta; $g.l. = 11.06$; $(-8.28, -0.00014)$; sí, ya que el intervalo no contiene al 0.
11. $s_2^2/s_1^2 = 36/25 \equiv 1.44 < 2$; es apropiada la estimación conjunta.

$$s_p^2 = \frac{96\,319(25) + 81\,608(36)}{96\,320 + 81\,609 - 2} = 30.05;\ (-5.521, -5.419)$$

EJERCICIOS 9.5

1. $\bar{d} = 12.55$; $s_d = 24.47$; $(-25.92, 0.82)$; no, porque el intervalo contiene al 0.
3. H_0: $\mu_D \leq 0$, H_1: $\mu_D > 0$; $t = 2.76$; $0.001 < P < 0.005$; se rechaza H_0.
5. H_0: $\mu_D \leq 0$, H_1: $\mu_D > 0$; $t = 4.63$; $P < 0.0005$

EJERCICIOS 10.1

1. *a)* Son de particular interés para el investigador y no se seleccionan aleatoriamente.
   b) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
   c) $T_{1.} = 591.4$ $T_{..} = 1835.4$
   $T_{2.} = 460.4$ $\bar{X}_{..} = 36.71$
   $T_{3.} = 364.5$ $\sum\sum x^2 = 79\,896.22$
   $T_{4.} = 254.7$
   $T_{5.} = 164.4$
   d) $SS_{\text{Total}} = 12\,522.36$, $SS_{\text{Tr}} = 11\,274.32$, $SS_E = 1248.04$
   e) $MS_{\text{Tr}} = 2818.58$, $MS_E = 27.73$
   f) $F_{4,45} = 101.64$
   g) **sí**; $P < 0.01$ ($P \cong 0$ vía TI83)
   h) suponemos normalidad e igualdad de varianzas

3. ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	**4**	3.935	0.984	8.07
Error	**37**	4.497	0.122	
Total	**41**			

$P = P[F_{4,37} > 8.09]$; $P < 0.01$; se concluye que hay diferencias en el contenido medio de azufre entre los cinco yacimientos.

5. ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	*2*	7.73	3.865	107.36
Error	**29**	1.04	0.036	
Total	**31**			

se rechaza H_0: $\mu_1 = \mu_2 = \mu_3$; $P < 0.01$

EJERCICIOS 10.2

1. *a)*

$\bar{x}_{2\cdot}$	$\bar{x}_{1\cdot}$	$\bar{x}_{4\cdot}$	$\bar{x}_{3\cdot}$	$\bar{x}_{5\cdot}$
38.7	46.0	50.0	51.3	60.0

b) $\binom{6}{2} = 15$ *c)* $\binom{10}{2} = 45$

3. $\alpha' \leq 1 - (1 - 0.05)^3 = 0.1426; 0.5367; 0.9006$

5. *a)* ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	*2*	225.625	112.8125	4.9471
Error	21	478.875	22.8036	
Total	23			

$P = P[F_{2,21} > 4.9471]$; $0.01 < P < 0.025$; se rechaza H_0: $\mu_1 = \mu_2 = \mu_3$

b) $\frac{b}{c} = \frac{0.15}{3} = 0.05$

$|\bar{x}_{1\cdot} - \bar{x}_{2\cdot}|$ debe exceder $2.080\sqrt{22.8036\left(\frac{1}{8} + \frac{1}{9}\right)} = 4.83$

$|\bar{x}_{1\cdot} - \bar{x}_{3\cdot}|$ debe exceder $2.080\sqrt{22.8036\left(\frac{1}{8} + \frac{1}{7}\right)} = 5.14$

$|\bar{x}_{2\cdot} - \bar{x}_{3\cdot}|$ debe exceder $2.080\sqrt{22.8036\left(\frac{1}{9} + \frac{1}{7}\right)} = 5.01$

$\bar{x}_{2\cdot}$	$\bar{x}_{3\cdot}$	$\bar{x}_{1\cdot}$
16.33	22.0	23.125

7.

P	2	3
r_p	4.024	4.197
SSR_p	0.685	0.715

g.l. = 21 (usar 20)
MSE = 0.232
n = 8

$\bar{x}_1$	$\bar{x}_3$	$\bar{x}_2$
−0.9	0	0.2

9.

P	2	3	4
r_p	3.889	4.506	4.168
SSR'_p	0.135	0.156	0.144

g.l. = 36 (usar 30)
MSE = 0.012
n = 10

$\bar{x}_{5\cdot}$	$\bar{x}_{4\cdot}$	$\bar{x}_{2\cdot}$	$\bar{x}_{3\cdot}$
0.491	0.656	1.604	1.623

11. *a)* ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	*2*	189.22	94.61	0.44
Error	15	3193.06	212.87	
Total	17			

no es posible rechazar H_0: $\mu_1 = \mu_2 = \mu_3$

b) no aplicable

EJERCICIOS 10.3

1. ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	3	586.6	195.53	2.90
Error	76	512.2	67.42	
Total	79	5710.8		

$P = P[F_{3,76} > 2.90]$; $0.025 < P < 0.05$; se rechaza H_0: $\sigma^2_{Tr} = 0$

EJERCICIOS 10.4

1. a) no interacción; la diferencia del bloque 1 al bloque 2 es 2 para cada tratamiento, y la diferencia del bloque 2 al bloque 3 es -2 para cada tratamiento.
   b) interacción; para el tratamiento *A* la diferencia entre los bloques 1 y 2 es 3, mientras que para el tratamiento *C* la diferencia es 5.
   c) interacción; para el tratamiento *A* la diferencia entre el bloque 1 y el bloque 2 es 3, mientras que la diferencia es 2 para el tratamiento *B*.

3. a) Ya que $RE > 1$, la técnica de dividir en bloques es efectiva. El diseño completamente aleatorio requiere dos veces más observaciones que el de bloques completamente aleatorizados, para funcionar igual de bien.
   b) La construcción de bloques es efectiva, requiere 10 veces más observaciones.
   c) No es efectiva, no hay que dividir en bloques en el futuro.
   d) No es efectiva, no hay que dividir en bloques en el futuro.
   e) Los diseños son equivalentes.

5. *a)* $T_{1\cdot} = 1872.3$ $\quad \bar{X}_{1\cdot} = 936.15$ $\quad T_{\cdot 1} = 3967.9$ $\quad \bar{X}_{\cdot 1} = 991.975$
   $T_{2\cdot} = 1618.9$ $\quad \bar{X}_{2\cdot} = 809.45$ $\quad T_{\cdot 2} = 3183.5$ $\quad \bar{X}_{\cdot 2} = 795.875$
   $T_{3\cdot} = 1781.3$ $\quad \bar{X}_{3\cdot} = 890.65$
   $T_{4\cdot} = 1878.9$ $\quad \bar{X}_{4\cdot} = 939.45$

   $T_{\cdot\cdot} = 7151.4$ $\quad \bar{X}_{\cdot\cdot} = 893.925$ $\quad \sum_{i=1}^{4}\sum_{j=1}^{2} X_{ij}^2 = 6\,501\,860.16$

   N=S $\quad T_{\cdot\cdot}^2/N = 6\,392\,815.245$

   b) ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	3	22 004.455	7334.82	2.17
Bloque	1	76 910.42	76 910.42	
Error	3	10 130.04	3376.68	
Total	7	109 044.95		

   $P = P[F_{3,3} > 2.17]$; $P > 0.10$; imposible rechazar H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} = \mu_{4\cdot}$

   c) $C = 2(3)/7 = 0.857$; $RE = 0.857 + 0.143(22.78) = 4.11$; los bloques parecen ser útiles

7.

p	2	3	4
r_p	3.956	4.126	4.239
SSR_p	1.697	1.769	1.818

$g.l. = 27$ (usar 24)
$MSE = 1.84$
$n = b = 10$

$\bar{x}_{3.}$	$\bar{x}_{4.}$	$\bar{x}_{2.}$	$\bar{x}_{1.}$
19.9	51.1	59.5	89.0

9. *Protección estándar*

$T_{1.} = 74$ $\quad \bar{X}_{1.} = 4.625$ $\quad T_{..} = 159$
$T_{2.} = 39$ $\quad \bar{X}_{2.} = 2.4375$ $\quad N = 48$
$T_{3.} = 46$ $\quad \bar{X}_{3.} = 2.875$ $\quad T^2/N = 526.6875$
$T_{.1} = 7.5$ $\quad T_{.9} = 15.0$ $\quad \sum_{i=1}^{3}\sum_{j=1}^{16} X_{ij}^2 = 879.5$
$T_{.2} = 16.5$ $\quad T_{.10} = 14.5$
$T_{.3} = 11.0$ $\quad T_{.11} = 2.5$
$T_{.4} = 4.5$ $\quad T_{.12} = 15.5$
$T_{.5} = 15.0$ $\quad T_{.13} = 5.5$
$T_{.6} = 7.0$ $\quad T_{.14} = 14.5$
$T_{.7} = 16.5$ $\quad T_{.15} = 2.0$
$T_{.8} = 6.0$ $\quad T_{.16} = 5.5$

ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	2	42.875	21.4375	3.848
Bloque	15	142.8125	9.5208	
Error	30	167.125	5.5708	
Total	47	352.8125		

$P = P[F_{2,30} > 3.848]$; $0.01 < P < 0.025$; rechaza H_0: $\mu_{1.} = \mu_{2.} = \mu_{3.}$

Pueden realizarse $\binom{3}{2} = 3$ contraste T de tipo Bonferroni. Si cada uno se lleva a cabo con $\alpha = 0.05$ entonces $\alpha' \leq 1 - (0.95)^3 = 0.1426$.

H_0: $\mu_{1.} = \mu_{2.}$, $\quad H_1$: $\mu_{1.} \neq \mu_{2.}$; $t = \dfrac{\bar{d}}{\sqrt{S_d/n}} = 2.39$; $P = 0.03 < 0.05$ **(hallado con la TI83)**

Se rechaza H_0 y concluimos que la flexibilidad fue diferente antes y después de la protección.

H_0: $\mu_{1.} = \mu_{3.}$, H_1: $\mu_{1.} \neq \mu_{3.}$; $t = 2.04$; $P = 0.059 > 0.05$ (hallado con la TI83)

No se puede rechazar H_0: no pueden detectarse diferencias en la flexibilidad después de la protección y después de correr.

H_0: $\mu_{2.} = \mu_{3.}$, H_1: $\mu_{2.} \neq \mu_{3.}$; $t = -0.607$; $P = 0.55 > 0.05$ (hallado con la TI83)

No se puede rechazar H_0; no pueden detectarse diferencias en la flexibilidad después de la protección y después de correr.

Protección reforzada

$T_{1\cdot} = 77$ $\quad \bar{X}_{1\cdot} = 4.8125$ $\quad T_{\cdot\cdot} = 168.5$
$T_{2\cdot} = 37$ $\quad \bar{X}_{2\cdot} = 2.3125$ $\quad N = 48$
$T_{3\cdot} = 54.5$ $\quad \bar{X}_{3\cdot} = 3.40625$ $\quad T_{\cdot\cdot}^2/N = 591.51$
$T_{\cdot 1} = 13.5$ $\quad T_{\cdot 9} = -7.0$ $\quad \sum_{i=1}^{3}\sum_{j=1}^{16} X_{ij}^2 = 10$
$T_{\cdot 2} = 13.0$ $\quad T_{\cdot 10} = 4.0$
$T_{\cdot 3} = 12.5$ $\quad T_{\cdot 11} = 5.5$
$T_{\cdot 4} = 16.0$ $\quad T_{\cdot 12} = 22.5$
$T_{\cdot 5} = 10.5$ $\quad T_{\cdot 13} = 11.0$
$T_{\cdot 6} = 9.0$ $\quad T_{\cdot 14} = 11.5$
$T_{\cdot 7} = 17.0$ $\quad T_{\cdot 15} = 0.5$
$T_{\cdot 8} = 19.0$ $\quad T_{\cdot 16} = 10.0$

ANOVA

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	***2***	50.256	25.128	4.377
Bloque	15	261.74	17.449	
Error	30	172.244	5.741	
Total	47	484.24		

$P = P[F_{2,30} > 4.377]$; $0.01 < P < 0.025$; se rechaza H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$
H_0: $\mu_{1\cdot} = \mu_{2\cdot}$, H_1: $\mu_{1\cdot} \neq \mu_{2\cdot}$; $t = 3.03$; $P = 0.008 < 0.05$
Rechazamos H_0 y concluimos que hay diferencia entre la flexibilidad antes y después de la protección. H_0: $\mu_{1\cdot} = \mu_{3\cdot}$, H_1: $\mu_{1\cdot} \neq \mu_{3\cdot}$; $t = 1.63$; $P = 0.123 > 0.05$
No se pueden detectar diferencias en la flexibilidad antes de la protección y después de correr. H_0: $\mu_{i\cdot} = \mu_{3\cdot}$, H_1: $\mu_{2\cdot} \neq \mu_{3\cdot}$; $t = -1.28$; $P = 0.219 > 0.05$
No se pueden detectar diferencias en la flexibilidad después de la protección y después de correr. Los resultados para la protección estándar y reforzada son similares.

EJERCICIOS 10.5

1. *b)* $T_{\cdot\cdot}^2/N = (3366)^2/36 = 314\,721$; $SS_{\text{Total}} = 12\,710.42$
   c) $SS_{\text{Tr}} = 12\,489$; $SS_E = SS_{\text{Total}} - SS_{\text{Tr}} = 221.42$
   d) $SS_A = 10\,842$; $SS_B = 1225$; $SS_{AB} = = SS_{\text{Tr}} - SS_A - SS_B = 422$

   e/f) ANOVA

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	5	12 489	2 497.8	338.42
A	2	10 842	5 421	734.48
B	1	1225	1225	165.97
AB	2	422	211	28.59
Error	30	221.42	7.3807	
Total	35	12 710.42		

$\mathbf{H_0}$: $(\alpha\beta)_{ij} = 0$ (no existe interacción)
$\mathbf{H_1}$: $(\alpha\beta)_{ij} \neq 0$ (existe interacción)

Rechazamos H_0 con $P < 0.01$. Concluimos por tanto que existe interacción. En tal caso, compararemos los niveles del factor *A* en cada nivel del factor *B*.

ANOVA

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	*2*	7564	3782	460.097 *$P < 0.01$
Error	15	123.3	8.22	
Total	17			

Se rechaza $\mathbf{H_0}$: $\mu_{11\cdot} = \mu_{21\cdot} = \mu_{31\cdot}$ y $\mathbf{H_1}$: $\mu_{i1\cdot} \neq \mu_{k1\cdot}$ para algún *i y k*

p	2	3
r_p	4.168	4.347
SSR_p	4.879	5.088
$\bar{x}_{11\cdot}$	$\bar{x}_{21\cdot}$	$\bar{x}_{31\cdot}$
73.0	102.0	123.0

$g.l. = 15$
$MSE = 8.22$
$n = 6$

ANOVA

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	*2*	3700	1850	282.82 *$P < 0.01$
Error	15	98.12	6.54	
Total	17			

Se rechaza $\mathbf{H_0}$: $\mu_{12\cdot} = \mu_{22\cdot} = \mu_{32\cdot}$ y $\mathbf{H_1}$: $\mu_{i2\cdot} \neq \mu_{k2\cdot}$ para algún *i y k*

p	2	3
r_p	4.168	**4.347**
SSR_p	4.35	**4.54**
$\bar{x}_{12\cdot}$	$\bar{x}_{22\cdot}$	$\bar{x}_{32\cdot}$
71.0	86.0	106.0

$df = 15$
$MSE = 6.54$
$n = 6$

3. *b)* $T^2_{..}/N = (859)^2/20 = 36\,894.05$; $SS_{\text{Total}} = 953.21$
c) $SS_{\text{Tr}} = 471.45$; $SS_E = 481.76$
d) $SS_A = 151.25$; $SS_B = 0.2$; $SS_{AB} = 320.0$
e) sí; hay un cruce

f) **ANOVA**

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	3	471.45	157.15	5.22
A	1	151.25	151.25	5.02
B	1	0.20	0.20	0.007
AB	1	320.0	320.0	10.63
Error	16	481.76	30.11	
Total	19	953.21		

H_0: $(\alpha\beta)_{ij} = 0$ (no existe interacción)
H_1: $(\alpha\beta)_{ij} \neq 0$ (existe interacción)
$P = P[F_{1,16} > 10.63]$; se rechaza H_0 con $P < 0.01$
H_0: $\mu_{11\cdot} = \mu_{21\cdot}$, H_1: $\mu_{11\cdot} \neq \mu_{21\cdot}$; $f = 9.375$; $0.01 < P < 0.025$ se rechaza H_0
H_0: $\mu_{12\cdot} = \mu_{22\cdot}$, H_1: $\mu_{12\cdot} \neq \mu_{22\cdot}$; $f = 1.34$; $P > 0.10$; no es posible rechazar H_0

5. a) **ANOVA**

Fuente	***DF***	***SS***	***MS***	***F***
Tratamiento	5	1 201 750.417	240 350.08	87.96
A	2	102 443.17	51 221.58	18.75
B	1	1 020 250.08	1 020 250.08	373.39
AB	2	79 057.167	39 528.58	14.47
Error	6	16 394.48	2 732.41	
Total	11	1 218 144.9		

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2} X_{ijk}^2 = 3\,175\,929$$

$$T_{\cdot\cdot\cdot}^2/N = (4847)^2/12 = 1\,957\,784.083$$

b) H_0: $(\alpha\beta)_{ij} = 0$ (no existe interacción)
H_1: $(\alpha\beta)_{ij} \neq 0$ (existe interacción); se rechaza H_0 con $P < 0.01$
H_0: $\mu_{11\cdot} = \mu_{21\cdot} = \mu_{31\cdot}$ y
H_1: $\mu_{i1\cdot} \neq \mu_{j1\cdot}$ para algún i y j.

ANOVA

Se rechaza H_0 con $P = 0.0276$ (TI83)

$|X_{i1\cdot} - X_{j1\cdot}|$ debe exceder $3.182\sqrt{4989.83\ (\frac{1}{2} + \frac{1}{2})} = (224.77)$

$\bar{x}_{31\cdot}$	$\bar{x}_{11\cdot}$	$\bar{x}_{21\cdot}$
490.5	722.5	873.5

H_0: $\mu_{12\cdot} = \mu_{22\cdot} = \mu_{32\cdot}$ y H_1: $\mu_{i2\cdot} \neq \mu_{j2\cdot}$ para algún i y j.

ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	**2**	32 624.33	16 312.17	34.34
Error	**3**	1425	475	
Total	**5**			

Se rechaza $\mathbf{H_0}$ con $P < 0.01$.

$|X_{i2\cdot} - X_{j2\cdot}|$ debe exceder $3.182\sqrt{475\ (\frac{1}{2}+\frac{1}{2})} = 69.35$

$\bar{x}_{32\cdot}$	$\bar{x}_{22\cdot}$	$\bar{x}_{12\cdot}$
56.0	64.5	216.5

7. No se ha encontrado interacción. Se han hallado diferencias entre los lugares. No se han hallado diferencias entre las semanas.

EJERCICIOS 11.1

1. sí; la nube de puntos muestra una tendencia lineal. **3.** cuestionable

EJERCICIOS 11.2

1. $\sum x = 36.2 \quad \sum y = 85.6 \quad \bar{x} = 2.59$

$\sum x^2 = 105.66 \quad \sum xy = 244.8 \quad \bar{y} = 6.11$

$$b = \frac{14(244.8) - 36.2(85.6)}{14(105.66) - (36.2)^2} = \frac{328.48}{168.80} = 1.95$$

$a = 6.11 - 1.95(2.59) = 1.06$; $\mu_{Y|x} = 1.06 + 1.95x$;

$\mu_{Y|x=3.7} = 1.06 + 1.95(3.7) = 8.275$; $\hat{y} = 8.275$

3. $$b = \frac{106(75\,989.6) - 366.1(12\,623)}{106(2435.63) - (366.1)^2} = \frac{3\,433\,617.3}{124\,147.57} = 27.66$$

$a = 119.08 - 27.66(3.45) = 23.65$; $\hat{\mu}_{Y|x} = 23.65 + 27.66x$

$\hat{y} = 23.65 + 27.66(5.5) = 175.78$

no; el 16 está por encima de los valores usados para generar la línea de regresión

5. *a)* $\sum x = 56.6 \quad \sum y = 151.1 \quad \sum x^2 = 117.68 \quad \sum xy = 311.96$ *b)* $b = 1.996 \quad a = 1.361 \quad \hat{\mu}_{Y|x} = 1.361 + 1.996x$

c) $\hat{y} = 5.853$

7. *b)* $\hat{\mu}_{Y|x} = 7.227 - 0.03296x$ *c)* $\hat{\mu}_{Y|x=18} = \hat{y} = 6.634$

EJERCICIOS 11.3

1. ***b)*** próximo a 1

c) $\sum x = 15 \quad \sum y = 30.1 \quad \sum xy = 92.75 \quad \sum x^2 = 47.5 \quad \sum y^2 = 183.65$

d) $r \cong 0.99$; fuertemente positiva

3. *b)* próximo a 0; hay una pequeña tendencia lineal

c) $\sum x = 21 \quad \sum y = 38.2 \quad \sum xy = 114.46$
$\sum x^2 = 67.12 \quad \sum y^2 = 228.98$

d) $r \cong 0.015$; débil negativa

5. *a)* $\sum x = 2405 \quad \sum y = 2503 \quad \sum xy = 902\,475$
$\sum x^2 = 900\,775 \quad \sum y^2 = 919\,489$

b) $r \cong 0.978$

c) sí; da las mismas lecturas que el método manual y es más fácil de usar.

7. *b)* Cuando la preocupación aumenta, hay una tendencia a que la depresión aumente. Cuando la satisfacción aumenta, hay una tendencia a que la depresión disminua. Cuando la preocupación aumenta, hay una tendencia a que la satisfacción disminuya.

c) $H_0: \rho = 0, \quad H_1: \rho \neq 0$

EJERCICIOS 11.4

1. ***b)*** $\sum x = 1776 \quad \sum y = 3018 \quad \sum xy = 549\,705$
$\sum x^2 = 322\,062 \quad \sum y^2 = 941\,056 \quad r \cong 0.967$

c) $r^2 \cong 0.9351$; **el 93.51 %** de la variabilidad en el mejor levantamiento en cuanto limpieza y empuje está asociada con el peso corporal.

d) ANOVA

Fuente	*DF*	*SS*	*MS*	*F*
Tratamiento	1	28 280	28 280	116.4 *P< 0.01
Error	8	1 943.6	242.95	
Total	9	30 223.6		

$$S_{xy} = \frac{10(549\,705) - 1776(3018)}{10} = 13\,708.2$$

$$S_{xx} = \frac{10(322\,062) - (1776)^2}{10} = 6644.4$$

$$b = \frac{S_{xy}}{S_{xx}} = 2.063$$

$$a = -64.61$$

$$S_{yy} = \frac{10(941\,056) - (3018)^2}{10} = 30\,223.6$$

$$\hat{\mu}_{y|x} = -64.61 + 2.063x$$

$\hat{y} = 347.39$ libras

5. $r^2 \cong (0.978)^2 = 0.956$; sí
7. 0.09; 0.1296; 0.0256

EJERCICIOS 11.5

1. a) sí

   b) ANOVA

Fuente	*DF*	***SS***	*MS*	*F*
Tratamiento	1	4 269.44	4 269.44	**11**.26 ***P*< 0.01**
Error	9	3411.11	379.01	
Total	10	7 680.55		

$\sum x = 548$ $\sum y = 620$ $\sum xy = 28\,895$ $\bar{x} = 49.818$

$\sum x^2 = 28\,230$ $\sum y^2 = 42\,626$ $b = -2.143$

$S_{xy} = -1992.27$ $S_{xx} = 929.64$ $S = \sqrt{379.01} = 19.468$

   c) $a = 163.127$; $b = -2.143$

   d) $163.127 \pm 2.262\left[\dfrac{19.468\sqrt{28\,230}}{\sqrt{11(929.64)}}\right]$ o 163.127 ± 73.17; 95% de confianza en que α se encuentre en el intervalo (89.957; 236.297)

   $-2.143 \pm 2.262\left[\dfrac{19.468}{\sqrt{929.64}}\right]$ o -2.143 ± 1.444; 95% de confianza en que β se encuentre en el intervalo (-3.587, -0.699)

   e) $\hat{\mu}_{Y|x} = 163.127 - 2.143(35) =$ 88.122 días

   $88.122 \pm 2.262(19.468)\sqrt{\dfrac{1}{11} + \dfrac{(35 - 49.818)^2}{929.64}}$ o 88.122 ± 25.186; 95% de confianza en que $\mu_{Y|x=35}$ se encuentre en el intervalo (62.936, 113.308)

   f) $88.122;\ 88.122 \pm 2.262\,(19.468)\sqrt{1 + \dfrac{1}{11} + \dfrac{(35 - 49.818)^2}{929.64}}$ o 88.122 ± 50.73; 95% de confianza en que $Y|_{x=35}$ se encuentre en el intervalo (37.392, 138.852)

EJERCICIOS 11.6

1. $\hat{\mu}_{Y|x} = \hat{y} = 54.079 + 0.097(233) + 0.034(260) + 0.522(82) - 2.655(80) + 2.559(88) = 141.116$

3. disminuye la estimación en 2.655

5. a) 73 % de la variación en la porosidad es explicable por la asociación lineal con los tres regresores

   6) 0.0018,0.018

   c) -0.72(20) = -14.40

7. $2^5 - 1 = 31$

EJERCICIOS 12.1

1. La tabla dispuesta de otra forma en la que las filas son fijas.

Hepatitis			
Vacunados	Sí	No	
Sí	11 (41.06)	538 (507.94)	549 (fijo)
No	70 (39.94)	464 (494.06)	534 (fijo)
	81	1002	1083

$\chi_1^2 = 48.24$; $P < 0.05$; hay evidencia de que la proporción de vacunados con hepatitis no es la misma que la proporción de no vacunados con la enfermedad.

3.

Rubéola			
Cataratas	Sí	No	
Sí	14 (10.67)	6 (9.33)	20 (fijo)
No	10 (13.33)	15(11.67)	25 (fijo)
	24	21	

$\chi_1^2 = 4.018$; $0.025 < P < 0.05$; hay evidencia de que la proporción de los niños con cataratas cuya madre tuvo rubéola es diferente de la proporción de aquellos cuya madre no tuvo la enfermedad.

5.

Alergia presente			
Leucemia	Sí	No	
No	5 (10.39)	12 (6.61)	17 (fijo)
Sí	17(11.61)	2 (7.39)	19 (fijo)
	22	14	36

$\chi_1^2 = 13.62$; $P < 0.005$; la proporción de los que tienen alergia difiere entre los que tienen y los que no tienen leucemia.

7.

Edad			
Alérgico	3 ó menos	Más de 3	
Sí	32 (28.34)	30 (33.66)	62
No	80 (83.66)	103 (99.34)	183
	112	133	245

$\chi_1^2 = 1.16$; $P > 0.10$; no hay asociación entre la edad y la alergia a los huevos; independencia.

9. $\hat{p}_1 = \frac{52}{300} = 0.173;\ \hat{p}_2 = \frac{48}{320} = 0.150;\ \hat{p} = \frac{52+48}{620} = 0.1613$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.173 - 0.150}{\sqrt{0.1613(0.8387)\left(\frac{1}{300}+\frac{1}{320}\right)}} \cong 0.778$$

$(0.778)^2 = 0.605$; la diferencia entre Z^2 y χ_1^2 es debida a los redondeos.

EJERCICIOS 12.2

1. Pacientes: $\hat{p}_A = \frac{472}{1301} = 0.363$ Controles: $\hat{p}_A = \frac{2625}{6313} = 0.415$

$\hat{p}_B = \frac{102}{1301} = 0.078$ $\hat{p}_B = \frac{570}{6313} = 0.090$

$\hat{p}_{AB} = \frac{29}{1301} = 0.022$ $\hat{p}_{AB} = \frac{226}{6313} = 0.036$

Las proporciones ponen de manifiesto diferencias para todos los grupos sanguíneos excepto, tal vez, para el grupo B.

3. *a)* no *b)* independencia c) H_0: fragancia y color son independientes
H_1: fragancia y color no son independientes

d)

	Color de la flor			
Fragancia	**Blanca**	**Rosa**	**Naranja**	
Sí	**12** (40.3)	**60** (45.5)	**58 (44.2)**	**130**
No	**50** (21.7)	**10 (24.5)**	**10** (23.8)	**70**
	62	**70**	**68**	**200**

$\chi_2^2 = 82.29$; $P < 0.005$; se concluye que el color y la fragancia no son independientes. Dado que una flor tiene fragancia hay un 9.23 % de probabilidad de que sea blanca, un 46.15 % de que sea rosa y un 44.62 % de que sea naranja.

5. *a)* contraste de homogeneidad

b)

	Nivel de cloroplastos			
Nivel de SO_2	**Alto**	**Normal**	**Bajo**	
Alto	**3(5)**	**4** (8.33)	**13** (6.67)	**20**
Normal	**5(5)**	**10** (8.33)	**5** (6.67)	**20**
Bajo	**7(5)**	**11** (8.33)	**2** (6.67)	**20**
	15	**25**	**20**	**60**

$\chi_4^2 = 14.74$; $0.005 < P < 0.01$; se rechaza H_0; concluimos que el nivel de cloroplastos se ve afectado por el nivel de SO_2.

c) $\hat{p}_H = \frac{13}{20} = 0.65;\ \hat{p} = \frac{2}{20} = 0.10;\ \hat{p} = \frac{5}{20} = 0.25$

Se pone de manifiesto que el SO_2 tiende a disminuir el nivel de cloroplastos.

7.

	Especies							
Localización	A	B	C	D	E	F	G	
Arriba	37 (34.70)	12(11.31)	6 (9.80)	18(13.58)	7 (5.28)	6 (9.05)	0 (2.26)	86
Abajo	9(11.30)	3 (3.69)	7 (3.20)	0 (4.42)	0 (1.72)	6 (2.95)	3 (0.74)	28
	46	15	13	18	7	12	3	114

$\chi^2_6 = 28.35$; $P < 0.005$.
Estos datos no verifican los requisitos ya que hay una celda con frecuencia esperada menor que 1 y varias menores que 5. Se necesitan más datos para que el contraste sea válido. Sin embargo, de los datos se deduce que es probable que la planta está teniendo un efecto adverso con un impacto grande en las especies D y E.

EJERCICIOS 13.1

1. *a)* no se puede rechazar H_0; los datos no refutan la suposición de normalidad

   b) $t = 1.149$; $P > 0.10$

   Puesto que $P > 0.05$, H_0 no se rechaza al nivel $\alpha = 0.05$.

   c) $\chi^2_{14} = \dfrac{(0.000008)}{(0.0025)^2} = 17.94$; $0.10 < P < 0.25$

   Puesto que $P > 0.05$, H_0 no se rechaza al nivel $\alpha = 0.05$.

3. *b)* $t = 1.22$; $0.10 < P < 0.25$

   Puesto que $P > 0.10$, H_0 no se rechaza al nivel $\alpha = 0.10$.

   c) $\chi^2_{19} = \dfrac{20(104.93)^2}{(150)^2} = 9.79$; $0.025 < P < 0.05$

   Puesto que $P < 0.10$, H_0 se rechaza al nivel $\alpha = 0.10$.

EJERCICIOS 13.2

1. H_0: $M \geq 22$, H_1: $M < 22$
   Si H_0 es cierta $E[N'] = 10$; si H_1 es cierta, habrá muy pocos signos positivos.
   $N' = 5$; $P = P[N' \leq 5 | p = \frac{1}{2}] = 0.0207$; se rechaza H_0

3. H_0: $M \geq 9$, H_1: $M < 9$
   $E[N'] = 10$; $P = P[N' \leq 1 | p = \frac{1}{2}] \cong 0.0000$; rechazamos H_0
   Sí, porque P es extremadamente pequeño.

5. $W_+ = 42.5$, $|W_-| = 12.5$

   $W_+ + |W_-| = 55 = \dfrac{10(11)}{2}$

7.

x_i	0.71	0.13	0.16	0.65	0.21	1.00	0.51
$x_i - 8$	-0.09	-0.67	-0.64	-0.15	-0.59	0.2	-0.29
$\|x_i - 8\|$	0.09	0.67	0.64	0.15	0.59	0.2	0.29
Rango	1.5	13	12	3	10.5	4	5
Rango de signos	-1.5	-13	-12	-3	-10.5	4	-5

x_i	1.1	1.11	0.32	0.71	0.4	0.21	1.63
$x_i - 8$	0.3	0.31	-0.48	-0.09	-0.4	-0.59	0.83
$\|x_i - 8\|$	0.3	0.31	0.48	0.09	0.4	0.59	0.83
Rango	6	7	9	1.5	8	10.5	14
Rango de signos	6	7	- 9	-1.5	-8	-10.5	14

$\mathbf{H_0}$: $M \geq 0.8$, $\mathbf{H_1}$: $M < 0.8$
Si H_1 es cierta, esperamos muchos signos positivos y muy pocos negativos. El estadístico de contraste es W_+
$P = P[W_+ \leq 31]$; $P > 0.05$; no se rechaza $\mathbf{H_0}$ al nivel $\alpha = 0.05$ porque 31 > 26.

9.

x_i	35.5	44.5	39.8	33.3	51.4	51.3	30.5	48.9	42.1
$x_i - 45$	-9.5	-0.5	-5.5	-11.7	6.4	6.3	-14.5	3.9	-2.9
$\|x_i - 45\|$	9.5	0.5	5.5	11.7	6.4	6.3	14.5	3.9	2.9
Rango	22	2	13	24	16	15	25	10	9
Rango de signos	-22	-2	-13	-24	16	15	-25	10	-9

x_i	40.3	46.8	38	40.1	36.8	39.3	65.4	42.6	42.8
$x_i - 45$	-4.7	1.8	-7	-4.9	-8.2	-5.7	20.4	-2.4	-2.2
$\|x_i - 45\|$	4.7	1.8	7	4.9	8.2	5.7	20.4	2.4	2.2
Rango	11	5	17	12	20	14	30	8	6
Rango de signos	-11	5	-17	-12	-20	-14	30	-8	-6

x_i	59.8	52.4	26.2	60.9	45.6	27.1	47.3	36.6
$x_i - 45$	14.8	7.4	-18.8	15.9	0.6	-17.9	2.3	-8.4
$\|x_i - 45\|$	14.8	7.4	18.8	15.9	0.6	17.9	2.3	8.4
Rango	26	19	29	27	3	28	7	21
Rango de signos	26	19	-29	27	3	-28	7	-21

x_i	55.6	45.1	52.2	43.5
$x_i - 45$	10.6	0.1	7.2	-1.5
$\|x_i - 45\|$	10.6	0.1	7.2	1.5
Rango	23	1	18	4
Rango de signos	23	1	18	-4

$W_+ = 200$, $|W_-| = 265$; $P = P[W_+ \leq 200] > 0.10$ porque 200 > 152; Tipo II

11. *a)* $E[W_+] = \frac{70(71)}{4} = 1242.5$; Var $W_+ = \frac{70(71)(141)}{24} = 29\,198.75$;

$P[W_+ \leq 1000] = P[Z \leq -1.42] = 0.0778$

b) $E[|W_-|] = \frac{80(81)}{4} = 1620$; Var $|W_-| = \frac{80(81)(161)}{24} = 4347$;

$P[|W_-| \leq 1500] = P[Z \leq -0.58] = 0.2810$

EJERCICIO 13.3

1. *a)* número de signos positivos
   b) $P = P[N' \leq 3 | p = \frac{1}{2}] = 0.1719$; no se rechaza $\mathbf{H_0}$

3. $\mathbf{H_0}$: $M_{X,Y} \leq 0$, $\mathbf{H_1}$: $M_{X,Y} > 0$; $P = P[N \leq 1 | p = \frac{1}{2}] = 0.0107$; se rechaza $\mathbf{H_0}$

5. *a)*

d_i	9.05	10.52	1.86	0.11	3.04	9	2.3	-4.49	4.12	-5.36
$\|d_i\|$	9.05	10.52	1.86	0.11	3.04	9	2.3	4.49	4.12	5.36
Rango	9	10	2	1	4	8	3	6	5	7
Rango de signos	9	10	2	1	4	8	3	-6	5	-7

$W_+ = 42$, $|W_-| = 13$; $P = P[|W_-| \leq 13] > 0.05$ porque $13 > 11$; no se rechaza H_0 al nivel $\alpha = 0.05$

b) $P = P[N \leq 2 | p = \frac{1}{2}] = 0.0547$; no se rechaza al nivel $\alpha = 0.05$ porque $P > 0.05$

EJERCICIOS 13.4

1.

Observación	0.09	0.12	0.13	0.17	0.19	0.19	
Grupo	U	U	U	U	U	U	
Rango	1	2	3	4	5.5	5.5	
Observación	0.2	0.2	0.21	0.21	0.21	0.22	0.23
Grupo	U	M	U	M	M	M	M
Rango	7.5	7.5	10	10	10	12	13

$W_m = 58$; rechazamos si W_m es demasiado grande; $P < 0.025$; rechazamos H_0 y concluimos que el contenido total de proteinas tiende a ser mayor entre ratas privadas de NGF en el útero que entre las que fueron privadas del factor de crecimiento en la leche.

3.

Observación	604.1	646.8	688.1	739.4	760.5	793.5	
Grupo	<	<	<	≥	<	<	
Rango	1	2	3	4	5	6	
Observación	797.0	806.8	812.4	818.9	843.6	850.0	856.6
Grupo	≥	<	≥	≥	≥	≥	<
Rango	7	8	9	10	11	12	13
Observación	899.1	906.5	909.3	940.7	961.8	968.1	979.1
Grupo	<	<	≥	≥	≥	<	≥
Rango	14	15	16	17	18	19	20

Observación	1009.9	1100.6	1330.3	1335.8
Grupo	≥	≥	≥	≥
Rango	21	22	23	24

W_m = suma de los rangos del grupo <; se rechaza si W_m es demasiado pequeño; $W_m = 86$; se rechaza con $P < 0.025$ (el punto crítico es 91)

5. *a)* $E[W_m] = \dfrac{30(91)}{2} = 1365$; Var $W_m = \dfrac{30(60)(91)}{12} = 13\,650$;

$P[W_m \leq 1350] = P[Z \leq -0.13] = 0.4483$

b) $E[W_m] = \dfrac{40(91)}{2} = 1820$; Var $W_m = \dfrac{40(50)(91)}{12} = 15\,166.67$;

$P[W_m \geq 2000] = 1 - P[Z \leq 1.46] = 0.0721$

EJERCICIOS 13.5

1. *a)* $R_1 = 72.5$, $R_2 = 170$, $R_3 = 163.5$

b) $H = \dfrac{12}{28(29)}\left[\dfrac{(72.5)^2}{8} + \dfrac{(170)^2}{10} + \dfrac{(163.5)^2}{10}\right] - 3(29) = 4.92$;

$P = P[X_2^2 \geq 4.92]$; $0.05 < P < 0.10$

3. $R_1 = 35$, $R_2 = 23$, $R_3 = 102$, $R_4 = 103$, $R_5 = 62$

$H = \dfrac{12}{25(26)}\left[\dfrac{(35)^2}{5} + \dfrac{(23)^2}{5} + \dfrac{(102)^2}{5} + \dfrac{(103)^2}{5} + \dfrac{(62)^2}{5}\right] - 3(26)$

$H = 20.256$; $P = P[X_4^2 \geq 20.256]$; $P < 0.005$

5. $R_1 = 39$, $R_2 = 163$, $R_3 = 98$

$H = \dfrac{12}{24(25)}\left[\dfrac{(39)^2}{8} + \dfrac{(163)^2}{8} + \dfrac{(98)^2}{8}\right] - 3(25)$

$H = 19.235$; $P = P[X_2^2 \geq 19.235]$; $P < 0.005$

EJERCICIOS 13.6

1. $R_1 = 18$, $R_2 = 21.5$, $R_3 = 25$, $R_4 = 15.5$

$\dfrac{b(k+1)}{2} = \dfrac{8(5)}{5} = 20$

$S = [(18 - 20) + (21.5 - 20) + (25 - 20) + (15.5 - 20)]^2 = 51.5$

$X_3^2 = \dfrac{12S}{bk(k+1)} = \dfrac{12(51.5)}{8(4)5} = 3.8625$; $P = P[\chi_3^2 \geq 3.8625]$; $0.25 < P < 0.50$;

no se rechaza H_0

3. $R_1 = 25$

$\dfrac{b(k+1)}{2} = \dfrac{8(6)}{2} = 24$

$R_2 = 17$
$R_3 = 25$
$R_4 = 25$
$R_5 = 28$

$S = 68;\ X_4^2 = \dfrac{12(68)}{40(6)} = 3.4;\ P = P[\chi_4^2 \geq 3.4];\ 0.25 < P < 0.50;$ no se rechaza $\mathbf{H_0}$

EJERCICIOS 13.7

1. *b)* $\sum r_x = \sum r_y = 300; \quad \sum r_x^2 = \sum r_y^2 = 4900; \quad \sum r_x r_y = 3196;$

$$r_s = \frac{24(3196) - (300)^2}{\sqrt{[24(4900) - (300)^2][24(4900) - (300)^2]}} = -0.4817$$

c) $\sum d^2 = 3408;\ r_s = 1 - \dfrac{6(3408)}{24(575)} = -0.4817$

d) la diferencia es muy pequeña

e) $r_s^2 = 0.2320$; correlación negativa débil

3. $\sum r_x^2 = \sum r_y^2 = 649.5; \quad \sum r_x = \sum r_y = 78; \quad \sum r_x r_y = 588.25;\ r_s = 0.57;$

Hay una correlación positiva moderada entre el contenido de alcohol en sangre y el porcentaje de disminución del máximo de la velocidad de persecución suave.

EJERCICIOS 13.8

1. $n_1 = 8,\ n_2 = 10,\ n_3 = 10,\ N = 28$

$s_1^2 = 18.2124, \quad s_2^2 = 22.3162, \quad s_3^2 = 17.5645, \quad K = 3$

$$s_p^2 = \frac{7(18.2124) + 9(22.3162) + 9(17.5645)}{25} = 19.4565$$

$\log s_p^2 = 1.2891 \quad Q = 25(1.2891) - [7(1.2604) + 9(1.3486) + 9(1.2446)]$

$\log s_1^2 = 1.2604 \quad Q = 0.0659$

$\log s_2^2 = 1.3486$

$\log s_3^2 = 1.2446$

$$h = 1 + \frac{1}{3(2)}\left[\left(\frac{1}{7} + \frac{1}{9} + \frac{1}{9}\right)^{-1/25}\right] = 1.3584$$

$$B = \frac{2.3026(0.0659)}{1.3584} = 0.1117;\ P = P[X_2^2 \geq 0.1117];\ P > 0.90;$$ no se rechaza $\mathbf{H_0}$

EJERCICIOS 13.9

1. a) $P[X \leq 3] \cong P\left[Z \leq \dfrac{3.5 - 6}{\sqrt{4.2}}\right] = P[Z \leq -1.22] = 0.1112$(0.1071 binomial)

b) $P[3 \le X \le 6] \cong P\left[\frac{2.5-6}{\sqrt{4.2}} \le Z \le \frac{6.5-6}{\sqrt{4.2}}\right] = P[-1.71 \le Z \le 0.24] = 0.5512$
(0.5725 binomial)

c) $P[X \ge 4] = 1 - P[X \le 3] = 1 - P\left[Z \le \frac{3.5-6}{\sqrt{4.2}}\right] = P[-1.22 \le Z \le -0.73];$
0.1215(0.1304 binomial)

3. $E[X] = 30$; Var $X = 0.99997$

$$P[X \le 25] \cong P\left[Z \le \frac{25.5-30}{29.9991}\right] = P[Z \le -0.82] \cong 0.2061$$

$$P[25 \le X \le 35] \cong P\left[\frac{24.5-30}{\sqrt{29.9991}} \le Z \le \frac{35.5-30}{\sqrt{29.9991}}\right] =$$

$$= P[-1.000 \le Z \le 1.00] = 0.6826$$

5. $E[X] = 5$; Var $X = 4.9875$; $P[X \ge 1] \cong P\left[Z \ge \frac{0.5-5}{\sqrt{4.9875}}\right] = P[Z \ge -2.01] = 0.9778$

7. *a)* $P[X \ge 95] \cong P\left[Z \ge \frac{94.5-100}{10}\right] = P[Z \ge -0.55] = 0.7088$

$$P\left[Z \le \frac{80.5-100}{10}\right] = P[Z \le -1.95] = 0.0256$$

c) $P[90 \le X \le 110] \cong P\left[\frac{89.5-100}{10} \le Z \le \frac{110.5-100}{10}\right] =$

$$= P[-1.05 \le Z \le 1.05] = 0.7062$$

d) $P[X = 99] \cong P\left[\frac{98.5-100}{10} \le Z \le \frac{99.5-100}{10}\right] = P[-0.15 \le Z \le -0.05] =$

$$= 0.0397$$

9. $\lambda = 1$; $s = 0.5$; $\lambda s = 0.5$; $X =$ número de partículas asesinas por paramecio;
$P[X = 0] = 0.607$ (Tabla de Poisson)
$Y =$ número de paramecios mortales que emite una partícula mortal; suceso = emisión de una partícula mortal; $p = 0.393$
$E[Y] = 20(0.393) = 7.86$; Var Y=4.771

$$P[Y = 0] \cong P\left[\frac{-0.5-7.86}{\sqrt{4.771}} \le Z \le \frac{0.5-7.86}{\sqrt{4.771}}\right] = P[-3.83 \le Z \le 3.37] = 0.9996$$

$$P[5 \le Y \le 10] \cong P\left[\frac{4.5-7.86}{\sqrt{4.771}} \le Z \le \frac{10.5-7.86}{\sqrt{4.771}}\right] =$$

$$= P[-1.54 \le Z \le 1.21] = 0.8251$$

EJERCICIOS 13.10

1. a) 10.5 *b)* $P = P[X \ge 12 | p = 0.7] = 0.2969$; no se rechaza H_0

3. *a)* H_0: $p \le 0.20$, H_1: $p > 0.20$ ***b)*** 3.2

c) $P = P[X \ge 5 | p = 0.2] = 0.2018$; no se rechaza H_0

ÍNDICE

- Libro de **texto para los estudiantes de Estadística, tanto de Biología como de Ciencias de la Salud en general.**
- En él se hace una **exposición clara y sencilla de los conceptos** y contenidos **básicos** de la asignatura.
- Al mismo tiempo, aborda la materia con la **suficiente profundidad como para adaptarse a niveles de segundo y tercer ciclo** (optativas de segundo ciclo, doctorado, máster en investigación biomédica).
- Incluye **numerosos ejercicios y ejemplos** de los diferentes procedimientos estadísticos aplicados a las áreas biomédicas, con la solución a cada uno de ellos.

NOVEDADES DE LA EDICIÓN AMPLIADA

Se han añadido dos **ANEXOS** al final del libro, que incorporan ejercicios para el alumno con los programas de computación que generalmente se utilizan en hospitales ***(SPSS)*** y en las facultades de Medicina ***(Statgraphics).***

ISBN: 978-84-481-5996-2

9 788448 159962

The McGraw-Hill Companies

www.mcgraw-hill.es