Breaking the Law of Averages

Real-Life Probability and Statistics in Plain English

William M. Briggs

ISBN 978-0-557-01990-8

William M. Briggs. 300 E. 71st St. Apr. 3R. New York, NY, 10021. matt@wmbriggs.com.

http://wmbriggs.com/book

matt@wmbriggs.com

To the best teachers I ever had: Mrs. Keller (5th grade English), Dale Hardy (high school chemistry), Ralph Schweigert (high school music), and my Dad (everything else).

Contents

Preface

There is no such thing as a *law of averages*. If you are watching a roulette wheel, and it has just come up red twelve times in a row, the in no way is black "due" to show. That wheel has no memory; it cannot recall that it was just red for so long. There are no hidden forces that will nudge it back to black. Evidence, and logic, tell us that the probability black will be next is the same no matter how many times we saw red.

Real life events, like a ball landing on a certain color in roulette, will not always "even out." Just because it's possible to win the lottery does not mean, unfortunately, that if you keep playing you will eventually win. No, there is no law of averages. But there is such a thing as being too sure of yourself—as you will be if you try to make decisions under this mythical law.

You might then be surprised to learn that much of probability and statistics—as taught in college courses all over the world—are designed around a law-of-averages-like set of procedures. This means that if you use those traditional methods, then you will be too sure of your results, just as when you were too certain that black would show next.

This book is different than others in two major ways. The first is the focus on what is called *objective Bayesian* probability. This is a logical, non-subjective, evidence-driven way to understand probability. Chapter 1 details the merits of this approach, and the demerits of older ideas.

The second difference is more difficult to explain, and which will become clearer as you progress. Briefly, to create a probability requires fixing certain mathematical objects called *parameters*. Almost every statistical method in use focuses solely on these parameters. But here's the thing. These parameters do not exist, they cannot be measured, seen, touched, or tasted. Isn't is strange, then, that nearly every statistical method in use is designed to make statements about parameters. This book will show you how to remove the influence of parameters and bring the focus back to reality, to real, tangible, measurable, observables. Things you can touch and see. Doing so will give us a much clearer—and fairer—picture of what is going on in any problem. Incidentally, in mathematical circles, this approach goes by the fancy term *predictive inference* (e.g. Geisser, 1993; Lee et al., 1996).

Probability, and its stepchild statistics, exist to do one thing: help us to understand and quantify uncertainty. A lot of uncertainty can be quantified, and some cannot. We'll learn how to indentify both situations, to know when

we can use math and computers and when we will be left with nothing but our intuitions.

For Students

Most books contain pre-packaged datasets for you to analyze. These have some utility, but my experience is that you spend just as much time trying to understand the data as you do the methods of analyzing them. So this book only has one major and two minor datasets, which are used throughout the text to illustrate concepts.

The real difference is that all of you *must collect your own data for a major project*. The data should be in the form that it can be analyzed by linear or logistic regression. You will understand what these are later, but for now it is important to know that these are the two types of models that comprise the bulk of statistics as actually used by civilians. You should start thinking about what interests you as soon as the class begins, because collecting data takes time. An extended real-life appendicitis study is given as an example.

The process of defining an interesting idea, identifying and collecting data which will describe that idea, and then analyzing that data does a better job of teaching the subject than using any canned example can ever do. I have lead many students through projects and they invariably understand probability and statistics far better when they get to pick their own data.

Understanding logic and interpreting probability is, or should be, the largest portion of a statistics course. But usually, a lot of effort must be expended in mastering the actual mechanics of the methods: how to calculate a mean and so forth. A great deal of statistics evolved before easy access to computers became usual. Therefore, many of the methods, and even most or all that are traditionally taught in introductory courses were designed to be calculated by hand and by referrals to standardized tables. This often meant that certain crude assumptions were made which would greatly simplify the calculations involved. This was certainly a rational thing to do at the time.

Naturally, now that computers have become so cheap that even professors can afford them, the methods you learn in statistics classes will have changed to reflect this fact, right?

Uh...no, that's wrong, actually.

You will still find legions of students churning out sums of squares of X, sums of squares of Y, total sums of squares, computing means and variances by plugging in numbers into calculators, and looking up probabilities in tables with small print in the backs of ridiculously heavy books. I can think of two reasons for this: (1) All the textbooks are written in this old-fashioned way, and all the courses are already created around these textbooks. Do you have any idea how long it takes to develop a new college course? Not just writing new textbooks, but also creating thousands of new homework questions and exams, and answers for them, too, and on and on? Just don't

even ask. (2) A lot of us professors have progressed a long way toward fogeyhood, and we feel that, since we had to calculate sums of squares by hand, then, by golly, our students will too! It builds character.

This book therefore represents a hefty departure from tradition, not just in its more realistic focus, but away from hand calculation. You will not learn as many different methods in this book as you would from others, but you'll learn the few we have better and more completely. And you won't make the same mistakes in understanding that most others students make.

For Teachers

I have used this book succesfully as a one-semester course for groups of physicians and residents at Methodist, fresh undergraduates (mostly business students) at Central Michigan University, and for professional Masters students at the ILR school of Cornell. These groups share at least one thing in common: they need to use and understand probability on a daily basis but they do not know any math beyond multiplication and division.

Because of this, I emphasize understanding above all things, which necessarily means I de-emphasize math and the memorization of formulas. This strategy is sometimes met with disapproval by those who say "You should not dumb down a course just so people can pass it", a sentiment with which I heartily agree. If anything, though, this book is more difficult than many other introductory probability and statistics texts *because* there are no routines to have by rote. If the student does not assimilate each concept in succession, then he will have a nearly impossible time completing the course. That being said, I have rarely lost a student.

The book is designed to be read in order. All Chapters, except perhaps 6 and 15 if there is a lack of time, are meant to be read and gone over in class. All homework questions in each Chapter should be completed, except perhaps for those marked "EXTRA." The class is designed so that all students complete their own guided, but self-designed, data-analysis (linear or logistic regression) project. This means that the material in the book is usually covered in the first 80 to 90% of the time alloted to the course, with that remaining given to the class projects.

I usually have the students (in small groups if the class is large, or individually if there are 15 or fewer) present the project to the class in stages. The first presentation is for the students to describe their projects and to solicit ideas from the class. The second and third, which may be combined, are for describing the exploratory analysis and results. If the class is graduate level, an "abstract" or small paper is required. Satisfactory completion of the course is a demonstration the students know what they are doing and saying about their data.

Acknowledgments

Russ Zaretzki (U. Tennessee), Tilman Gneiting (U. Washington), and Marty Wells (Cornell) all provided ears while I described my ideas. This is not to say that they are to blame. Neil Mower (CMU), David Ruppert

(Cornell), and Dan Wilks (Cornell) put up with me for many years. Mary Charlson (Weill Cornell) generously gave me the opportunity to begin this work. The entire math department at Central Michigan University kindly let me experiment on their students. Bob Birkhahn and the folks at Methodist kept me in hops while writing.

Several classes were forced to read through earlier drafts of this book. Especial thanks to ILR 2008 for providing so much useful feedback, and for the last class George Thorogood-themed show and tell. Pat McGreal's meticulous Army training allowed him to spot more typos than was thought humanly possible.

The regular readers of my blog, where parts of this book previewed piece by piece, provided razor sharp editing and keen questioning and kept me from making major blunders. So thanks to (screen names) Mike D, JH, Harvey, Joy, Noahpoah, Harry G, Bernie, Lucia, Luis Dias, Noblesse Oblige, Charlie (Colorado), Dan Hughes, Mr C Physics, Jinnah Mohammed, Ari, Steve Hempell, Wade Michaels, Raphael, TCO, Sylvain, Schnoerkelmanon, and many others (sorry if I left you out!). Any mistakes left are obviously their fault.

Lastly, I would like to thank two men whom I never met (and who are now dead). David Stove and Edwin T. Jaynes. The books *The Rationality of Induction* by Stove and *Probability Theory: The Logic of Science* by Jaynes quite literally changed my entire way of thinking about uncertainty, probability, and philosophy. I owe these men a debt of gratitude that I can only hope to repay by passing on their insights to new generations of students. I cannot recommend these books strongly enough. Those who are familiar with these texts will instantly see where I have stolen liberally from them. Students who go on in probability and statistics (and have absorbed some calculus) should visit Jaynes after this book. Those in philosophy should read everything written by Stove, particularly the books *On Enlightenment*, *The Plato Cult*, and *Scientific Irrationalism*.

CHAPTER 1

Logic

1. Certainty & Uncertainty

There are some things we know with *certainty*. These things are true or false given some evidence or just because they are obviously true or false. There are many more things about which we are *uncertain*. These things are more or less probable given some evidence. And there are still more things of which nobody can ever quantify the uncertainty. These things are nonsensical or paradoxical.

First I want to prove to you there are things that are true, but which cannot be proved to be true, and which are true based on *no* evidence. Suppose some proposition A is true. A might be shorthand for "I am a citizen of Planet Earth"; writing just 'A' is easier than writing the entire statement; the proposition is everything between the quotation marks. Also suppose some proposition B is true. B might be "Some people are frightfully boring". Then this proposition: "A and B"—meaning "I am a citizen of Planet Earth *and* some people are frightfully boring"—is true, right? But also true is the proposition "B and A". We were allowed to reverse the letters A and B and the joint proposition stayed true. Why? Why doesn't switching make the new proposition false? Nobody knows. It is just assumed that switching the letters is valid and does not change the truth of the proposition. The operation of switching does not change the truth of statements like this, but nobody will ever be able to prove or explain why switching has this property. If you like, you can say we take switching's truth-preserving quality on faith.

That there are certain propositions which are assumed true based on no evidence will not be surprising to you if you have ever studied mathematics. The basis of all mathematics rests on beliefs which are assumed to be true but cannot be proved to be true. These beliefs are called *axioms*. Axioms are the foundation: theorems, lemmas, and proofs are the bricks which build upon the base using rules (like the switching propositions rule) that are also assumed valid. The axioms and basic rules cannot, and can never, be proved to be true. Another way to say this is, "We hold these truths to be self-evident." I say it this way to hint that there are non-mathematical truths that cannot be proved but are held to be true, too. It is important to understand that some truths cannot be proved true—but this also means

that they some propositions are false but cannot be proved false.[1] Thus, there are some, even many, propositions that we have to accept based on no evidence except our intuition.

Here, for example, is one of them; an axiom of arithmetic. For all natural numbers[2] x and y, if $x = y$, then $y = x$. Obviously true, right? It is just like our switching statements rule above, except applied to numbers and not propositions. There is no way to prove this axiom is valid. You have to take it on faith. But from this axiom and a couple of others, all of mathematics arises. There are other axioms[3]—two, actually—that define probability. Here, due to Cox (1961), is one of those axioms: The probability of a statement on given evidence determines the probability of its contradictory on the same evidence. I'll explain these terms as we go, but this essentially means that the probability that something is false is one minus the probability it is true.

It is the job of logic, probability, and statistics to quantify the amount of certainty any given proposition has. An example of a proposition which might interest us: "This new drug improves memory in Alzheimer patients by at least ten percent." How probable is it that that proposition is true given some specific evidence, perhaps in the form of a clinical trial? Another proposition: "This stock will increase in price by at least two dollars within the next thirty days." Another: "Marketing campaign B will result in more sales than campaign A over the next month." In order to specify how probable these statements are, we need evidence, some of which comes in the form of *data*. Manipulating data to provide coherent evidence is why we need statistics.

Manipulating data, while extremely important, is in some sense only mechanical. We must always keep in mind that our goal is to make sense of the world and to quantify the uncertainty we have in given problems. So we will hold off on playing with data for several chapters until we understand exactly what probability really means.

2. Logic

We start with simple logic. Here is a classical logical argument, slightly reworked:

---

[1] For example, take any proposition that is known to be true based on no evidence, call it T, and then the proposition "T is false" is false but cannot be proved to be so based on any evidence.

[2] Natural numbers are $1, 2, 3, \ldots$.

[3] See Halpern (1999a,b) for criticisms of the original Cox axioms and Dupre and Tipler (2007) and references therein for a defense.

All statistics books are boring.

Breaking the Law of Averages is a statistics book.

Therefore, *Breaking the Law of Averages* is boring.

The structure of this argument can be broken down as follows. The two propositions above the horizontal line are called *premises*; they are our *evidence* for the proposition below the line, which is the *conclusion*. We can use the words "premises" and "evidence" interchangeably. We want to know the probability that the conclusion is true given these two premises. Given the evidence listed, it is 1 (probability is a number between, and including, 0 and 1). The conclusion is true *given* these premises. Another way to say this is the conclusion is *entailed* by the premises (or evidence)[4]. It is crucial to understand that the conclusion might not be true given different premises (see the homework).

Thus, you are no doubt tempted to say that the probability of the conclusion is not 1, that is, that the conclusion is *not* certain, because, you say to yourself, statistics books like this are nothing if not fun. But that would be missing the point. You are not free to add to the evidence (premises) given. You *must* assess the probability of the conclusion given *only* the evidence provided.

This method of writing arguments is important because it lays bare the exact evidence we will use to assess the conclusion. It also shows you that there are things we can know to be true *given* certain evidence but that might not be true given different evidence. Another way to say this, which is commonly used in statistics, is that the conclusion is true *conditional* on given evidence.

Here is another argument, courtesy (in form, at least) of David Hume (2003):

All the reality TV shows I have observed before have been ridiculous.

This is a (new) reality show before me.

Therefore, this reality show will be ridiculous.

The conclusion here does not follow from the premises; that is, the conclusion is not certainly true, nor is it certainly false (its probability is not 1 nor 0). You may be surprised to learn this, but the universe is not set up to guarantee that all reality TV shows will be ridiculous. It may be that, for

[4]Yet another way to say this is that the argument is *valid*

whatever unknown reason, that *this* new show will not be ridiculous. The conclusion, then, is *contingent* on certain facts (about network executives, uncontrollable weeping of contestants, viewers' habits, etc.), and any conclusion that is contingent (on certain conditions about the universe holding) is never certainly true nor certainly false.[5]. So what is the probability that the conclusion is true? Pretty high, but not 1 and not 0. We don't need to, and there is nothing in the universe that guarantees that we can, put an exact number of this probability. It is fine to leave it vague! We can say, "Given this evidence, it is pretty likely that this show will be ridiculous." In fact, many arguments in life do not require numerical probability values be assigned to their conclusions (see Keynes (2004); Franklin (2001b)).

Another argument:

> I will roll a die, which has six sides, only one of which will show.
>
> Just 1 side of the six is labeled "6.".
>
> ---
>
> Therefore, the side that shows will be a "6."

The conclusion here is also not certain, as will be plainly obvious to any of us. The conclusion is contingent (not certainly true or false) given just the evidence in the two premises. Here we can assign an exact number to the probability that the conclusion is true: it is is 1 in 6, or about 0.17. You knew this before reading this book, but you might not have seen it written out like this before. That we can assign probabilities this way is one of the principles of logical probability.

Here is a very difficult argument to understand, but it is important, so we will take our time with it:

> T
>
> ---
>
> M

T is any tautology, which is a proposition that is necessarily true, or always true no matter what: an example of a tautology is T = "Either Joe is a pain in the ass, or he is not." The proposition (all the stuff inside the quotation marks) T is always true, is it not? Another tautology, T = "Tomorrow it will rain or it will not." In this book, whenever we see "T" is means a true statement; thus the probability of T being true is always 1. A shorthand way to say this is the probability of T is 1 (we can leave out the "being true"). There are other true statements that are not necessarily true. One appeared in the argument before this, "Just 1 side of the six is

[5]This fact comes to us from Aristotle.

labeled '6'" is a true statement just in case there is a die that has one side with a 6 on it. This statement is an *observation*, and a true one, but it is not necessarily true. The die *could* have no sides with a 6 on it. Later, we will call observation propositions like this *data*.

M is some proposition—it could be anything—which I'll leave undefined for a moment to make a point. It should be obvious that if we know nothing about M, we cannot state any probability about it because there is no direct evidence about M in T. So if you let T be, for example, the tautology about Joe, and I did not tell you anything about M, then the probability of M being true is undefined. Thus, it is possible that some propositions have no probability (of being true) at all. M is empty in this sense. But let M be any proposition you want (make one up in your head right now) and ask "What is the probability M is true?" It has no probability! You can *never* ask the (unconditional) probability of any M. You can *only* ask what is the probability of M *given* some evidence.

Let's change our tautology to T = "M is true or false," which is another way of saying, T = "M will happen or it won't." These tautologies have buried within them implicit information[6] about M, which is that M can happen or not. So it must be physically possible for M to be true. If I add evidence that M is physically possible, but not certain, then we are saying some positive thing about M, it is information about M that is useful. We are no longer in complete ignorance about M. With this new information about M (implicit) in the premises, we can then state a probability of M being true. However, the evidence is pretty weak: saying something *might* be true doesn't say much. So the best we can do is to say the probability of M is greater than 0 but less than 1.

This is the same as saying M is *contingent*. Given only that we know a proposition is contingent the best we can do is to say that the probability the proposition is true is greater than 0 but less than 1. Pause here, because this is a great truth. It is a reminder that all contingent statements, which we do not know the truth or falsity of, have a probability between 0 and 1. As long as the premises we have do not entail or negate M, then we will *never* know whether M is true or false; the best we can ever do is to say M is more or less probable (Briggs, 2007).

Much later we will meet arguments that look like this (though, sadly, they will rarely be written this way):

[6]Tautologies can easily contain implicit information because of the flexibility and nuances of human language. For example, T = "Either utopians are deluded or they are not." Or we can show that the simple T_1 = "Poor people have less money" is equivalent to the ridiculous, but common, T_2 = "Rising heating costs hit poor hardest." Thus, we generally should ignore this kind of headline because it is always true no matter what.

I have collected a bunch of data, which I will call x and y and about which I know certain things.

I want to describe my uncertainty in future values of x and y using probability models M_x and M_y.

Future $f(x) > f(y)$

It will be our job to quantify the uncertainty of the statement $f(x) > f(y)$, which is just some mathematical proposition about future data; for example, "there will be more xs than ys" ($f()$ is any function of the data of interest to us; and if you don't know what a function is, stick around). Our job as statisticians is to collect the evidence (the data) and create the models that make up the two premises of the argument. The job of defining the statement which is the conclusion (some interesting question about the data) is usually given to us by the powers that be. We use probability to tie the whole process together. All statistics problems fit this framework.

3. Theories of probability?

DEFINITION 1. **Logical probability:** *A measure of the logical relation between a list of premises (or observation statements, or evidence) and some conclusion (which is a proposition or event).*
Keynes (see 2004); Carnap (see 1950); Adams (see 1998); Benenson (see 1984).

This is not the only interpretation of probability.[7] There are many more, but only two of them have large followings stemming from different axioms (a fact which proves that not all people agree on the truth of all axioms, which itself should not be surprising to you). The largest (in statistics) is called *frequentism*, which was firmly set on its mathematical path by Andrei Nikolaevich Kolmogorov, but who was preceded by a slew of mathematician-gamblers like the famous Jakob Bernoulli. The newer and second largest group follows *Bayesianism*, named after Thomas Bayes, the man who first proved a theorem we will meet in Chapter 2. Bayes overlapped Pierre-Simon marquis de Laplace, who arguably was the better and more original probabilist (de Laplace, 1996); but somehow the name "Laplacianism" was never proposed. This situation is an instance of statistician Stephen *Stigler's Law of Eponymy* "No scientific discovery is named after its original discoverer." Incidentally, the so-called Gaussian, or normal probability distribution which we shall later meet, was originally discovered by Abraham de Moivre. Go figure.

[7] See Plato (1998); Franklin (2001b); Fine (1973) for technical histories and theories of probability.

The Bayesian camp is divided into two sects, the *subjective*,[8] Bayesian and the logical or *objective Bayesian*, which we follow in this book. The practical differences between the two flavors of Bayesianism are slight, but the philosophical distinctions are large and have to do with whether probabilities are matters of human belief. Subjectivists believe that events can be given quantitative probabilities merely by recourse to introspection; objectivists can show that probabilities are statements of logic. But these are mere quibbles in some sense, because the math and the methods are mainly the same for both flavors of Bayesianism, at least in those cases where actual "hard data" is being analyzed. In Chapter 15 we'll talk about how reliance on subjective probability can make you more certain than you should be when venturing away from data, and below is an argument showing you why probabilities cannot be subjective.

There is an enormous, yawning philosophical gap between Bayesians and frequentists. Frequentist theory rose to prominence in the early 20th century in part because of the (understandable) distaste of its originators with the non-objective and relativistic nature of early subjective Bayesianism. The classical theorists who invented frequentism wanted to develop probability on a solid, objective ground, freed from human opinion, and this is much to their credit. However, people are starting to realize that this approach has failed (Howson and Urbach, 1993; Berger and Selke, 1987; Little, 2006; Hájek, 1997). Frequentists also rejected Bayesianism because they did not accept arguments like this: All the flames I have seen before have been hot and this is a flame in front of me; therefore this flame will be hot. This is an *inductive* argument, a kind of argument which frightened a great host of twentieth century philosophers and statisticians because they were thought to be "groundless." The history of his odd behavior in philosophy is detailed in Stove (1982, 1986); Williams (1947) and in statistics in Briggs (2006); Campbell and Franklin (2004).

4. Why probability isn't relative frequency

For frequentists, probability is defined to be the frequency with which an event happens in the limit of "experiments" where that event can happen; that is, given that you run a number of "experiments" that approach infinity, then the ratio of those experiments in which the event happens to the total number of experiments is *defined* to be the probability that the event will happen. This obviously cannot tell you what the probability is for your well-defined, possibly unique, event happening now, but can only give you probabilities in the limit, after an infinite amount of time has elapsed for all those experiments to take place. Frequentists obviously never speak about propositions of unique events, because in that theory there *can be no unique*

[8]See Howson and Urbach (1993); Jeffrey (2004); Kyburg and Smokler (1964) for defenses of subjective probability.

events. Because of the reliance on limiting sequences, frequentists can never know, with certainty, the value of any probability.

There is a confusion here that can be readily fixed. Some very simple math shows that if the probability of A is some number p, and it's physically possible to give A many chances to occur, the relative frequency with which A does occur will approach the number p as the number of chances grows to infinity. This fact—that the relative frequency sometimes approaches p—is what lead people to the backward conclusion that probability *is* relative frequency.

Logical probabilists say that sometimes we can deduce probability (Franklin, 2001a), and both logical probabilists and frequentists agree that we can use the relative frequency (of data) to help guess something about that probability if it cannot be deduced[9]. We have already seen that in some problems we can deduce what the probability is (the dice throwing argument above is a good example). In cases like this, we do not need to use any data, so to speak, to help us learn what the probability is. Other times, of course, we cannot deduce the probability and so use data (and other evidence) to help us. But this does not make the (limiting sequence of that) data *the* probability.

To say that probability is relative frequency means something like this. We have, say, observed some number of die rolls which we will use to inform us about the probability of future rolls. According to the relative frequency philosophy, those die rolls we have seen are embedded in an infinite sequence of die rolls. Now, we have only seen a finite number of them so far, so this means that most of the rolls are set to occur in the future. When and under what conditions will they take place? How will those as-yet-to-happen rolls influence the actual probability? Remember: these events have not yet happened, but the totality of them *defines* the probability. This is a very odd belief to say the least.

If you still love relative frequency, it's still worse than it seems, even for the seemingly simple example of the die toss. What *exactly* defines the toss, what explicit reference do we use so that, if we believe in relative frequency, we can define the limiting sequence?[10] Tossing *just this* die? *Any* die? And how shall it be tossed? What will be the temperature, dew point, wind speed, gravitational field, how much spin, how high, how far, for what surface hardness, what position of the sun and orientation of the Earth's magnetic field, and on and on to an infinite list of exact circumstances, none of them having any particular claim to being the right reference set over any other.

You might be getting the idea that *every* event is unique, not just in die tossing, but for everything that happens— every physical thing that

[9]The guess is usually about a parameter and *not* the probability; we'll learn more about this later.

[10]The book by Cook (2002) examines this particular problem in detail.

happens does so under very specific, unique circumstances. Thus, nothing can have a limiting relative frequency; there are no reference classes. Logical probability, on the other hand, is not a matter of physics but of information. We can make logical probability statements because we supply the exact conditioning evidence (the premises); once those are in place, the probability follows. We do not have to include every possible condition (though we can, of course, be as explicit as we wish). The goal of logical probability is to provide conditional information.

In his *A Philosophical Essay on Probabilities* Laplace (1996), also quoted in (Tipler, 2008), opened his remarks with:

> All events, even those which on account of their insignificance do not seem to follow the great laws of nature, are a result of it just as necessarily as the revolutions of the sun. In ignorance of the ties which unite such events to the entire system of the universe, they have been made to depend upon final causes or upon [chance][11], according as they occur and are repeated with regularity, or appear without regard to order; but these imaginary causes have gradually receded with the widening bounds of knowledge and disappear entirely before sound philosophy, which see in the only the expression of our ignorance of the true causes.

That is, probability is a measure of ignorance, or information, and is not a physical entity.

The confusion between probability and relative frequency was helped because people first got interested in frequentist probability by asking questions about gambling and biology. The man who initiated much of modern statistics, Ronald Aylmer Fisher Fisher (1970, 1973),[12] was also a biologist who asked questions like "Which breed of peas produces larger crops?" Both gambling and biological trials are situations where the relative frequencies of the events, like dice rolls or ratios of crop yields, can very quickly approach the actual probabilities. For example, drawing a heart out of a standard poker deck has logical probability 1 in 4, and simple experiments show that the relative frequency of experiments quickly approaches this. Try it at home and see.

Since people were focused on gambling and biology, they did not realize that some arguments that have a logical probability do not equal their relative frequency (of being true). To see this, let's examine one argument in

[11]Laplace actually used the word *hazard*.

[12]While an incredibly bright man, Fisher showed that all of us are imperfect when he repeatedly touted a ridiculously dull idea. Eugenics. He figured that you could breed the idiocy out of people by selectively culling the less desirable. Since Fisher also has strong claim on the title Father of Modern Genetics, many other intellectuals—all with advanced degrees and high education—at the time agreed with him about eugenics.

closer detail. This one is from Stove (1986, 1973) (we'll explore this argument again in Chapter 15):

Bob is a winged horse

Bob is a horse

The conclusion given the premise has logical probability 1, but has no relative frequency because there are no experiments in which we can collect winged horses named Bob (and then count how many are named Bob). This example, which might appear contrived, is anything but. There are many, many other arguments like this; they are called *counterfactual arguments*, meaning they start with a premise that we know to be false. Counterfactual arguments are everywhere. At the time I am writing, a current political example is "If Barack Obama did not get the Democrat nomination for president, then Hillary Clinton would have." A sad one, "If the Detroit Lions would have made the playoffs last year, then they would have lost their first playoff game." Many others start with "If only I had..." We often make decisions based on these arguments, and so we often have need of probability for them. This topic is discussed in more detail in Chapter 15.

There are also many arguments in which the premise is not false and there does or can not exist any relative frequency of its conclusion being true; however, a discussion of these brings us further than we want to go in this book.[13]

Hájek (1997) collects many other examples why frequentism fails, most of which are more technical than what we can look at in this book. As he says in that paper, "To philosophers or philosophically inclined scientists, the demise of frequentism is familiar". But word of its demise has not yet spread to the statistical community, which tenaciously holds on to the old beliefs. Even statisticians who follow the modern way carry around frequentist baggage, simply because, to become a statistician you are *required* to first learn the relative frequency way before you can move on.

These detailed explanations of frequentist peculiarities are to prepare you for some of the odd methods and the even odder interpretations of these methods that have arisen out of frequentist probability theory over the past $\sim$ 100 years. We will meet these methods later in this book, and you will certainly meet them when reading results produced by other people. You will be well equipped, once you finish reading this book, to understand common claims made with classical statistics, and you will be able to understand its limitations.

[13]For more information see Chapter 10 of Stove (1986).

5. Why probability isn't subjective

If 3 out of 4 dentists agree that using Dr Johnston's Whitening Powder makes for shiny teeth, what is the probability that *your* dentist thinks so? (You are asking this question before you learn what your doctor prefers.) Given *only* the evidence that 3 out of 4 etc., then we know the probability is 0.75 that your dentist likes Dr Johnston's Whitening Powder.

But what if you learned your dentist had just attended an "informational seminar" (with free lunch) sponsored by Galaxy Pharmaceuticals, the manufacturer of Dr Johnston's Whitening Powder? This introduces new evidence, and will therefore modify the probability that your doctor would recommend Dr Johnston's.

It may suddenly seem that probability *is* a matter of belief, of subjective feeling, because different people will have different opinions on how the free lunch will effect the doctor's endorsement. Probability cannot be a matter of free choice, however. For example, knowing only that a die has 6 sides, and knowing *nothing else* except that the outcome of the die toss is contingent, then the probability of seeing a 6 is 1 in 6, or about 0.17, regardless of what you or I or anybody thinks. You are not free to choose another probability when the evidence and the conclusion are specific like this.

After you learn of your doc's cozying up to the pharmaceutical representative, *you* would be inclined to increase your probability that he would tout Dr Johnston's to, say, the extent of 0.95. *I* may come to a different conclusion, say, 0.76 (just slightly higher). Why? Because we are now using *different sets or collections of information*, different evidence or premises, which naturally change our probability assessments. You might know more about pharmaceutical companies than I do, for example, and this causes you to be more cynical, whereas I know more about the purity and selflessness of doctors, and this causes me to be trusting.

But, if I agreed with you exactly about the new evidence, and I felt it was as relevant as you did, then we must share the same probability that the conclusion was true. This, of course, is very unlikely to happen (see the homework). Rarely will two people agree on a list of premises when the argument involves human affairs, and so it is natural that for most complex things, people will come to different probabilities that the conclusions are true. Does this remind you of politics?

Because people never agree on the set of premises—and they cannot loosely agree on them, they have to agree on them *exactly*—probabilities will differ. In this sense, probabilities are subjective—rather, it is the choice of premises that is subjective, probability itself is not. The probabilities assigned to a conclusion *given* a set of premises is not. The probability of a conclusion always follows logically from the given premises.

Let's further highlight this with another example: E_1 = "An item which will be tossed once has n sides and just one side is labeled S." The proposition of interest is A = "We see an S." We already know that, via logical

probability, $\Pr(\text{A}|\text{Evidence}) = 1/n$. A subjectivist might agree with this but has some problems: justifying why he does so, why anybody should agree with him, and explaining why he does not pick another number. The logical probabilistic *must* choose $1/n$, regardless of what he wants the number to be. I invite anybody who is a subjectivist to argue either for $1/n$ or against it.

Even more illuminating is this classic: E_2 = "All men are mortal and Socrates is a man." The proposition of interest is B = "Socrates is mortal" and $\Pr(\text{B}|\text{E}_2) = 1$.

A subjectivist is allowed to argue that the probability was some other number than 1, an impossibility in logic(al probability). To prove what I just said is false requires you to show how it is impossible for a subjectivist to supply a different answer. Note that just saying "It's a valid argument" doesn't work, because then you have to answer say why valid arguments do not fall under subjective probability rules.

I was once a referee on a paper (for *Weather and Forecasting*) where the author was trying to introduce the use of Bayesian subjective probability in a problem with a beta prior. All you have to know here is that the beta is a probability used in certain kinds of arguments and that it has two parameters which must be specified: different choices will lead to different answers. The usual values for the parameters are 1 and 1 (or 1/2 and 1/2) chosen by (semi) logical arguments, but the author, on a whim, chose 10 and 3. I tried to argue with him and the editor that this was silly, but the author countered that since Bayes was subjective, he was free to choose whatever prior he wanted. There's no defeating that argument, not ever, given the premise that probabilities *are* subjective. (The paper, incidentally, was published with the author's odd values.)

The choice of probability models for observables and the probability models for the parameters of the those models make a huge difference in the final answer, which no one disputes. Old-school frequentists rightly fear that people, arguing subjectivity allows them anything, could select priors so as to produce desired or pre-determined results. This can happen, so the fears of the old guard are real (though perhaps exaggerated). Anyway, in this book, we use logical probability all the way.

6. Randomness

There is a great deal of nonsense written and said about randomness. Although it's never stated directly, there is a certain mysticism about randomness which is supposed to "bless" statistical results. Samples must be "random" to be statistically valid, it is often said. This view is false—we will take care of these beliefs when we meet them.

When we say some outcome (proposition) is random, all we are saying is that we don't know what that outcome will be certainly. If that outcome is, for example, the result of a dice roll, then we are saying that we do not

know in advance what face will show. That is, the outcome is "random." Not knowing what something is, is saying that the truth of that thing is unknown, or "random." Technically, then, randomness means ignorance. Only this, and nothing more.

Classical statistics talks a great deal about "random variables." We'll meet these creatures later, but for now, we can remember that every time we see that term, it only means "unknown value." Likewise, a favorite term is *randomized trial*, which are the only kind of experiment accepted in some quarters, all other forms of trial deemed untermenchen (tell this to physicists and chemists etc. who regularly run *un*randomized trials, yet have managed to learn a great deal about the world). "Randomized trials" only means a trial where some of the outcomes will be unknown, and others will be known or controlled. Chapter 14 will talk more about this.

There is no inherent physical quantity that is randomness that we can suck out of the dice, for example, or out of any other thing. So randomness isn't needed to assign logical probabilities to physical events (Briggs, 2006, 2007). Barring quantum mechanics,[14] about which Richard Feynman said fairly that nobody understands, there is nothing spooky or mysterious behind events that are random. Certainly there is no agreement yet even about what quantum mechanical measurements are; for example, Tipler (2008) and other authors claim, in the Many Worlds quantum interpretation, that the physical world is fully deterministic (see also Cook, 2002).

Nevertheless, some pine for physical randomness, if only to bulk up their faith in frequentism. For example, Jaynes (2003) claims that the originators of frequentism wanted there to be a inherently non-deterministic basis to the universe because of the mutation theory of natural selection. Mutations were said to be "random." But something caused the mutations, and just we do not know the causes does not mean that they do not exist nor that they are due to some ethereal property of randomness.

7. A taste of Boolean algebra

George Boole, in his 1854 book *An Investigation of the Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probabilities*, introduced a calculus for working with statements which is now called *Boolean Algebra*, (Brown, 2003). We have been using this calculus so far, although you didn't know it because the manipulations have been mostly intuitive. But there are some formal rules that we'll need before we get much further. Our only real problem will be notation. You can look in a dozen books and see thirteen different notations, all of which mean the same thing, but which you must assimilate anew each time before you can understand what the authors are saying. There are some consistencies, but very few. I

[14] How I regret the term *quantum*. If the physics describing movement of small objects were instead called *discrete mechanics*, we'd have far less silliness in the world.

try to stick with the most-used symbols, but be warned: if you read another book, you will likely see a different notation.

We use capital Latin letters to stand for propositions, which may be true or false. So A = "Real estate agents never lie" is a statement which is either true or false, but we might not know whether it is true or false. If we have another proposition, B, and we write AB, this means the joint proposition "A *and* B". So if B = "The *New York Times* always reports stories in an absolutely unbiased manner", then AB means "Agents never lie *and* the *Times* is always unbiased." Both parts of the proposition must be true for AB to be (jointly) true.

When we write A $\cup$ B, it means "A *or* B." If either A *or* B is true, or both are true, then the proposition "A $\cup$ B" is true.

There are, unfortunately, many ways to write that a proposition is false. None of them is especially beautiful. One way, the one we'll use here, is A^F. Thus, if A = "Real estate agents never lie", then A^F = "Real estate agents sometimes lie." If A = "A Head will show when I flip this coin", then A^F = "A Head will not show when I flip this coin", which is a fancy way of saying we'll see a tail. Other ways of writing A is false that you'll see in other books: A^c (the *c* means complement, or opposite), $\neg$A, $\sim$A, and *not*A.

Sometimes we need a truth, a proposition which is always true. We'll label these statements T. There are propositions that are true just because the universe was in a certain position, meaning we observed the proposition to be true. Like above, it might be true that T = "you saw a dog on your lawn last week" just in case you actually did see a dog on your lawn last week. All observed data statements are truths in this sense. For example, T = "In patient 37, I measured a systolic blood pressure of 182 mmHg" just in case I actually did measure 182 mmHg on patient 37. There are other propositions which are necessarily true, which are true regardless of anything. Tautologies, which we met earlier, are the most common examples of these truths.

8. Homework

(1) Rewrite the first logical argument (with the conclusion "*Breaking the Law of Averages* is boring") using one or more different premises such that the conclusion has probability 0. Rewrite it again so that it has a probability between 0 and 1.
(2) Rewrite the second argument (with the conclusion "This reality show will be ridiculous") using one or more different premises such that the conclusion has probability 1.
(3) What is the probability of drawing the Jack of Hearts from a standard deck of playing cards? Write your argument in the same form as the dice example.
(4) In the dice argument, the *only* evidence was "I will roll a die, which has six sides, only one of which will show" and "Just 1 side of the six is labeled '6.'" The conclusion, "We see a '6'" had logical probability 1/6. Why don't we need the premise "The die is fair"? Similarly, think of a

coin flip, which has similar premises: "Flip a coin and only one side has an H" to the conclusion "See an H", which has logical probability 1/2. Why is it not necessary to add the premise "This is a fair coin"?

(5) Alice hands you a deck of playing cards which she says are well shuffled. Bob hands you another deck and says nothing. What is the probability of drawing the Jack of Hearts from Alice's deck and what is it from Bob's deck? Explain your answer.

(6) Charlie hands you a third deck, but as he does so, he gives you a wink. What is the probability of drawing the Jack of Hearts from Charlie's deck? Write your answer in the form of a logical argument. Be clear about your premises.

(7) The logic of advertisements: (a) An ad states that you can "Save up to 50%!". Logically, what is the (entire) range of savings?; (b) Stanford Financial took out a full page ad in the *Wall Street Journal* with a picture of golfer Vijay Singh listing his enormous number of tournament wins with the words "Vijay Means Victory." Given this evidence, what is the probability Stanford Financial won't lose money on your investment?

(8) New York City "Health Czar" Thomas Frieden (D), who successfully banned smoking and trans fat in restaurants and who now wants to add salt to the list, said in an issue of *Circulation: Cardiovascular Quality and Outcomes*,[15] that "cardiovascular disease is the leading cause of death in the United States." Describe why no government or no person, no matter the purity of their hearts, can *ever* eliminate the leading cause of death.

(9) My insurance company recently disputed a claim I had made. In order for them to pay, they said that I had to provide proof that I did not have other health insurance. What is the probability I could provide such proof?

(10) There is a famous, if not tedious, statement that goes L = "This statement is false." What is the probability that L is true? Explain how you arrived at your answer.

(11) Right before I come to class, I put either a quarter or a dime in my pocket. Once I get there, I pull out the coin and conceal it from your view. What is the probability that I reveal the quarter? Write your answer in the form of a logical argument. Be clear about your premises.

(12) Bounding probabilities. Is it possible to translate the statement "Given evidence E (about the past performance, knowledge of the starting lineup, etc.), I conclude the Detroit Tigers will most likely win tomorrow's game" into a numerical value? Explain how you arrived at your answer.

(13) Wiley, a prominent textbook publisher, is keen that their books contain no *bias*. In their author guidelines, they give this example of how to avoid bias. "**Biased**: Almost everyone likes his bacon crisp. **Unbiased**: Most people like their bacon crisp." Let both of these serve as premises in two different arguments. What can you say about the probability of A = "Joe likes his bacon crisp" given either of these two premises?

(14) Create your own tautology T (different from those in the text).

(15) B = "The sun rises in the west." What it the probability of BT and B$\cup$T, where T is the tautology from the previous problem.

[15]DOI: 10.1161/CIRCOUTCOMES.108.791954

(16) What is the probability that A = "Wearing white shoes after Labor Day is wrong"? Explain.
(17) Write out a list of premises (be explicit) for the Dr Johnston's Whitening Powder example supposing you learned your doctor did have that free lunch, then give a guess for the probability of the conclusion. Compare your list with other people in the class. Do any two lists exactly match?
(18) A statement of moral relativism, often called upon in Postmodern philosophy and by highly-educated people, is C = "There is no truth." What is the probability C is true? This reminds me of an appearance of Leonard Nimoy on *The Simpsons* where he said, "The following tale of alien encounters is true. And by true, I mean false. It's all lies. But they are entertaining lies. And in the end, isn't that the real truth? The answer: No."
(19) If A = "Real estate agents never lie", then A^F = "Real estate agents sometimes lie." Why isn't A^F = "Real estate agents *always* lie"?
(20) Why is the probability that D is true given the evidence you listed *not* evidence that probability is subjective?
(21) EXTRA In the argument T, therefore M, where T is the tautology "M will happen or it won't", why isn't the probability of M 1/2?
(22) EXTRA A current theme in statistics is that we should design our procedures in the modern way but such that they have good relative frequency properties. That is, we should pick a procedure for the problem in front of us that is not necessarily optimal for that problem, but that when this procedure is applied to similar problems the relative frequency of solutions across the problems will be optimal (see Little, 2006). Show why this argument is wrong.

CHAPTER 2

Probability

1. Probability rule number 1

We always write propositions, which are observable or definable events, with Latin letters. For example, let E = "3 out of 4 dentists recommend Dr Johnston's" and A = "My dentist will recommend Dr Johnston's." The notation we use to write probability is

$$\Pr(\mathrm{A}|\mathrm{E}) = \frac{3}{4}. \tag{1}$$

The shorthand in equation (1) means, in English, "The probability that A is true *given* the evidence E (is true), is 0.75." Don't let the "E" or "A" confuse you: these are just place holders so we can avoid typing all that stuff about the dentist each time; we could have used any other letters, and equation (1) would be just the same. This is just like the arguments that we wrote in a long-hand fashion before; here, it is written briefly, the line separating the premises went from horizontal to vertical, but nothing really changed except that it is compact and easier to work with.

Try this one: a die has 6 sides and we want to know the chance that we see a 6 on the next throw. Our evidence E = "This die has 6 sides, and we can see only side at a time". We want to know B = "See a 6". This is

$$\Pr(\mathrm{B}|\mathrm{E}) = \frac{1}{6}.$$

Note that I re-used the letter E for the new evidence; the specific letter just does not matter. OK. Given the evidence E, what is the probability we C="see a 5 or 6"? Given our experience with dice, we know that we can only see one side at a time on any throw (which is evidence that is implicitly part of E), that we'll see one of $\{1, 2, 3, 4, 5, 6\}$, and that the evidence between seeing sides 5 or 6 is irrelevant, we can form this rule:

$$\Pr(\mathrm{C}|\mathrm{E}) = \Pr(5 \text{ or } 6|\mathrm{E}) = \Pr(5|\mathrm{E}) + \Pr(6|\mathrm{E}) = \frac{2}{6}.$$

This is the rule: In propositions like C with evidence E, the "or"s in English turn into the "+"s in the math.

In general, given an event, like a dice throw, that can be broken down into discrete sub-events, and there is evidence that these discrete sub-events can only happen one at a time, the probabilities of the individual sub-events sum together. For example, if the proposition or event A can be broken down

into discrete propositions or events $\mathrm{A} = \{\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n\}$, then

$$1 = \Pr(\mathrm{A}|\mathrm{E}) = \Pr(\mathrm{A}_1|\mathrm{E}) + \cdots + \Pr(\mathrm{A}_n|\mathrm{E}), \tag{2}$$

Here, A means "The number 1 or 2 or 3 or 4 or 5 or 6 will show". Notice that the sum of all sub-events always equals 1, for *something must happen*: this is because the equation means "A_1 or A_2 or ... A_n will happen." For the die, A = "a die is rolled" and A_1 ="a 1 shows", A_2 ="a 2 shows" etc. Incidentally, it is *not* always the case that each $\Pr(\mathrm{A}_i|\mathrm{E}) = 1/n$ for every A we can think of. That is, not all events will break apart into equally likely pieces.

Another way to view the problem is to think about the probability of *not* seeing a 5 or 6. The probability of seeing one of $\{1, 2, 3, 4, 5, 6\}$ is, of course, 1. So the probability of seeing a 5 or 6 must be 1 minus the probability of not seeing 5 or 6. This sort of "negative" thinking can come in very useful in solving problems.

How about the probability of seeing a number greater than 3? Well, what are the possibilities? Namely, 4, or 5, or 6. Turn the "or"s to "+"s. How about greater than or less than 3? The possibilities are 4, 5, or 6 as before, and then 1 or 2. Another way to express this is the probability of *not* seeing a 3.

The trick for these sorts of problems comes in turning the English into math. Never try to jump to the answer. You cannot go wrong by just writing out everything that can happen explicitly. The answer will then become obvious. Use this technique all throughout the book.

2. Probability rule number 2

What is the probability, in the throw of two dice (one after the other, or two together), of seeing a pair of 6s? We can guess that any given throw does not influence the results from any other throw; or, at least, that knowledge of the results of any given throw are *irrelevant* to knowledge of other throws. A classical way to say this is that the throws are *independent*.

In general, if $\mathrm{B}_1, \mathrm{B}_2, \ldots, \mathrm{B}_n$ are separate propositions or events, and the knowledge (or evidence) E of what happens on any B_i is irrelevant to what happens on any other B_j (for $j \neq i$; what happens what $j = i$?), then the probability that B_1 *and* B_2 *and* etc. B_n is true is

$$\Pr(\mathrm{B}_1\mathrm{B}_2 \ldots \mathrm{B}_n|\mathrm{E}) = \Pr(\mathrm{B}_1|\mathrm{E}) \Pr(\mathrm{B}_2|\mathrm{E}) \ldots \Pr(\mathrm{B}_n|\mathrm{E}). \tag{3}$$

In English, getting two 6s means getting a 6 on one *and* a 6 on the other. The rule is: the "and"s of English turn to the "×"s of math. Thus, $\frac{1}{6} \times \frac{1}{6}$. Getting three 6s in a row is $\Pr(6|\mathrm{E})^3 = 1/6^3 \approx 0.005$, and so on.

How about if you were watching the roulette wheel and saw that red hit 20 times in a row? What is the probability of that happening, you ask yourself. Well, the probability of red on any one spin is about 1/2, or close enough to 1/2 to do a rough calculation (this is our evidence). The probability of seeing red twice in a row must be, by our rule, "red on the first *and*

red on the second", or about $(1/2)^2 = 1/4$. Thus, twenty times in a row is $(1/2)^{20}$, which is about 1 in a million. That is so small a probability that it almost can't happen, so you decide to bet on black for the next spin, since you reason that black is certainly due. And that's just what the casino is hoping you'll do.

Again, the real difficulty in these problems—in *all* problems—is translating the English back into the mathematical rules you know. It is a non-trivial skill and takes lots and lots of practice, so do not be dismayed if you have a trying time at first.

3. Probability rule number 3

What is the probability that A= "somebody is older than 42"? What evidence (E) is given here? Well, none is explicitly stated, but there is some implicit evidence. We know at least the fact that we are asking about a human. North American human? We don't know, so we can assume E="all humans." That's fine, that is enough information to make a guess. Now here is a different question: what is the probability that somebody is A="older than 42" *given* that they B ="are older than 40" and that they are E (we *always* need some evidence to make a probability statement)? Well, someone can be 43 or older, so that it is possible that A and B can both be true; or someone can be 39 or younger, so that it is possible that A and B can both be false. But somebody can be 41, which means that A can be false and B true. Thus, it is never the case that B can be false and A true.

It is not always true that probability rule 2 holds, that is, the formula $\Pr(\mathrm{AB}|\mathrm{E}) = \Pr(\mathrm{A}|\mathrm{E})\Pr(\mathrm{B}|\mathrm{E})$ does not always work. That formula only works when what we know about B is irrelevant to what we know about A, and of A about B. If knowledge of B *is* relevant to our knowledge of A, and vice versa, then

$$\begin{aligned} \Pr(\mathrm{AB}|\mathrm{E}) &= \Pr(\mathrm{A}|\mathrm{BE})\Pr(\mathrm{B}|\mathrm{E}) \\ &= \Pr(\mathrm{B}|\mathrm{AE})\Pr(\mathrm{A}|\mathrm{E}), \end{aligned} \tag{4}$$

because logically, AB = BA. In words, the probability that A and B is true is the probability that A is true first *given* that B is true, times the probability that B is true—everything is conditional on some evidence E, of course. We first think about how true A is if B were true, then we ask how true is B (given E). As noted, the relationship is symmetric: we can first take the probability that B is true given A.

This rule lets us make a guess about the age problem, conditional on E, which means we're considering all humanity. The probability that somebody is older than 42 *given* they are older than 40 is pretty high, maybe 0.8 as a guess. The probability that somebody is older than 40 is maybe 0.4. So the probability of AB is roughly $0.8 \times 0.4 \approx 0.3$. There is no need to be more precise than one digit for our answer, we are just guessing anyway. Do not fall into the common trap of writing more digits to give the appearance of precision where it does not exist. I'll harp on this point later.

TABLE 1. Table of ASVAB scores for a room full of 155 recruits.

	Score < 38	Score ≥ 38
Air Force	10	25
Army	25	5
Marines	40	0
Navy	25	25

Meanwhile, let's demonstrate the rule. Pretend that you are a military recruiter and you have been ordered to find an electronics weapon technician for training. As always, you are in a terrible hurry. The person you select can be from any branch of the service, and, luckily, there is a room full of recruits next door. The job requires an intelligent person, and the military measures smartness with a test called the Armed Services Vocational Aptitude Battery, or ASVAB. Higher scores are better. The room contains the following men (which is our evidence E):

What is the probability that a recruit in the room B = "Scores over 38" *and* A = "Is a marine"? That is, what is the probability of BA? Try using $\Pr(\mathrm{B}|\mathrm{AE})\Pr(\mathrm{A}|\mathrm{E})$ first. What is $\Pr(\mathrm{A}|\mathrm{E})$? There are 155 recruits in the room, and 40 are marines, so this must be $40/155 \approx 0.26$. How about $\Pr(\mathrm{B}|\mathrm{AE})$? The A on the right hand side means, of course, that the recruit is a marine, so all we have to count are marines and nobody else. Again, there are 40 of them and none scored over 38. So $\Pr(\mathrm{B}|\mathrm{AE}) = 0$, which means $\Pr(\mathrm{BA}|\mathrm{E}) = 0 \times 0.26 = 0$. We could have solved this the opposite way, using $\Pr(\mathrm{A}|\mathrm{BE})\Pr(\mathrm{B}|\mathrm{E})$. What is $\Pr(\mathrm{B}|\mathrm{E})$? Well, 55 men scored over 38, so this is $55/155 \approx 0.35$. And $\Pr(\mathrm{A}|\mathrm{BE})$? The conditioning information is B, those with scores over 38, which are just those 55 men. None of them were A (none were marines), so $\Pr(\mathrm{A}|\mathrm{BE}) = 0$, and $\Pr(\mathrm{BA}|\mathrm{E}) = 0 \times 0.35 = 0$.

Suppose B can happen in one of n different ways. That is $\mathrm{B} = \mathrm{B}_1$ or B_2 or $\ldots \mathrm{B}_n$, then $\Pr(\mathrm{B}_1$ or B_2 or $\ldots$ or $\mathrm{B}_n|\mathrm{E}) =$
$\Pr(\mathrm{B}|\mathrm{E}) = \Pr(\mathrm{B}_1|\mathrm{E}) + \Pr(\mathrm{B}_2|\mathrm{E}) + \cdots + \Pr(\mathrm{B}_n|\mathrm{E}) = 1$. In the example above, B is a military recruit, and, for instance, B_3 was a marine; the other B_i are the other branches. Or think of B as a roll of a die, and the sides are B_i; *one* of the sides will come up, so B itself is always true, but only one of the B_i will be true. Since B is true, it is the case that A=AB, shorthand for A *and* B is true whenever A is true, and false whenever A is false. Hold it right there! Make sure you understand what is happening here, because it is tricky. Let B be any true statement. Then regardless whether A is true, the probability that A is true is equal to the probability that AB is true. This means that $\Pr(\mathrm{A}|\mathrm{E}) = \Pr(\mathrm{AB}|\mathrm{E})$. Incidentally, if C were also true, then $\Pr(\mathrm{A}|\mathrm{E}) = \Pr(\mathrm{ABC}|\mathrm{E})$, too.

We can prove this by recourse to rule number 3. $\Pr(\mathrm{AB}|\mathrm{E}) = \Pr(\mathrm{A}|\mathrm{BE})\Pr(\mathrm{B}|\mathrm{E}) = \Pr(\mathrm{A}|\mathrm{BE}) = \Pr(\mathrm{A}|\mathrm{E})$ because $\Pr(\mathrm{B}|\mathrm{E}) = 1$ and BE=E (remembering the Boolean algebra rules).

We can now use the simple consequence of logic that A=AB to help us calculate the probability of A using something called *total probability*:

$$\begin{aligned}
\Pr(\mathrm{A}|\mathrm{E}) &= \Pr(\mathrm{AB}|\mathrm{E}) \\
&= \Pr(\mathrm{A}(\mathrm{B}_1 \text{ or } \mathrm{B}_2 \text{ or } \ldots \text{ or } \mathrm{B}_n)|\mathrm{E}) \\
&= \Pr(\mathrm{AB}_1 \text{ or } \mathrm{AB}_2 \text{ or } \ldots \text{ or } \mathrm{AB}_n|\mathrm{E}) \\
&= \Pr(\mathrm{AB}_1|\mathrm{E}) + \cdots + \Pr(\mathrm{AB}_n|\mathrm{E}) \\
&= \Pr(\mathrm{A}|\mathrm{B}_1\mathrm{E})\Pr(\mathrm{B}_1|\mathrm{E}) + \cdots + \Pr(\mathrm{A}|\mathrm{B}_n\mathrm{E})\Pr(\mathrm{B}_n|\mathrm{E}). \qquad (5)
\end{aligned}$$

Sometimes we do not directly know $\Pr(\mathrm{A}|\mathrm{E})$ but we do know each $\Pr(\mathrm{A}|\mathrm{B}_i\mathrm{E})$, and when this is the case, we can use this formula.

Even though we know all the probabilities in the ASVAB example, let's see how total probability works with it. What is the probability of B = "The recruit is an Airman, Soldier, Marine, or Sailor"? It is 1, right? So what is the probability of A = "The recruit scores less than 38 on the ASVAB?" We can compute it right from the table: 100 recruits scored less than 38 out 155, so the probability is 100/155. Now use total probability. $\Pr(\mathrm{A}|\mathrm{E}) = \Pr(\mathrm{A}|\mathrm{B}_1\mathrm{E})\Pr(\mathrm{B}_1|\mathrm{E}) + \Pr(\mathrm{A}|\mathrm{B}_2\mathrm{E})\Pr(\mathrm{B}_2|\mathrm{E} + \Pr(\mathrm{A}|\mathrm{B}_3\mathrm{E})\Pr(\mathrm{B}_3|\mathrm{E} + \Pr(\mathrm{A}|\mathrm{B}_4\mathrm{E})\Pr(\mathrm{B}_4|\mathrm{E} = (10/35)(35/155) + (25/30)(30/155) + (40/40)(40/155) + (25/50)(50/155) = 100/155$ (notice all the cancellations: for example, the 25s go in $(10/25) \times (25/155)$ leaving 10/155).

4. Probability rule number 4: Bayes's rule

Since $\Pr(\mathrm{AB}|\mathrm{E}) = \Pr(\mathrm{A}|\mathrm{BE})\Pr(\mathrm{B}|\mathrm{E}) = \Pr(\mathrm{B}|\mathrm{AE})\Pr(\mathrm{A}|\mathrm{E})$, then it is true that

$$\Pr(\mathrm{B}|\mathrm{AE}) = \frac{\Pr(\mathrm{A}|\mathrm{BE})\Pr(\mathrm{B}|\mathrm{E})}{\Pr(\mathrm{A}|\mathrm{E})} \qquad (6)$$

We can go farther, using total probability, since $\Pr(\mathrm{A}|\mathrm{E}) = \Pr(\mathrm{A}|\mathrm{B}_1\mathrm{E})\Pr(\mathrm{B}_1|\mathrm{E}) + \cdots + \Pr(\mathrm{A}|\mathrm{B}_n\mathrm{E})\Pr(\mathrm{B}_n|\mathrm{E})$, then

$$\Pr(\mathrm{B}|\mathrm{AE}) = \frac{\Pr(\mathrm{A}|\mathrm{BE})\Pr(\mathrm{B}|\mathrm{E})}{\Pr(\mathrm{A}|\mathrm{B}_1\mathrm{E})\Pr(\mathrm{B}_1|\mathrm{E}) + \cdots + \Pr(\mathrm{A}|\mathrm{B}_n\mathrm{E})\Pr(\mathrm{B}_n|\mathrm{E})} \qquad (7)$$

which at this point is just some formula, though an important one, called *Bayes's rule*, named for the gentlemen who first wrote it out. We can make it less abstract by using an example. The example we'll use is a cliche—it's found in nearly every introductory probability book—but for good reason, because the example will always be relevant, especially when you get older (and you will, you will).

Suppose that you and a friend dine at Uncle Wong's Chinese restaurant. Unbeknownst to your companion, you have read Penn & Teller's *How to Play With Your Food* and have loaded your friend's fortune cookie with one of that book's tear-out predictions[1], and which reads *That lump is cancer*. Your friend opens this cookie. Now, since most people these days are paranoid

[1]These are on a sheet in the back of the book.

about their health, and your friend is certainly most people, he decides to take seriously the warning of his Chinese-American pastry, and runs to the doctor to have his lump checked out.

So he does, and has a blood test which looks for the presence of a certain chemical, which is known to be associated with lump cancer. Then the bad news comes: the A_+ = "test is positive"! The very natural thing your friend now wants to know is what is the probability that he B_+ = "has cancer." We write "has cancer" as B_+, because it is also possible that your friend does not have cancer, or B_-. The test might also have been negative, which we write as A_-. That is, $B = B_+ \cup B_-$, and $A = A_+ \cup A_-$; remember "$\cup$" is the mathematical way of writing "or." This means that $\Pr(B|E) = \Pr(B_+ \cup B_-|E) = \Pr(B_+|E) + \Pr(B_-|E) = 1$: a fancy way of stating that the probability of having cancer or not having cancer is 1. B is true for every single individual on this planet, right? (Don't read further until you agree with this.) This means $A_+ = A_+B = A_+B_+ \cup A_+B_-$ because the statement B is always true.

Equation (7) in this situation becomes—**Hold it here for a moment.** It is my experience that students start freaking out just about now. We used As and Bs etc. so far, but all of a sudden we have strange creatures like B_- and A_+. Never forget that these are *just symbols*, place-holders for actual statements, and we are free to substitute any symbols we want for the original ones—so the new equation is (stare at this awhile to make sure you get it):

$$\Pr(B_+|A_+E) = \frac{\Pr(A_+|B_+E)\Pr(B_+|E)}{\Pr(A_+|B_+E)\Pr(B_+|E) + \Pr(A_+|B_-E)\Pr(B_-|E)} \tag{8}$$

To solve this, you obviously need to know what the numbers on the right-hand-side are. Since this is such a common situation, all those probabilities have official names (which are given in the next section). For now, we'll just state things in words.

$\Pr(B_+|E)$ is the probability of having cancer given E. Your friend's E is probably something like "All Americans" or maybe "All Americans who are as old as I am" or whatever. *Your* E is "The whole thing is a sick, sick (but increasingly funny) joke." Let's say the probability of cancer, given your friend's E (but not given any information about the test), of lump cancer is 1 in a 1000.

The other probabilities are key. The first is the probability that somebody gets a positive test given that they have cancer. You might think this is 1, or near it. But let me tell you—and I know, I work in a hospital—it isn't, and sometimes it's not even close to 1. These tests are not, are far from, perfect. Mistakes creep in any old way. A not unusual value is 9 out of 10, which is high, but is not 10 out of 10.

This leaves the probability of somebody getting a positive test given that they do not have cancer. More flaws are possible this way, too, and a

common value is, say, 1 out of 100. We can then calculate

$$\Pr(\mathrm{B}_+|\mathrm{A}_+\mathrm{E}) = \frac{\frac{9}{10}\frac{1}{1000}}{\frac{9}{10}\frac{1}{1000} + \frac{1}{100}\frac{999}{1000}} \approx 0.083.$$

Yes, only an 8% chance of cancer given a positive test and the evidence E. Is that surprising?

Of course, using *your* E, the probability that your friend has lump cancer is near 0, so you should really pick up the check.

5. Extra: More Bayes's rule and beyond!

The example given in the Section above is used on a daily basis, particularly in medicine. Here is what the various pieces of that formula are called:

$\Pr(\mathrm{B}_+\|\mathrm{E})$	*Base rate*, or the probability of having the disease (B_+) in the population specified by E.
$\Pr(\mathrm{A}_+\|\mathrm{B}_+E)$	*Sensitivity*, or the probability of having a positive test (A_+) given the patient has the disease.
$\Pr(\mathrm{A}_-\|\mathrm{B}_-E)$	*Specificity*, or the probability of having a negative test (A_-) given the patient does *not* have the disease.
$\Pr(\mathrm{B}_+\|\mathrm{A}_+E)$	*Positive predictive value*, or the probability of having the disease given a positive test.
$\Pr(\mathrm{B}_-\|\mathrm{A}_-E)$	*Negative predictive value*, or the probability of not having the disease given a negative test.

The most touted statistic is sensitivity, although this, as you know by now, does not answer the question the patient has, which is "What is the probability that *I have the disease*?" A test, after all, can be perfectly sensitive, i.e. $\Pr(\mathrm{A}_+|\mathrm{B}_+E) = 1$, but this does not guarantee that $\Pr(\mathrm{B}_+|\mathrm{A}_+E) = 1$.

Before leaving this Chapter it is important to understand that we have only listed a paltry few of an ever-growing list of probability tools. These bare three allow us to solve many common problems, it is true, but they are not nearly enough to solve even most. Do not fool yourself into being too confident that the rule you apply to some new situation is just the right one, or that you have applied it correctly. It is a common mistake. However, with these three and the next one, we'll have all we need to understand the most common problems.

Just as an example of how easy it is to mislead yourself, let A = "It snows in December in New York City" and B = "Roll and die in December and see anything but a 1". Let E be whatever evidence we need to give the

probabilities of A and B, for example, historical weather reports and our standard knowledge of dice. Then $\Pr(\text{A}|\text{E}) \approx 0.9$ and $\Pr(\text{B}|\text{E}) = 5/6 \approx 0.8$. So what is $\Pr(\text{A or B}|\text{E})$? If you naively used Rule 1, then that "or" turns to pluses and $\Pr(\text{A or B}|\text{E}) \approx 0.9 + 0.8 = 1.7$ and we are in trouble. But we have misapplied Rule 1 because A and B are not part of the same event.

Now, it turns out that it is very easy to to use the rules we already know and show that

$$\Pr(\text{A or B}|\text{E}) = \Pr(\text{A}|\text{E}) + \Pr(\text{B}|\text{E}) - \Pr(\text{AB}|\text{E}) \tag{9}$$

and since $\Pr(\text{AB}|\text{E}) = \Pr(\text{A}|\text{BE})\Pr(\text{B}|\text{E})$ then (see the homework)

$$\Pr(\text{A or B}|\text{E}) = \Pr(\text{A}|\text{E}) + \Pr(\text{A}^F|\text{E})\Pr(\text{B}|\text{E}) \tag{10}$$

and because of symmetry this is also

$$\Pr(\text{A or B}|\text{E}) = \Pr(\text{B}|\text{E}) + \Pr(\text{B}^F|\text{E})\Pr(\text{A}|\text{E}). \tag{11}$$

Isn't that wild? In words, the chance of "A or B" being true is the chance of A being true plus the chance of B being true times the chance that A is false. Or it's the probability (all given E) that "A is true or A is false and B is true." Or, the probability (given E) that "We do not see a 1 when we roll the die or We do see a 1 and it snows". I don't know about you, but I don't find this result immediately intuitive, which is why you have to be careful is assessing probabilities!

An excellent (classical) book to learn more about probability rules is Ross (1988).

6. Homework

(1) Suppose aliens from the planet Thorsten have just developed a new and improved probe, which they decide to test on humans. They only want the best and brightest, and so have chosen to sample humans from an introductory statistics class. In this class are 12 females and 8 males. On their first run, the aliens abduct just 1 person. What is the aliens' E, and what are the chances the student they snatch is a male?

(2) On their second run, after fixing an unfortunate side effect of the probe that was discovered after the first abduction went awry, they decide to get a larger sample, and want to be sure to get both males and females, but they also don't want to re-use the guy they took before, for obvious reasons. They want two females and two males. F_1 and F_2 (two females who have numbers for names), are best friends. What are the chances that they are taken together (from their perspective)? Their brothers, M_1 and M_2, are also in the class. What are the chances that they are *also* taken? HINT: Take a deep breath and relax; you can do this problem. First start with considering that, out of the females, they abduct F_1; then consider, having got her, the aliens abduct F_2, and so on.

(3) A year has passed and the memory blocks the aliens have used to stop the abductees from blabbing have worked. They will grab one more student and bring this student back to the Home World for their zoo. Gort argues that they should "randomly sample" the students, and Klaatu is fine with

just plain "grabbing one" from them. What is the probability that the student named C is abducted under Gort's plan and under Klaatu's plan; explain.

(4) Let your evidence be E = "This is a coin, only one side of which is an H." Suppose you toss this coin and see that it turns up heads 40 times out of 50. You want to toss is one more time. What is your estimate of $\Pr(H|E)$ and why?

(5) Prove to me that, given the same E as before the probability of any coin toss sequence, in n tosses, is the same. An example of a sequence of three tosses is HHT; one for four tosses is TTHT, and so on.

(6) What is the probability of getting no heads in $n = 4$ tosses of a coin. Assume the same E.

(7) EXTRA Find a general formula for the probability, when tossing a 20-sided die,[2] of getting *at least* one "20" in any n tosses. HINT The probability that A is true is one minus the probability that A is false.

(8) List one situation where you have an event or observation statement A and another B such that the probability of A being true is irrelevant to knowledge that B is true. That is, that $\Pr(A|BE)\Pr(B|E) = \Pr(A|E)\Pr(B|E)$. Write out your A, B, and E carefully. Most gambling situations fit this scenario well.

(9) List one situation where you have an event or observation statement A and another B such that the probability of A being true is *relevant* to knowledge that B is true. That is, that $\Pr(A|BE)\Pr(B|E) \neq \Pr(A|E)\Pr(B|E)$. Write out your A, B, and E carefully. Many physical situations fit this scenario well.

(10) In the fortune cookie example, we estimated that $\Pr(B_+|A_+E) \approx 0.08$. What if instead of positive, the test came back negative? That is, what is $\Pr(B_+|A_-E)$? Where "A_-" means a negative test. First state what this probability means in words.

(11) If you really understand the fortune cookie problem, then you'll be able to tell me what was your friend's estimate of the probability that he has lump cancer *before* he went to the doctor and *before* he read the fortune *and* given your friend knew the sensitivity, specificity, etc. that went into calculating $\Pr(B_+|A_+E)$? HINT: This is not a trick question.

(12) *Odds* are one-to-one functions of probability. To calculate the odds, use this formula $odds = \frac{p}{1-p}$ where p is the probability of interest. What are the odds of $\Pr(B_+|AE)$ and $\Pr(B_+|A^FE)$.

(13) *Odds ratios* are simply the odds of one thing divided by another. What is the odds ratio of having a cancer given a positive test versus not having a positive test? That last quantity measures the multiplicative increase in the odds.

(14) EXTRA Prove the formulas (10) and (11) are correct.

[2]Sure these exist; haven't you ever played Dungeons and Dragons?

CHAPTER 3

How to Count

1. One, two, three...

Youtube.com has a video at this URL

`http://www.youtube.com/watch?v=wcCw9RHI5mc`

The important part is that "`v=wcCw9RHI5mc`" business at the end, which essentially means "this is video number `wcCw9RHI5mc`". This video is, of course, different than number `wcCw9RHI5m`*d*, and number `wcCw9RHI5m`*e* and so on. This might be a new way to use the word *number* to you, but these are numbers just the same, only the symbols used to write them have changed from the familiar 0-9. We can notice that the video number contains 11 different slots (count them), each of which is filled with a number or upper or lower case Latin letter, which means the number is case sensitive; `A` differs from `a`. The question is, how many different videos can YouTube host given this numbering scheme? Are they going to run out of numbers anytime soon?

That problem is hard, so we'll start on a simpler one. Suppose the video numbering scheme only allowed one slot, and that this slot could only contain a single-digit number, chosen from 0-9. Then how many videos could they host? They'd have `v=0`, `v=1` and so on. Ten, right? Now how about if they allowed two slots chosen from 0-9. Just 10 possibilities for the first, and 10 for each of the 0-9 of the first (10 for 0, 10 for 1, etc.), a confusing way of saying 10×10. For three slots it's $10 \times 10 \times 10$. But you already knew how to do this kind of counting, didn't you? That's how we write numbers!

Suppose the single slot is allowed only to be the lower case letters `a,...,z`? This is `v=a`, `v=b`, etc. How many in two such slots? Using what we just learned, it is $26 \times 26 = 676$. This result was had by the same way we got 100 in two slots of the numbers 0-9.

So if allow any number, plus any lower or upper case letter in any slot, we have $10 + 26 + 26 = 62$ different possibilities per slot. That means that with 11 slots we have $62 \times 62 \cdots \times 62 = 62^{11} \approx 5 \times 10^{19}$, or 50 billion billion different videos that YouTube can host.

2. Arrangements

How many ways are there of arranging things? In 1977, George Thorogood remade that classic John Lee Hooker song, "One Bourbon, One Scotch,

and One Beer." This is because George is, of course, the spirituous counterpart of an oenophile; that is, he is a connoisseur of fine spirits and regularly participates in tastings. Further, George, who is way past 21, is not an idiot and never binge drinks, which is about the most moronic of activities that a person could engage in. He very much wants to arrange his coming week, where he will taste, each night, one bourbon (B) , one scotch (S), and one beer (R). But he wants to be sure that the order he tastes these drinks doesn't influence his personal ratings. So each night he will sip them in a different order. How many different nights will this take him? Write out what will happen: Night 1, BSR; night 2, BRS; night 3, SBR; night 4, SRB; night 5, RBS; night 6, RSB. Six nights! Luckily, this still leaves Sunday free for contemplation.

Later, George decides to broaden his tasting horizons by adding Vernors (the tasty ginger ale aged in oak barrels that can't be bought in New York City) to his line up. How many nights does it take him to taste things in different order now? We could count by listing each combination, but there's an easier way. If you have n items and you want to know how many different ways they could be grouped or ordered, the general formula is:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 \tag{12}$$

The term on the left, $n!$, reads "n factorial." With 4 beverages, this is $4 \times 3 \times 2 \times 1 = 24$ nights, which is over three weeks! Good thing that George is dedicated.

3. Being choosy

It's the day before Thanksgiving and you are at school, packing your car for the drive home. You would have left a day earlier, but you didn't want to miss your favorite class—statistics. It turns out that you have three friends who you know need a ride: Larry, Curly, and Moe. Lately, they have been acting like a bunch of stooges, so you decide to tell them that your car is just too full to bring them along. The question is, how many different ways can you arrange your friends to drive home with you when you plan to bring none of them? This is not a trick question; the answer is as easy as you think. Only one way—that is, with you driving alone.

But, they are your friends, and you love them, so you decide to take just one. Now how many ways can you arrange your friends so that you take just one? Since you can take Larry, Curly, or Moe, and only one, then it's obviously three different ways, just by taking only Larry, or only Curly, or only Moe. What if you decide to take two, then how many ways? That's trickier. You might be tempted to think that, given there are 3 of them, that the answer is $3! = 6$, but that's not quite right. Write out a list of the groupings: you can take Larry & Curly, Larry & Moe, or Moe & Curly. That's three possibilities. The grouping "Curly & Larry," for example, is just the same as the grouping "Larry & Curly." That is, the order of your friends doesn't matter: this is why the answer is 3 instead of 6. Finally, all

these calculations have made you so happy that you soften your heart and decide to take all three. How many different groupings taking all of them are possible? Right. Only one: by taking all of them.

You won't be surprised to learn that there is a formula to cover situations like this. If you have n friends and you want to count the number of possible groupings of k of them when the order does not matter, then the formula is

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{13}$$

The term on the left is read "n choose k". By definition (via some fascinating mathematics) $0! = 1$.

Here are all the answers for the Thanksgiving problem:

$$\binom{3}{0} = \frac{3!}{3!0!} = 1 \qquad \binom{3}{1} = \frac{3!}{2!1!} = 3$$
$$\binom{3}{2} = \frac{3!}{1!2!} = 3 \qquad \binom{3}{3} = \frac{3!}{1!3!} = 1$$

Don't worry about the math. The computer will do it for you (we'll talk about how in Chapter 5). But there are some helpful facts about this combinatorial function that are useful to know. The first is that $\binom{n}{0}$ always equals 1. This means, out of n things, you take none; or it means there is only one way to arrange no things, namely no arrangement at all. $\binom{n}{n}$ is also always 1, regardless of what n equals. It means, out of n things, you take all. $\binom{n}{1}$ always equals n, and so does $\binom{n}{n-1}$: these are the number of ways of choosing just 1 or just $n-1$ things out of n. As long as $n > 2$, $\binom{n}{2} > \binom{n}{1}$, which makes sense, because you can make more groups of 2 than of 1.

4. The Binomial distribution

We started the Thanksgiving problem by considering it from *your* point of view. Now we take Larry, Moe, and Curly's perspective, who are waiting in their dorm room for your call. They don't yet know whether which, or if any of them, will get a ride with you. Because they do not know, they want to quantify their uncertainty and they do so using probability. We are now entering a different realm, where counting meets probability. Take your time here, because the steps we follow will the same in *every* probability problem we ever do.

Moe, reminiscent, recalls an incident wherein he was obliged to poke you in the eyes, and guesses that, since you were somewhat irked at the time, the probability that you take any one of the gang along is only 10%. That is, it is his judgment that the probability that you take him, Moe, is 10%, which is the same as you would also (independently) take Curly and so on. So the boys want to figure out the probability that you take none of them, take one of them, take two of them, or take all three of them.

Start with taking all three. We want the probability that you take Larry *and* Moe *and* Curly, where the probability of taking each is 10%. Remember probability rule #2? Those "ands" become "times": so the probability of taking all three is $0.1 \times 0.1 \times 0.1 = 0.001$, or 1 in a 1000. Keep in mind: this is from *their* perspective, not yours. This is their guess of the chances; because you may already have made up your mind—but they don't know that. Remember that all probability is conditional on evidence, and that the Stooges' evidence is difference that yours.

What about taking none of them? This is the chance that you do not take Larry *and* you do not take Moe, *and* you do not take Curly. This is because taking Larry *means* not taking the other two. The key word is still "and;" which makes the probability $(1-0.1)\times(1-0.1)\times(1-0.1) = 0.9^3 \approx 0.73$, since the probability of not taking Larry etc. is one minus the probability of taking him etc. It is, too, because you can either take Larry or not; these are the only two things that can happen, so the probability of taking Larry or not *must* be 1. We can write this using our notation: let A = "Take Larry", then A^F = "Don't take him". Then $\Pr(\mathrm{A} \cup \mathrm{A}^F|\mathrm{E}) = \Pr(\mathrm{A}|\mathrm{E}) + \Pr(\mathrm{A}^F|\mathrm{E}) = 1$, using probability rule #1. So if $\Pr(\mathrm{A}|\mathrm{E}) = 0.1$, then $\Pr(\mathrm{A}^F|\mathrm{E}) = 1 - \Pr(\mathrm{A}|\mathrm{E}) = 0.9$. In this case, E is the information dictated by Moe (who is the leader), which caused him to say $\Pr(\mathrm{A}|\mathrm{E}) = 0.1$.

How about taking just one? Well, one way this can happen is that you can take Larry, not take Moe, and not take Curly, and the chance of that is (using rules #1 and #2 together) $0.1 \times (1 - 0.1) \times (1 - 0.1) \approx 0.08$; but you could just as easily have taken Moe and not Larry, or Curly and not Larry, and the chance you do either of these is just the same as you taking Larry and not the other two. For shorthand, write M as "Take M" and so on, and M^F as not take M and so on. Thus you could "$\mathrm{LM}^F\mathrm{C}^F$ *or* $\mathrm{L}^F\mathrm{MC}^F$ *or* $\mathrm{L}^F\mathrm{M}^F\mathrm{C}$" (this is written using the Boolean algebra). Using probability rule #1, we break up this statement into three pieces ("$\mathrm{LM}^F\mathrm{C}^F$"), and then use probability rule #2 on each piece ("ands" turn to times), then add the whole thing up.

You could do all that, but there is an easier way. You could notice there are three different ways to take just one—which we remember from our choosing formula, eq. (13). This makes the probability $\binom{3}{1}0.08 = 3 \times 0.08 = 0.24$. Since we already know the probability of taking one of those combinations, we just multiply it by the number of times we see it. We could have also written the answer like this:

$$\binom{3}{1}0.1^1(1-0.1)^2 = 0.24.$$

And we could also written the first situation (taking all of them) in the same way

$$\binom{3}{0}0.1^3(1-0.1)^0 = 0.001.$$

where you must remember that $a^0 = 1$ (for any a you will come across, even $a = 0$).

You see the pattern by now. This means we have another formula to add to our collection. This one is called the binomial and it looks like this:

$$\Pr(k|n, p, \mathrm{E}_B) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{14}$$

There is a subtle shift in notation with this formula, made to conform with tradition. "k" is shorthand for the statement, in this instance, K = "You take k people." For general situations, k is the number of "successes": or, K = "The number of successes is k". Everything to the right of the "|" is still information that we know. So n is shorthand for N = "There are n possibilities for success", or in your case, N = "There are three brothers which could be taken." The p means, P = "The probability of success is p". We already know E_B, written here with a subscript to remind us we are in a binomial situation. This new notation can be damn convenient because, naturally, most of the time statisticians are working with numbers, and the small letters mean "substitute a number here," and if statisticians are infamous for their lack of personality, at least we have plenty of numbers. This notation can cause grief, too. Just how that is so must wait until later.

Don't forget this: in order for us to be able to use a binomial distribution to describe our uncertainty, we need three things. (1) The definition of a *success*: in the Thanksgiving example, a success was a person getting a ride. (2) The probability of a success is always the same for every opportunity. (3) The number of chances for successes is fixed.

5. Homework

(1) It turns out that YouTube actually allows more than just numbers and letters in their video numbers. They also use the symbol "_" (the underscore). Now how many videos can they host?
(2) In the 23 April 2008 *Wall Street Journal*, on the front page, Ford Motor Company CEO Alan Mulally complained about the "mind-boggling level of vehicle customization, which jacked up costs. Until recently, for instance, the Lincoln Navigator offered 128 options on its console alone." How many differently optioned Lincoln Navigators can be built if all the vehicles are the same except for differences in the console?
(3) The daily lottery in New York requires you to pick three different numbers to win. How many different combinations of three numbers are there? What are the chances you win?
(4) You just got a new dorm room, and have three roommates, and two bunk beds. How many different sleeping arrangements are there, assuming, I hope it isn't necessary to say, one per bed? Later, one of your roommates (a football linesman who rarely bathes) insists, in an emphatic way, that he *must* have the top bunk facing East. How many arrangements now?
(5) The FAA uses three-capital-letter designators for airport codes; for example, LGA is La Guardia and DTW is Detroit Metro. How many unique airport codes can there be?

(6) You are staying away from home at college for the first time, and have decided to re-invent yourself. Nobody here knows that you were the kid that mistakenly ate part of your classmate's biology project. Time to start fresh. So you buy an entire new wardrobe, consisting of six shirts, and three pants. Assuming you'll have to wear one shirt and one pair of pants to create an ensemble, how many different ensembles can you wear?

(7) You have discovered that pair of pants 1 does *not* go with shirts 3 and 4; and that pants 2 does not go with shirts 1, 5, or 6. How many ensembles are now possible?

(8) You are a generous soul and decide to forgive Moe and decide to take all the gang with you. How many seating arrangements are there, assuming you drive the car?

(9) Part of the binomial formula is $p^k(1-p)^{n-k}$. Explain how this part comes about, that is, why is it this and not something else? I mean, why are the exponents on p and $(1-p)$ k and $n-k$. Can you explain it in terms of the probability rules you already know? Using the Three Stooges example as a starting point.

(10) Another lottery question. The multiple-state Mega Millions drawing requires you to guess 5 different numbers from 1 to 56. It also requires you to pick a "mega" number (after those 5) from 1 to 46. In the first case, what is the chance that you guess the first number correctly? And the second? And third through fifth? And the "mega"?

(11) EXTRA: Obviously, the order you guess the balls do not matter: if you match all 6 you will win. The question before assumes you guess the numbers in the order that they were drawn. Can you think of a way to calculate the probability of winning, that is, of matching all 6, where the order the numbers were drawn do not matter.

(12) EXTRA: Greek sororities and fraternities are designated by two or three Greek letters, like $\Gamma\Omega$ or $\Sigma\Pi\Delta$. How may Greek unique societies are possible?

(13) EXTRA: Suppose Moe estimates the probability that you take him as 0.1, and Larry too. But since Curly knows he's so lovable, he estimates his probability of going at 0.8. What is the probability (they estimate) that you take none, one, two, or all three. HINT: Do *not* use the binomial.

(14) EXTRA: You can see that $\binom{n}{2} > \binom{n}{1}$. With some playing around, it's easy to see that $\binom{n}{3} > \binom{n}{2}$. But it's also true that $\binom{n}{n-1} > \binom{n}{n}$ and $\binom{n}{n-2} > \binom{n}{n-1}$. Can you find an m such that $\binom{n}{m}$ is larger than any other $\binom{n}{k}$ where $k \neq m$? If you cannot find it, at least make a guess and give a reason why you chose that guess.

CHAPTER 4

Distributions

1. Variables

Recall that *random* means *unknown*. Suppose x represents the number of times the Central Michigan University football team wins next year. Nobody knows what this number will be, though we can, of course, guess. Further suppose that the chance that CMU wins any individual game is 2 out of 3, and that (somewhat unrealistically), a win or loss in any one game is irrelevant to the chance that they win or lose any other game. We also know that there will be 12 games. Lastly, suppose that this is *all* we know. Label this evidence E. That is, we will ignore all information about who the future teams are, what the coach has leaked to the press, how often the band has practiced their pep songs, what students will fail their statistics course and will thus be booted from the team, and so on. What, then, can we say about x?

We know that x can equal 0, or 1, or any number up to 12. It's unlikely that CMU will lose or win every game, but they'll probably win, say, somewhere around 2/3s, or 6-10, of them. Again, the exact value of x is random, that is, unknown.

Now, if last chapter you weren't distracted by texting messages about how great this book is, this situation might feel a little familiar. If we instead let x (instead of k—remember these letters are place holders, so whichever one we use does not matter) represent the number of classmates you drive home, where the chance that you take any of them is 10%, we know we can figure out the answer using the binomial formula. Our evidence then was E_B. And so it is here, too, when x represents the number of games won. We've already seen the binomial formula written in two ways, but yet another (and final) way to write it is this:

$$x|n,p,\mathrm{E}_B \sim \text{Binomial}(n,p). \tag{15}$$

This (mathematical) sentence reads "Our uncertainty in x, the number of games the football team will win next year, is best represented by the Binomial formula, where we know n, p, and our information is E_B." The "$\sim$" symbol has a technical definition: "is distributed as." So another way to read this sentence is "Our uncertainty in x is distributed as Binomial where we know n, etc." The "is distributed as" is longhand for "quantified." Some people leave out the "Our uncertainty in", which is OK if you remember it is

there, but is bad news otherwise. This is because people have a habit of imbuing x itself with some mystical properties, as if "x" itself had a "random" life. Never forget, however, that it is just a placeholder for the statement X = "The team will win x games", and that this statement may be true or false, and it's up to us to quantify the probability of it being true.

In classic terms, x is called a "random variable". To us, who do not need the vague mysticism associated with the word random, x is just an unknown number, though there is little harm in calling it a "variable," because it *can* vary over a range of numbers. However, all classical, and even much Bayesian, statistical theory uses the term "random variable", so we must learn to work with it.

Above, we guessed that the team would win about 6-10 games. Where do these number come from? Obviously, based on the knowledge that the chance of winning any game was 2/3 and there'd be twelve games. But let's ask more specific questions. What is the probability of winning no games, or X = "The team will win $x = 0$ games"; that is, what is $\Pr(x = 0|n, p, \mathrm{E}_B)$? That's easy: from our binomial formula, this is $\binom{n}{k}p^k(1-p)^{n-k} = \binom{12}{0}0.67^0(1-0.67)^{12} = (1 - 0.67)^{12} \approx 2$ in a million. We don't need to calculate $\binom{n}{0}$ because we know it's 1; likewise, we don't need to worry about 0.67^0 because we know that's 1, too. What is the chance the team wins all its games? Just $\Pr(x = 12|n, p, \mathrm{E}_B)$. From the binomial, this is $0.67^{12} \approx 0.008$ (check this). Not very good!

Recall we *know* that x can take any value from zero to twelve. The most natural question is: what number of games is CMU most likely to win? Well, that's the value of x that makes $\binom{12}{x}0.67^x(1 - 0.67)^{12-x}$ the largest, i.e. the most probable. This is easy for a computer to do (you'll learn how in Chapter 5). It turns out to be 8 games, which has about a one in four chance of happening. We could go on and calculate the rest of the probabilities, for each possible x, just as easily.

What is the most likely number of games the team will win is the most natural question for us, but in pre-computer classical statistics, there turns out to be a different natural question, and this has something to do with creatures called *expected values*. That term turns out to be a terrible misnomer, because we often do not, and cannot, expect any of the values that the "expected value" calculations give us. The reason expected values are of interest has to do with some mathematics that are not of especial interest here; however, we will have to take a look at them because it is expected of one to do so.

Anyway, the expected value for any *discrete* distribution, like the binomial, is calculated like this:

$$E_x(x) = 0 \times \Pr(x = 0|\mathrm{E}) + 1 \times \Pr(x = 1|\mathrm{E}) + \cdots + n \times \Pr(x = n|\mathrm{E}) \quad (16)$$

where discrete means that x can only take on measurable, actual values (there are other distributions that are called *continuous* which I'll describe below). The *expectation* (another name for it) is the sum of every value that x

can be times the probability that x takes those numbers. Think of it as a sort of probability-weighted average of the xs. The little sub $_x$ on the expected values means "calculate the expected value of the variable with respect to x"; that is, calculated $E(x)$ with respect to the probability distribution of x. Incidentally, we can also calculate $E_x(x^2)$ or $E_x(g(x))$, where $g(x)$ is some function of x that might be of interest to us, and sometimes it can get confusing what we're doing, hence placing the subscript as a reminder. As always, it is important to be precise.

Turns out that there is a shortcut for the binomial, which is $E_x(x) = np$. So, for the CMU team, $E_x(x) = 12 \times \frac{2}{3} = 8$...which sounds like I'm complaining about nothing, because this is the same as the most likely number of games won! But what if the probability of winning individual games was 3/5 instead of 2/3? Then (a computer shows us) the most likely number of games won is 7, but the expected value is $E_x(x) = 12 \times \frac{3}{5} = 7.2$. Now, according to the rules of football as I understand them, you can only win whole games; that is, winning the expected number of games is an impossibility.

There is another quantity related the expected value called the *variance*. It has a similar birth story and a precise mathematical definition, which for discrete distributions is (don't get overly concerned about the math)

$$\begin{aligned} V_x(x) &= E_x((x - E_x(x))^2) \\ &= (0 - E_x(x))^2 \times \Pr(x = 0|\mathrm{E}) + \\ &\quad \cdots + (n - E_x(x))^2 \times \Pr(x = n|\mathrm{E}). \end{aligned} \tag{17}$$

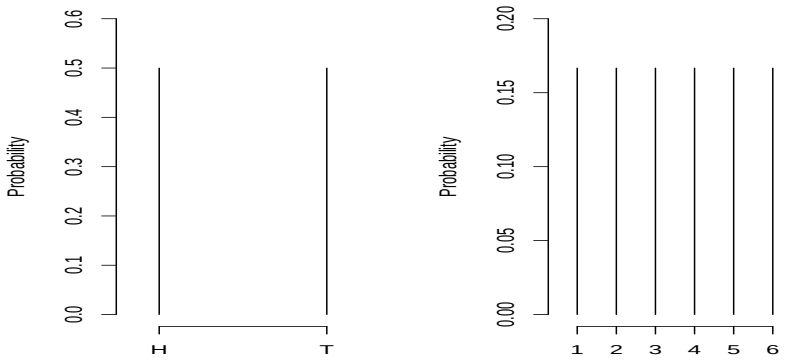
It's purpose is to give some idea of the precision of the expected value. Look at the definition: it is a function of the value of x minus the "expected" value of x, for each possible value of x (that's the outer expectation). High values of variance, relative to the expected value, imply that the expected value is imprecise; low values have the opposite implication. There is a binomial shortcut to the variance: $V_x(x) = np(1-p)$. For the CMU football example, $V(x) = 12 \times 0.67 \times 0.33 \approx 2.7$.

Why talk about expected values and variances when they are not terribly informative? Well, let's be generous and recall that these theoretical entities had great value in the days before computers. Nowadays, we can easily calculate the probability that x equals any number, but back in the technolithic days this could only have been done with great effort. Besides, the expected value is not too far from the most likely value, and is even the same sometimes. The variance gives an idea of the plus and minus range of the expected value, that is, the most likely values x could take. And you could do it all on the back of an envelope! But since expectations still fill pages of nearly every statistics book, you at least have to be aware of them. Next, we learn how to quantify uncertainty the modern way.

2. Probability Distributions

Remember what will be our mantra: *if we do not know the truth of a thing, we will quantify our uncertainty in that thing using probability.* Usually, we will use a probability distribution, like the binomial. A probability distribution *gives us the probability for every single thing that can happen in a given situation.*

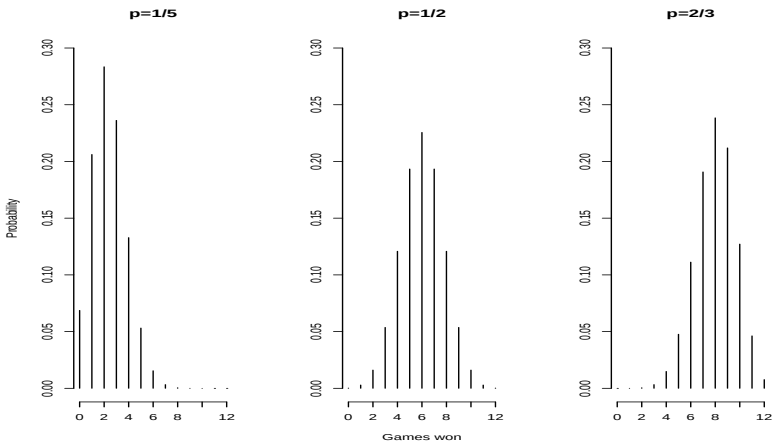
You already know lots of probability distributions (they go by the technical name "probability mass functions" for discrete data), you just didn't know they were called that. Here are two you certainly have memorized, shown in pictures:



The first is for a coin flip, where every single thing that can happen in a H(ead) or T(ail). The information we are given is E_{coin} = "This is a coin with two sides labeled H and T, and a flip will show one of them." Given *this* information and no other, we get the picture on the left, which shows the distribution of probability for every single thing that can happen. Easy, right? It's just a spike at 0.5 for an H, and another at 0.5 for a T. The total probability is the sum of the spikes, or 1.

The second is for the roll of a die, where every single thing that can happen is a 1, or 2, or etc. The information is E_{dice} = "This is a die with six sides labeled 1, 2,...,6, and a roll will show one of them." Given just this information, we get the picture with a spike of 1/6 for every possible number. Again, the total probability is the sum of the spikes, which is still 1. It is *always* equal to 1 for any probability distribution.

We can also picture the binomial for the CMU football victories.



Here, it is drawn for three possible values of p: $p = 1/5$, $p = 1/2$, and $p = 2/3$. Every single thing that can happen is that CMU wins 0 games, 1 games, etc., up to 12 games. The information we are given is E_B = "The probability of winning any individual game is fixed at exactly p (=1/5, 1/2, or 2/3), there are $n = 12$ games, winning or losing any game gives no information about winning or losing any others, and we will use a binomial distribution to represent our uncertainty." If $p = 1/5$, you can see that there is at least a reasonable chance, about 7%, that CMU wins no games, while winning all games is so improbable that it looks close to 0.

Wait a minute, though. It is *not* 0. It just looks like 0 on this picture. The total number of games won by CMU is contingent on certain facts of the universe being true (like the defense not being inept, the quarterback not being distracted by job proposals or cheerleaders, and so on). Remember that the probability of any contingent event is between 0 and 1; it is never exactly 0 or 1. So even though the picture shows that winning all games when $p = 1/5$ looks 0, it is not, because that would mean that winning all 12 is impossible. To say something is impossible is to say it has probably 0, which we know we cannot be so for a contingent event. Incidentally, using the computer shows that the probability of winning at 12 games is about 4e-09, which is a decimal point, followed by eight 0s, then a 4, or 0.000000004. Small, but greater than 0.

The most likely number of games won, with $p = 1/5$, is 2—there is about a 28% chance of this happening. What is the expected value? And variance? (Not very interesting numbers, are they.)

Notice that when we switch to $p = 1/2$, the chance of winning games becomes *symmetric* around 6, the most likely number won. This means that the chance of winning all 12 is equal to the chance of winning none. Does it also mean the chances of winning 1 is the same as winning 11?

When $p = 2/3$, the most likely number of games won is again 8, but right behind that in probability is 9 games, which is actually more likely than winning 7, and so on.

The reason to show three pictures at different values of p is because we don't know what the value of p is, but E_B requires that we specify a known value of p, else we cannot draw the picture. We learn how to guess the value of p later.

3. What is Normal?

What will be tomorrow's high temperature? This value is, of course, unknown. But we can always guess. Suppose we guess x° C. Are we *certain* it will be x° C? No. It may be a little higher, it may be a little lower. It's unlikely to be too high or too low, or too far from x° C. So, the question you're undoubtedly asking yourself is: "Hasn't some brilliant and intriguing statistician come up with a way that I can quantify my uncertainty in x?" Why, yes, of course (and aren't all statisticians brilliant and intriguing?). It's called the normal, sometimes a Gaussian, distribution.

This distribution is different than a binomial in many ways. With the binomial, we had a fixed number of chances, or trials, for successes to occur. With the normal, there is no such thing as a success, and no fixed number of chances, except for one: the outcome itself. The binomial was useful for discrete numbers, while the normal is used for...something else, to be discussed below.

Here is one of the ways we can write it:

$$x|m, s, \mathrm{E}_N \sim \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-m)^2}{2s^2}}. \tag{18}$$

m is called the *central parameter* and s^2 is called the *spread parameter*: sometimes, s, the square root of s^2, is called the *standard deviation parameter* or *standard deviation parameter*. Some books will, using sloppy language, call m *the* mean and s *the* standard deviation. You will *never* make this mistake! The mean and standard deviation are entirely different creatures (we learn about them later). The e is equal to about 2.718, and π is about 3.142. Anyway, it is a sufficiently complicated formula such that we'll never calculate it by hand.

Let's review this to make certain where we are. Just like using a binomial, the x is shorthand for X = "The value of the temperature will be x." Certain information is given as known: the m and the s^2, plus E_N = "We use a normal distribution to quantify our uncertainty in X." Looking at the formula (18) might not show it, but there are some screwy things going on with the normal. First recall that the probability distribution gives us the probability of every single thing that can happen. So just what is every single thing that can happen in a normal distribution?

Well, (this is true for any situation that uses a normal and not just the temperature example), x can equal, say, 18.0000001, or 18.0000000001, or -19.124828184341, and on and on. Turns out, with the mathematical device used for creating normal distributions, an infinity of things can happen: every number between negative and positive infinity is possible with a normal.

How many numbers is that? So many that they can't be counted. Numbers like this are said to be *continuous*, that is, there is an unbroken continuity between any two numbers a and b. How many numbers are there between $a = 17$ and $b = 82$? An infinity. How many between $a = 6.01$ and $b = 6.1$? An infinity. The binomial, on the other hand, used discrete numbers, where there is a definite space between numbers (the outcome could *only* be certain, fixed numbers, there was no continuity).

Normal distributions are used to specify the uncertainty of a wide range of variables: from blood pressures, to real estate prices, to heat in chemical equations, to just about anything. But there is a huge problem with this. Recall our primary purpose in using probability distributions: they are meant to quantify our uncertainty in some thing about which we do not know the value. Normal distributions, though ubiquitous, *never* accurately capture our uncertainty in any real life X.

This is because the uncertainty in any real-life thing cannot be exactly quantified by a normal, because no real thing has an infinite number of possible values. Also, no real thing, like temperature, has a maximum going toward positive infinity, nor a minimum going toward negative infinity. We can accurately measure outdoor temperature to maybe a tenth of even a hundredth of a degree (eg. 18.11°C). But we cannot measure it to infinite precision. And the temperature of any thing can never be less than absolute zero (given certain physical arguments), and certainly cannot be infinitely high.

All these complications mean that equation (18) isn't real a probability distribution at all: instead, it is called a *density*. We first have to make it into a probability (via some hidden mathematics). When we do, any normal distribution says that

$$\Pr(x|m, s, \mathrm{E}_N) = 0.$$

In English, the probability that x takes any value is always 0, no matter what the value of x is, no matter what m is, and no matter what s is. The probability that x equals *any* number is always 0 (no continuous number can be measured to infinite precision). To help see this, imagine I pick a number out of an infinite number of choices. What are the chances that you guess this number correctly? Zero. Even worse, I cannot even really pick my number! Some (an infinite amount of) continuous numbers cannot even be measured, though we know how to compute them; that is, nobody can ever fully write one down, because that would require writing down an infinite number of digits. Worse still, most (another, larger kind of infinity of) continuous numbers, we don't even know how to calculate their digits! Incidentally, not all mathematicians are happy about using these kinds of numbers. After all, if you cannot actually write down or discover a number, it has little use in measuring real things. See the books by Chaitin (2005) and Kline (1980) for more information on this odd subject.

Continuous numbers are a major burden, which seems to put us at an impasse. Since we can't answer questions about the truth of statements like X = "The value of tomorrow's maximum temperature will be x", we are forced to change the question and instead ask about intervals. For example, X = "The value of tomorrow's maximum temperature will be *less than* x." X no longer makes a statement about a single number, but a range of numbers, namely, all those less than x (how many numbers are less than x?). Other examples of intervals: all the numbers between 0 and 1; all numbers smaller than 4; all numbers less than 17 *and* greater than 52; etc. Pick any two numbers, and as long as they are not the same, you have an interval. Then, for example,

$$\Pr(x < 4|m, s, \mathrm{E}_N) = a$$

(where a is some real number) can be answered. Again, to emphasize, we cannot ascertain the truth of statements like X = "The value of tomorrow's maximum temperature will be 20°C." We can only quantify the uncertainty of statements about intervals like X = "The value of tomorrow's maximum temperature will be *less than or equal to* 20°C."

If the normal can't handle questions we need answered, like giving us the probability of single numbers, why is it used? The biggest reason is habit, another is ignorance of any alternative. But there's more to it than that. Let's go back to our temperature example to see why. We know, say, in our situation that we can measure temperature to the nearest tenth of a degree. We can even suppose that temperature can only even *be* at every tenth degree[1], so that the temperature can be 20°C or 20.1°C, but it *cannot* be, say, 20.06°C or any other numbers that aren't even tenths of a degree. Using a normal distribution to represent our uncertainty will give probability to statements like Y = "Tomorrow's temp will be between, and not including, 20°C or 20.1°C." We then *know* that this probability is 0, which is to say, the statement is false, which we know based on our knowledge that temperature can only be at tenths of a degree. But the normal will say something like $\Pr(20^\circ < y < 20.1^\circ|m, s, \mathrm{E}_N) = 0.0001$ (it is saying there *is* a probability of seeing values that are impossible). Although this is a mistake, is it a big one?

"Ah, so what," you say to yourself, "this is so small a probability as not to be worthy of my attention. The normal distribution will give me an answer that is *close enough*." You might be right, too. In later Chapters, we'll have to see if the normal makes a reasonable approximation to what we really need. Besides, if you don't want to use a normal distribution, you still have to use something. What?[2] Using a normal distribution does allow you to bypass two very tricky problems. Remember that a normal distribution, regardless

[1]There is plenty of evidence the universe is set up so that temperature, and every other physical variable, is discrete like this: that is, continuous numbers are mathematical, not physical, creatures.

[2]This is a poor argument. Because we don't know what distribution to use, does not mean we should use anything we can get our hands on.

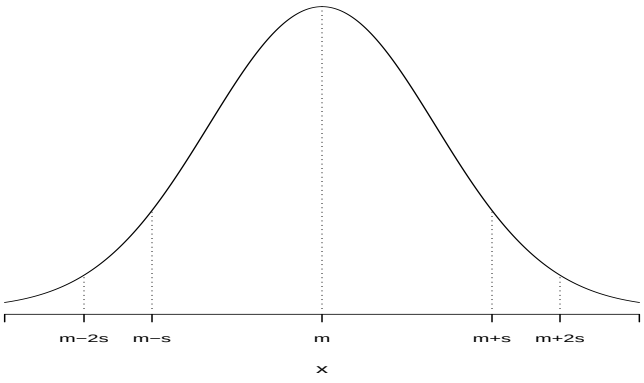
of the values of m or s, says something about *all* numbers going towards positive and negative infinity. If you eschew the normal in the temperature example, then you at least have to say what are the maximum and minimum possible temperature. Do you know? I mean, do you know with certainty?

Actually, in many cases, you will know with certainty, or to within a certain tolerance. Temperature has to be measured by a device, like a thermometer. That thermometer will not be able to register temperatures below some number, nor will it be able to register above some number. Many cases are like this: the situation itself dictates the limits. Later, we'll look at test scores, which also have a built in limit (0 and 100).

But if you cannot ascertain the limits, then you have, in a sense, double uncertainty: the future value of the temperature, plus some uncertainty in the distribution you use in representing that uncertainty. This situation is already beyond most statistics books, even the tough ones, so for now, until we talk about the subject of modelling, we'll ignore this question and say that the normal is "close enough." But we will *always* remember that it will *never* be best.

Whew. A lot of facts and we haven't even thought about our example. So why bring them up? To show you now that people are too certain of themselves. Normal distributions are so often used by that great mass of people who compute their own statistics, that you might think there are no other distributions. Since we now know that normals can only be, at best, approximations, when we hear somebody authoritatively state a statistical result must be believed, and we learn they used a normal to quantify their uncertainty, we know they are too confident. We'll meet a lot more of this kind of thing as we go along.

On to something more concrete. Let's look at an abbreviated picture of a normal distribution and see what we can learn (it is abbreviated because we cannot picture the whole thing). The point m is the central point, and is the most likely value (not forgetting that no single value is actually possible—just weird, right?); m plus or minus s contains about 68% of all possible values of x, and m plus or minus about 2 times s contains about 95% of all possible values of x.

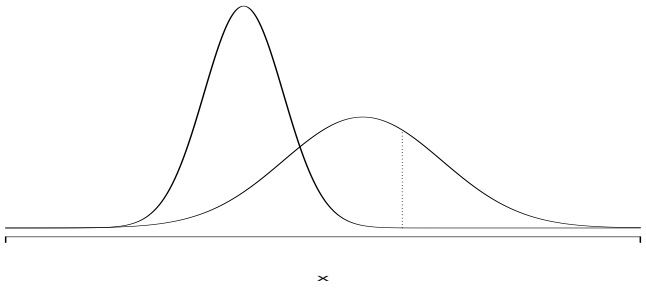


What does this mean? Specifically, that $\Pr(m - s < x < m + s|m, s, \mathrm{E}_N) \approx 0.68$ and $\Pr(m-2s < x < m+2s|m, s, \mathrm{E}_N) \approx 0.95$. The normal is symmetric about the point m: meaning there is equal probability of being above m as being below it. Incidentally, the expected value of x is m, and the variance s^2, which is easy to remember.

You might have noticed that there is no y-axis on the picture. This is to remind you that the curve itself does *not* represent probability. The (missing) y-axis is instead something called the *density*, which is the continuous-number equivalent of probability. It cannot picture the probability, because the probability of x equaling *any* number is 0. Because of this, these pictures are only useful for estimating probability by area under the curve. The entire area under the curve equals 1, just as it did with the coin flip and dice example, because this picture shows the probability of every single thing that can happen, and every thing that can happen is any number between $-\infty$ and $+\infty$. Since the actual value of x will take place somewhere in this interval, the area must equal 1. An example: what is $\Pr(x < m|m, s, \mathrm{E}_N)$. By symmetry, it is 0.5, because it is half the area under the curve, from the point m to $-\infty$ (everything to the left of the thin, dotted curve starting at m). Another example, $\Pr(x > m + 2s|m, s, \mathrm{E}_N)$, which we know is about 0.025, is that tiny area from the point $m+2s$ to $+\infty$ (everything to the right of the thin, dotted curve starting at $m + 2s$). Why do we know it is about 0.025? Because the probability of numbers in the interval $(m-2s, m+2s)$ is 0.95, which leaves 0.05 probability of numbers outside that interval. Then, since the normal is symmetric, this leaves 0.025 probability for numbers less than $m - 2s$ and 0.025 for numbers greater than $m + 2s$.

One more picture showing two normal distributions. The curve to the left has a smaller spread parameter *and* central parameter than the one to the right. The area under either curve is 1 (it is always equal to 1 for any distribution). Notice that the one with the larger spread parameter is wider, meaning you are less certain about the values of x. Obviously, then, smaller

spread parameters mean you are more certain, in the sense that the most of the probability is for a narrower ranges of xs.



If you quantified the uncertainty in x using the distribution to the left, and I used the one to the right, which of us thinks there is a higher probability of large values of x? For example, pick the point indicted by the dotted vertical line. Obviously, I am more certain that I will see values this large or larger because the area under the curve to the right of the dotted line is larger for me than for you. You can see that we can answer lots of questions like this by reference to pictures. Next Chapter, we'll learn how to do this on a computer.

Nonsense alert! You sometimes hear this, "Our observation was drawn from a normal distribution" (with given parameter values). When this person says "drawn" they do not mean they drew a picture like we just did. They instead intend that nature (the modern equivalent of a deity) "randomly generated" the observation using a normal. Somehow, through randomness, the value appeared, almost as if by magic. Well, dear reader, something caused that observation to take the value it did. If you knew the exact casual mechanism, including the initial conditions, the starting point that gives rise to a particular value, then you would have known in advance exactly what x would be. However, just because *you* do not know, does not mean the cause did not exist. If you want clarification of this, see the gorgeous Chapter 10 of Jaynes (2003) wherein he discusses the physics and probabilities of coin flips.

4. Homework

(1) The CMU football score was called a "random variable". Write down five other random variables.

(2) I teach a class called Statistics 101. There are 36 students signed up for this class. Before I come to class each day, I guess how many students will actually show up (I know you will be shocked, but some people actually miss class!). Obviously, I do not know *for certain*, the exact number. How do I express my uncertainty in this number? What is everything that can happen in this case? Which probability distribution would best represent my uncertainty?

(3) Why, in a probability distribution, does the sum of the probabilities of every single thing that can happen add up to 1?
(4) A famous statistical problem came from a lady who claimed she could tell by its taste whether her tea was poured into the cup before the milk, or that the milk was poured first and the tea afterwards. This is, of course, an extraordinary claim, and it was decided to test her in an experiment. Five servings of tea were given to her, some with the tea poured first, some with the milk first (she obviously could not see which). She guessed correctly (which was poured first) four times. So, what do you think? Does she have the ability she claims? Why or why not?
(5) Think of a situation where the outcome is "random" but where your uncertainty is approximately quantified with a normal distribution. Justify your answer!
(6) I earlier said that if I had a bucket filled with continuous numbers, such as are represented by normal distributions, I could not even pick one out. That is, not only is the probability of seeing any number 0, no continuous number could ever be picked. Why?
(7) Draw a picture with two normal distributions (both on the same picture) where they both have the same central parameter, but different spread parameters. Call the one with the larger spread parameter B (the smaller is A). Which distribution, A or B, has a higher probability of having very large values? Meaning, if you were to come across two observations, where the uncertainty in the first is quantified by distribution A and the second by distribution B, which is of these observations is more likely to be larger? How about very small values?
(8) Now draw two normals, but this time they both have the same spread parameter but different central parameters (label the one with the larger central parameter A). Which one now has a higher probability of very large values and very small values.
(9) It is often said that the distribution of students' grades in statistics classes is a "bell curve", that is, that they are normally distributed. Is that strictly true? Why or why not?
(10) Calculate the expected number of a die roll (dice have the numbers 1 through 6 on them, as you know, with equal probability of seeing each side). Comment on this number.
(11) The exquisite game of Petanque starts by throwing a tiny ball, a *cochonet*, a distance of six to ten meters. Players, in teams, then take turns throwing manly steel balls[3], or *boules*, which are about the size of a apples, and try to get them as close to the *cochonet* as possible. Points are awarded by the number of *boules* of your team that are closer to the *cochonet* than the nearest opponent's *boule*. Throws are rarely perfect, of course; meaning, that the *boules* aren't always right next to the *cochonet*; they lie some distance away. Suppose, then, that I step up and throw my first *boule*. Do you know how far away my *boule* will be from the *cochonet*? How would you quantify your uncertainty in this distance? Which probability distribution best describes your uncertainty?

[3]This is contrasted with the similar Italian game Bocce, which uses meek wooden balls.

(12) The picture of the normal distribution was said to be abbreviated because we cannot picture the whole thing. Why?
(13) EXTRA See if you can find, in the popular media, somebody using the normal distribution (or bell curve). Fully cite the source. Do *not* cite a source that has anything to do with the results of a scientific study (these are obviously far too common). Try looking in business areas (advertising, marketing, and so on).
(14) EXTRA In the CMU football example, with $p = 1/2$, we said the probability of winning games was symmetric about 6 games, and that the probability of winning no games was the probability of winning all 12. Try and prove this mathematically using the binomial formula. Then show that the probability of winning 1 game is the same as winning 11.

CHAPTER 5

R

1. R

R (R Development Core Team, 2008) is a fantastic, hugely supported, rapidly growing, infinitely extensible, operating-system agnostic, free and open source statistical software platform. Nearly everybody who is anybody uses R, and since I want you to be somebody, you will use it, too. Some things in R are incredibly easy to do; other tasks are bizarrely difficult. Most of what makes R hard for the beginner is the same stuff that makes any piece of software hard; that is, getting used to expressing your statistical desires in computerese. As such an environment can be strange and perplexing at first, some students experience a kind of peculiar stress that is best described by example. Here is a video from a Germany showing a young statistics student who experienced trouble understanding R:

`http://youtube.com/watch?v=PbcctWbC8Q0`

Be sure that this doesn't happen to you. Remember what Douglas Adams said: *Don't panic.*

The best way to start is by going to `r-project.org` and click the `CRAN` under the `Download` heading. You can't miss it. After that, you have to choose a *mirror*, which means one of the hundreds of computers around the world that host the software. Obviously, pick a site near you. Once that's done, and choose your platform (your operating system, like Linux or one of the others), and then choose the `base` package. Step-by-step instructions are at this book's website: `wmbriggs.com/book`. It is no more difficult to install than any other piece of software.

This is not the place to go over all the possibilities of R; just the briefest introduction will be given, because there are far better places available online (see the book website for links). But there are a few essential commands that you should not do without in Table 1.

The Appendix gives a fuller list of R commands.

It is important to understand that R is a *command-line* language, which we may interpret as meaning that all commands in R are *functions* which must be typed into the console. These are objects that are a command name plus a left and right parenthesis, with variables (called arguments) stuck in between, thus: `plot(x,y)`. Remember that you are dealing with computers, which are literal, intolerant creatures, and so cannot abide even the slightest deviation from its expectations. That means, if instead of `plot(x,y)`, you

TABLE 1. Useful R commands.

Command	Description
`help(command)`	Does the obvious: always scroll down to the bottom of the help to see examples of the command.
`?command`	Same as `help()`
`apropos('string')`	If you cannot remember the name of a command—and I always forget—but remember it started with `co`–something, then just type `apropos('co')` and you'll get a complete list of commands that have `co` anywhere in their names.
`c()`	This is the concatenation function: typing `c(1,2)` concatenates a 2 to 1, or sticks on the end 1 the number 2, so that we have a *vector* of numbers.

type `lot(x,y)`, or `plot x,y)`, or `plot(,y)`, or `plot(x,y` things will go awry. `R` will try to give you an idea of what went wrong by giving you an error message. Except in cases like that last typo, which will cause you to develop stress lines, because all you'll see is this

```
+
```

and every attempt of yours to type anything new, or hit `enter` 100 times, will not do a thing except give you more lines of `+` or other screwy errors. Because why? Because you typed `plot(x,y`; that is, you typed a left parenthesis (right before the `x`) and you never "closed" it with a right parenthesis, and `R` will simply wait forever for you to type one in.

The solution is to enter a right parenthesis, or hit

```
ctrl+c
```

which means the control key plus the `c` key simultaneously, which "breaks" the current computation.

Using `R` means that you have to memorize (!) and type in commands instead of using a graphical user interface (GUI), which is the standard point-and-click screen with which you are probably familiar. It is my experience that students who are not used to computers start freaking out at this point; however, there is no need to. I have made everything very, very easy and all you have to do is copy what you see in the book to the `R` screen. All will be well. I promise.

GUIs are very nice things, incidentally, and `R` has one that you can download and play with. It is called the `R Commander`. Like all GUIs, some very basic functionality is included that allows you to, well, point and click and get a result. Problem is, the very second you want to do something

TABLE 2. R commands for binomial distributions.

`dbinom`	The probability of *d*ensity function: given the `size`, or n, and `prop`, or p, this calculates the probability that we see `x` successes; this is equation (14).
`pbinom`	The distribution function, which calculates the probability that the number of successes is less than or equal to some `a`.
`qbinom`	This is the "quantile" function, which calculates, given a probability from the distribution function, which value of `q` it is associated with. This will be made clear with some examples with the normal distribution later.
`rbinom`	This generates a "random" binomial number; and since random means unknown, this means it generates a number that is unknown in some sense; we'll talk about this later.

different than what is available from the GUI, you are stuck. With statistics, we often want to do something differently, so we will stick with the command line.

The best reason for typing in—and saving—your commands, and not point and clicking, is that *you can save them!* Just today, a client emailed me and said row 122 of the data was a mistake and should be removed. Now, if I had to go through her analysis again, mousing up to the analysis selection, pointing and clicking this and that, I'd be in a world of pain, because what I did for her was complex. But with the commands I used all neatly typed and stored, all I had to do was read in the data again, cut and paste my commands, and I was finished in minutes. Every time you get up from the computer, when using a GUI, you have to start from scratch. Typing—and saving—your commands saves you more time than you will know what to do with.

2. R binomially

By now, you are eagerly asking yourself: "Can R help up with those binomial calculations like in the Thanksgiving example?" Let's type `apropos('bino')` and see, because, after all, 'bino' is something like binomial. The most likely function is called `binomial`, so let's type `?binomial` and see how it works. Uh oh. Weird words about "family objects" and the function `glm()`, and that doesn't sound right. What about one of the functions like `dbinom()`? Jackpot. We'll look at these in detail, since it turns out that this structure of four functions is the same for every distribution. The functions are in Table 2.

Let's go back to the Thanksgiving example, which used a binomial. Moe can calculate, given $n = \texttt{size} = 3, p = \texttt{prob} = 0.1$, his probabilities using `R`:

```
dbinom(0,3,.1)
```

which gives the probability of taking nobody along for the ride. The answer is `[1] 0.729`. The "`[1]` in front of the number just means that you are only looking at line number 1 of the output. If you asked for dozens of probabilities, for example, `R` would space them out over several lines. Let's now calculate the probability of taking just 0, just 1, etc.

```
dbinom(c(0,1,2,3),3,.1)
```

where we have "nested" two functions into one: the first is the concatenation function `c()`, where we have stuck the numbers 0 through 3 together, and which shows you the `dbinom()` function can calculate more than one probability at a time. What pops out is

```
[1] 0.729 0.243 0.027 0.001;
```

that is, the exact values we got in Chapter 3 for taking 0 or 1 or 2 etc. along for the ride. Now we can look at the distribution function:

```
pbinom(c(0,1,2,3),3,.1);
```

and we get

```
[1] 0.729 0.972 0.999 1.000.
```

This is the probability of taking 0 or less, 1 or less, 2 or less, and 3 or less. The last probability very obviously has to be 1, and will always be 1 for any binomial (as long as the last value in the function `c(0,1,...,n)` equals n).

There turns out to be a shortcut to typing the concatenation function for simple numbers, and here it is:

```
c(0,1,2,...,n) = 0:n.
```

So we can rewrite the first function as `dbinom(0:3,3,.1)` and get the same results.

We can nest functions again and make pretty pictures

```
plot(dbinom(0:3,3,.1))
```

And that's it for any binomial function. Isn't that simple? (The answer is *yes*.) The commands never change for any binomial you want to do.

3. `R` normally

Can `R` do normal distributions as well? *Can* it! Let's type in `apropos('normal')` and see what we get. A lot of gibberish, that's what. Where's the normal distribution? Well, it turns out that computer programmers are a lazy bunch, and they often do not use all the letters of a word to name a function (too much typing). Let's try apropos('norm') instead (which no matter what should give us at least as many results, right? This is a question of logic, not computers.). Bingo. Among all the rest, we see `dnorm` and `pnorm` etc.,

just like with the binomial. Now type `?dnorm` so we can learn about our fun function. Same layout as the binomial; only difference being we need to supply a "`mean`" and "`sd`" (the m and s). Sigh. This is an example of `R` being naughty and misusing the terminology that I earlier forbade: m and s are *not* a mean and standard deviation. They are parameters. It's a trap too many fall into. We'll work with it, but just remember "`mean`" and "`sd`" actually imply our *parameters* m and s.

You will recall from our discussion of normals that we cannot compute a probability of seeing a single number (and if you don't remember, shame on you: go back and read Chapter 4). The function `dnorm` does *not* give you this number, because that probability is always 0; instead, it gives you a "density", which means little to us. But we *can* calculate the probability of values being in some interval using the `pnorm` function. For example, to calculate $\Pr(x < 10|m = 10, s = 20, \mathrm{E}_N)$, use

```
pnorm(10,10,20)
```

and you should see `[1] 0.5`. But you already knew that would be the answer before you typed it in, right? (*Right?*) Let's try a trickier one: $\Pr(x < 0|m = 10, s = 20, \mathrm{E}_N)$; type `pnorm(0,10,20)` and get `[1] 0.3085375`. So what is this probability: $\Pr(x > 0|m = 10, s = 20, \mathrm{E}_N)$ (x *greater* than 0)? Think about it. x can either be less than or greater than 0; the probability it is so is 1. So $\Pr(x < 0|m = 10, s = 20, \mathrm{E}_N) + \Pr(x > 0|m = 10, s = 20, \mathrm{E}_N) = 1$. Thus, $\Pr(x < 0|m = 10, s = 20, \mathrm{E}_N) \Pr(x > 0|m = 10, s = 20, \mathrm{E}_N) = 1 - \Pr(x < 0|m = 10, s = 20, \mathrm{E}_N)$. We can get that in `R` by typing

```
1-pnorm(0,10,20)
```

and you should get `[1] 0.6914625`, which is $1 - 0.3085375$ as expected.

By the way, if you are starting to feel the onset of a freak out, and wonder "Why, O why, can't he give us a point and click way to do this!" Because, dear reader, a point and click way to do this does not exist. Stop worrying so much. You'll get it.

What is $\Pr(15 < x < 18|m = 15, s = 5, \mathrm{E}_N)$ (which might be reasonable numbers for the temperature example)? Any interval splits the data into three parts: the part less than the lower bound (15), the part of the interval itself (15-18), and the part larger than the upper bound (18). We already know how to get $\Pr(x < 15|m = 15, s = 5, \mathrm{E}_N)$, which is `pnorm(15,15,5)`, and which equals 0.5. We also know how to get $\Pr(x > 18|m = 15, s = 5, \mathrm{E}_N)$, which is `1-pnorm(18,15,5)`, and which equals 0.2742531. This means that $\Pr(x < 15 \text{ or } x > 18|m = 15, s = 5, \mathrm{E}_N)$, using probability rule number 1, is $0.5+0.2742531 = 0.7742531$. Finally, 0.7742531 is the probability of *not* being in the interval, so the probability of being *in* the interval must be one minus this, or $1 - 0.7742531 = 0.2257469$. A lot of work. We could have jumped right to it by typing

```
pnorm(18,15,5)-pnorm(15,15,5).
```

This is the way you write the code to compute the probability of any interval—remembering to input your own m and s of course!

4. Advanced

. You don't need to do this section, because it is somewhat more complicated. Not much, really, but enough that you have to think more about the computer than you do the probability.

Our goal is to plot the picture of a normal density. The function `dnorm(x,15,5)` will give you the value of the normal density, with an $m = 15$ and $s = 5$, for some value of `x`. To picture the normal, which is a picture of densities for a range of x, we somehow have to specify this range. Unfortunately, there is no way to know in advance which range you want to plot, so getting the exact picture you want takes some work. Here is one way:

```
x = seq(-4,4,.01)
```

which gives us a `seq`uence of numbers from -4 to 4 in increments of 0.01. Thus, $x = -4.00, -3.99, -3.99, \ldots, 4$. Calculating the density of each of these values of x is easy:

```
dnorm(x)
```

where you will have noticed that I did not type a m or s. Type `?dnorm` again. It reads `dnorm(x, mean=0, sd=1, log = FALSE)`. Ignoring the `log = FALSE` bit, we can see that `R` supplies helpfully *default* values of the parameters. They are default, because if you are happy with the values chosen, you do not have to type in your own. In this case, $m = 0$ and $s = 1$, which is called a *standard normal*. Anyway, to get the plot is now easy:

```
plot(x,dnorm(x),type='l')
```

This means, for every value of `x`, plot the value of `dnorm` at that value. I also changed the plot type to a line with `type='l'` , and which makes the graph prettier. Try doing the plot without this argument and see what you get.

5. Homework

(1) Google the `R reference card` and download it. This wonderful sheet, by Tom Short, of frequently used commands is invaluable. You won't use 98% of these, and the sheet is bound to look a little scary at first, but you will love it. All commands are organized into topics, making them easy to find.
(2) Type `demo(graphics)` then `demo(persp)` in the command window, and follow the instructions. This is to show you the wide range of pretty pictures you can get with `R`.
(3) Open a text file and call it `myRcode.R`. Save it anywhere you like, but save it as a *text* file. **Do not** save it as a–God help us–Microsoft Word file. In that file, you will type all the commands that you want `R` to run. When you do want `R` to run a particular line, cut and paste that line from the file `myRcode.R` to `R`. Why do this? So you can have a record of all your

commands, and so you don't have to retype them over and over again. Microsoft Word files insert many, many hidden characters that will screw up `R`. Why can't you see them? They are hidden. Don't ask questions like this.

(4) Calculate the probability that the CMU football teams wins $x = 0, x = 1, \ldots, x = 12$ games, using the information from the previous Chapter. Round your answers! Please do not try to write out 43 digits for each probability. Cut and paste the results and print them if you like.
(5) `dbinom(0,10,.5)` calculates the probability of getting 0 heads in 10 flips of a coin (right?). Write down the `R` code to calculate the probability of *getting at least one head*. HINT: getting at least one head means *not* getting zero heads.
(6) What is the probability of seeing an $x = 7$ where your uncertainty in x is represented by a normal distribution with parameters 7 and $\sqrt{10}$?
(7) Suppose before you decide to take anybody for the Thanksgiving trip, you learn that Moe and Curly's brother Shemp showed up, so now there are 4 people who need a ride. Describe the situation in terms of probability and calculate every single thing that can happen.
(8) What is $\Pr(x < -10 \text{ or } x > 5 | m = -5, s = 3, \mathrm{E}_N)$
(9) What is `pnorm(0)`? Try to do this in your head first. Either way, explain why the number is what it is.

CHAPTER 6

Normalities & Oddities

1. Standard Normal

Suppose $x|m, s, \mathrm{E}_N \sim \mathrm{N}(m, s)$, then there turns out to be a trick that can make x easier to work with, especially if you have to do any calculations by hand (which, nowadays, will be rarely). Let

$$z = \frac{x - m}{s}$$

Then $z|m, s, \mathrm{E}_N \sim \mathrm{N}(0, 1)$. It works for any m and s. Isn't that nifty? Lots of fun facts about z can be found in any statistics textbook that weighs over 5 pounds (these tidbits are usually in the form of impenetrable tables located in the back of the books; for no reason except professorial inertia, nearly all statistics students still have to learn to read this archaic tables).

What makes this useful is that $\Pr(z > 2|0, 1, \mathrm{E}_N) \approx \Pr(z > 1.96|0, 1, \mathrm{E}_N) = 0.025$ and $\Pr(z < -2|0, 1, \mathrm{E}_N) \approx \Pr(z < -1.96\ |0, 1, \mathrm{E}_N) = 0.025$: or, in words, the probability that z is bigger than 2 or less than negative 2 is about 0.05, which is a magic (I mean real voodoo) value in classical statistics. We already learned how to do this in `R`, Chapter 5.

In Chapter 4, a homework question explained the rules of petanque, which is a game more people should play. Suppose the distance the boule lands from the cochonette is x centimeters. We do not know what x will be in advance, and so we (approximately) quantify our uncertainty in it using a normal distribution with parameters $m = 0$ cm and $s = 10$ cm. If $x > 0$ cm it means the boule lands beyond the cochonette, and if $x < 0$ cm is means the boule lands in front of the cochonette. You are out on the field playing, far from any computer, and the urge comes upon you to discover the probability that $x > 30$ cm. First thing to do is to calculate z which equals $(30\text{cm} - 0\text{cm})/10\text{cm} = 3$ (the cm cancel). What is $\Pr(z > 3|0, 1, \mathrm{E}_N)$? No idea; well, some idea. It must be less than 0.025, since we have all memorized that $\Pr(z > 2|0, 1, \mathrm{E}_N) \approx 0.025$. The larger z is, the more improbable it becomes (right?). Let's say as a guess 1%. When you get home, you can open `R` and plug in `1-pnorm(3)` and see that the actually probability is 0.1%, so we were off by an order of magnitude (a power of 10), which is a lot, and which proves once again that computers are better at math than we are.

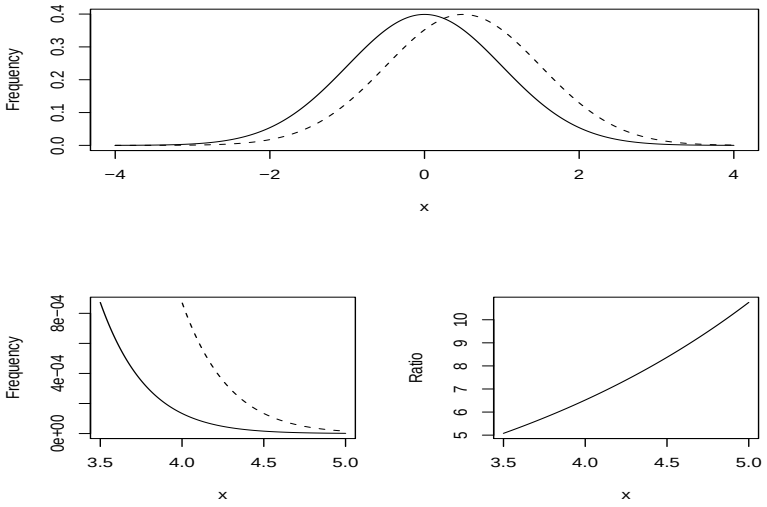


FIGURE 1. Upper panel: two normal distributions. The dotted line is one with are larger central parameter. The lower left panel expands the region from 3.5 to 5. The lower right divides the probability of the dotted line by the solid line.

2. Nonstandard Normal

The standard normal example is useful for developing your probabilistic intuition. Since normal distributions are used so often, we will spend some more time thinking about some consequences of using them. Doing this will give you a better feel for how to quantify uncertainty.

Figure 1 is a picture of two normal distributions. The one with the solid line has $m_1 = 0$ and $s_1 = 1$; the dashed line has $m_2 = 0.5$ and also $s_2 = 1$. In other words, the two distributions differ only in their central parameter, they have the same spread parameter. Obviously, large values are more likely according to distribution 2, and smaller values are more likely given distribution 1, as a simple consequence of $m_2 > m_1$. However, once we get to values of about $x = 4$ or so, it doesn't look like the distributions are that different. (Cue the spooky music.) *Or are they?*.

Under the main picture are two others. The one on the left is exactly like the main picture, except that it focuses only on the range of $x = 3.5$ to $x = 5$. If we blow it up like this, we can see that it is still more likely to see large values of x using distribution 2. How much more likely? The picture on the right divides the probabilities of seeing x or larger with distribution 2 by distribution 1, and so shows how much more likely it is to see larger values with distribution 2 than 1. For example, pick $x = 4$. It is about 7.5 times more likely to see an $x = 4$ or larger with distribution 2. That's a lot! By the

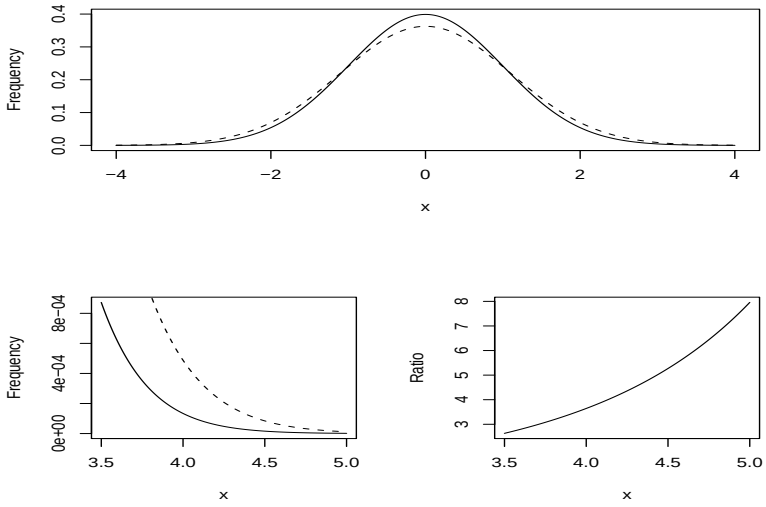


FIGURE 2. Like Figure 1, but for two normal distributions, the dashed line with a larger spread parameter.

time we get out to $x = 5$, we are 12 times more likely to see values this large with distribution 2. The point is that even very small changes in the central parameters lead to large differences in the probabilities of "extreme", values of x.

Figure 2 again shows two different distributions, this time with $m_1 = m_2 = 0$ with $s_1 = 1$ and $s_1 = 1.1$. In other words, both distributions have the same central parameters, but distribution 2 has a spread parameter that is slightly larger. The normal density plots do not look very different, do they? The dashed line, which is still distribution 2, has a peak slightly under distribution 1's, but the differences looks pretty small.

The bottom panels are the same as before. The one on the left blows up the area where $x > 3.5$ and $x < 5$. A big difference still exists. And the ratio of probabilities is still very large. It's not shown, but the plot of the right would be duplicated (or mirrored, actually) if we looked at $x > -5$ and $x < -3.5$. It is more probable to see extreme events in either direction (positive or negative) using distribution 2.

The surprising consequence is that very small changes in either the central parameter or the spread parameter can lead to very large differences at the extremes. Examples of these phenomena are easily found in real life, but my heightened political sensitivity precludes me from publicly pointing any of these out.

3. Intuition

We have learned probability and some formal distributions, but we have not yet moved to statistics. Before we do so, let us try to develop some intuition about the kinds of problems and solutions we will see before getting to technicalities. There are a number of concepts that will be important, but I don't want to give them a name, because there is no need to memorize jargon, while it is incredibly important that you develop a solid understanding of uncertainty.

The well-known Uncle Ted's[1] chain of Kill 'em and Grill 'em Vension Burger restaurants sell both Coke and Pepsi, and their internal audit shows they sell about an equal amount of each. The busy Times Square branch of the chain has about 5000 customers a day, while the store in tiny Gaylord, Michigan sees only about 100 customers. Which location is more likely to sell, on any given day, at least 2 times more Pepsi than Coke?

A useful technique for solving questions like this is *exaggeration*. For instance, the question is asking about a difference in location. What differs between those places? Only one thing, the number of customers. One site gets about 5000 people a day, the other only 100. Let's exaggerate that difference and solve a simpler problem. For example, suppose Times Square still gets 5000 a day, but Gaylord only gets 1 a day. The information is that selling a Coke is roughly equal to the probability of selling a Pepsi. This means that, at Gaylord, to that 1 customer on that day, they will either sell 1 Coke or 1 Pepsi. If they sell a Pepsi, Gaylord has certainly sold more than 2 times as much Pepsi as Coke. The chance of that happening is 50%. What is two times as much Pepsi as Coke at Times Square? A lot more Pepsi, certainly. So it's far more likely for Gaylord to sell a greater proportion of Pepsi *because* they see fewer customers. The lesson is that when the "sample size" is small, we are more likely to see extreme events.

Here is another common type of situation. What is the length of the first Chinese Emperor Qin Shi Huangdi's nose? You don't know? Well, you can make a guess. How likely is it that your guess is correct? Not very likely. Suppose that you decide to ask everybody you know to also guess, and then average all the answers together in an attempt to get a better guess. How likely is it that this averaged-guess is perfectly correct? No more likely. If you haven't a clue about the nose, and nobody else does either, than averaging ignorance is no better than single ignorance. The lesson is that just because a large group of people agree on an opinion, it is not necessarily more probable that that opinion, or average of opinions, is correct. Uninformed opinion of a large group of people is not necessarily more likely to be correct than the opinion of the guy on the corner. Think about this the next time you hear the results of a poll or survey.

You already posses other probabilistic intuition. For example, suppose, given some evidence E, the probability of A is 0.0000001 (A is something

[1]Uncle Ted Nugent, that is.

that might be given many opportunities to happen, e.g. winning the lottery). How often will A happen? Right, not very often. But if you give A a lot of chances to occur, will A eventually happen? It's very likely to. See the homework for a continuation of this.

Combining probability questions is also common. Every player in petanque gets to throw three boules. What are the chances that I get all three within 5 cm? This is a compound problem, so let's break it apart. How do we find out how likely it is to be within 5 cm of the cochonette? Well, that means the boule can be 5 cm in front of the cochonette, right near it, or up to 5 cm beyond it. The chance of this happening is $\Pr(-5cm < x < 5cm | m = 0cm, s = 10cm, \mathrm{E}_N)$. We learned how to calculate the probability of being in an interval last chapter:

```
pnorm(5,0,10)-pnorm(-5,0,10).
```

This equals about 0.38, which is the chance that one boule lands within, or +/- 5 cm, from the cochonette. What is the chance that all of them land that close? Well, that means the first one does *and* the second one *and* the third. What probability rule do we use now? The second, which tells us to multiple the probabilities together, which is $0.38^3 \approx 0.14$. The important thing to recall, when confronted with problems of this sort: do not panic. Try to break apart the complex problem into bite-size pieces.

4. Regression to the mean

There is a thing called the *Sports Illustrated* curse. It is supposed to befall an athlete immediately after he appears on the cover of the magazine. One minute the jock is riding high, adulation appears from all corners, he gets his mug on the front page of *SI*, and then—*poof!*—next week he's down in the dumps. Now, before I tell you why there is no such thing as this curse, can you think of how probability can explain it?

No athlete performs the same week to week. This is obvious. The chance that he performs very poorly, or less than average, to middle of the road, to above average, to exceptional well can all be quantified by a probability distribution. No athlete will always perform at his worst nor at his best. But when he does perform at his best, and he does so for, say, a week or two in a row, he will come to the notice of the writers at *SI*. Those writers have to put *somebody* on the cover, and so they choose our man.

Since performing at his absolute best, and for a week or two in a row, is improbable, our man will very likely perform merely above average or worse in the weeks after he gets his picture on the cover. The curse will have struck again!

Regression to the mean is also the name for the empirical observation that, for example, very tall men tend to have shorter children. Key word is "tend". I am tall and had one average-height son and one that was taller than me. If genetics dictated that men's children would be identical to him (in some characteristic, or share some gene), there would be no regression

to the mean. Since genetics tells us that it is only probable but not certain that a man's child would be identical to him (in the sense of sharing a gene), regression to the mean happens.

5. Forecasting

A central tenet of this book is that all of statistics is about using old data and evidence to make probability statements about *data not yet seen.* Example: an efficacy drug trial. After the trial is over we want to predict how *different* patients will fare using the drug. Usually, but not always, these data are observables that will be measured in the future, so, if you like, we can call this process *forecasting.* That term is typically reserved for the statistics of observables that occur at regular intervals through time, an area also called *time series analysis.* However, the distinction is unfortunate because in making it we tend to lose sight of the real purpose of statistics: which is making predictions.

There are, of course, many technical methods to do this predicting—we'll be learning some of them in two Chapters–but oftentimes we make informal guesses about unseen data and it's good to understand how these guesses fit in with probability statements. Not all the informal predictions we make are equally good. Around the 4th of July, here in the States, there is a tendency for weather forecasts to show a probability of precipitation that is lower than it should be. By that I mean, it rains more than the forecasters guess it would.

The same thing inverted happens around December 25th (the Federally Recognized Holiday That Shall Not Be Named): the forecasts tend to give too high a probability of precipitation. It snows less than the forecasters guess it would.

This phenomena is well recognized in meteorology where it has long gone by the name of —it wishcasting. This describes the tendency of the forecaster to tilt his guess toward the outcome which he would like to see, or toward the outcome he knows his viewers would like to see.

Good weather forecasters, obviously, are aware of this tendency and do their best to lessen its influence. But even the best of them tend to get excited when a big storm is on its way, these being matters of great and evident importance, and sometimes they issue forecasts which exaggerate the chance of severe weather. Still, the influence of wishcasting is small among most professionals, mostly because of the routine evaluation of forecast performance and criticism of peers. People like to pick on weather forecasters, but among any professional group, I have not found any to be better or more reliable than the National Weather Service.

Before we go further, let me answer an objection which might have occurred to you. Why not exaggerate the probability of a storm causing damage since "its better to be safe than sorry"? To do this takes the decision out of the hands of person who will experience the storm and puts it into the

hands of the forecaster. And that is the wrong thing to do: the forecaster does not know better than his audience what decisions are best. Every person in the path of a storm knows what losses he will face if a storm hits, and how much it will cost him to protect against one. Sometimes the cost of protecting against a storm are too much, and it will be better for a person to do nothing if the chance of storm isn't too high. If people are routinely given exaggerated forecasts, then they will pay the cost of protecting more than they should. You cannot use the forecast as a tool to warn people of dangers which are unimportant to them. It will make them less likely to believe forecasters when real dangers arise. The lesson of Chicken Little is pertinent.

While the Weather Service forecasters do a great job, this is not so among all professions. Reporters and politicians, for example, routinely wildly overstate potential dangers, even for mundane events. Well, reporters and politicians shading the truth, embroidering facts, neglecting pertinent information, and at times outright lying is by now of no surprise. People have learned to "divide by 10" any statement issued from a newsroom, so journalists cause less harm than they would if they were taken at face value.

Wishcasting is by no means restricted to weather predictions. Forecasting who will win an election, for example, is fraught with emotion. It is difficult to remove the prejudices you have for one candidate or the other and give a good guess. If you love candidate X, you are likely to increase your guess of the chance of him winning. If you fear candidate Ys promised reforms, that might increase your guess of the chance of him winning if you are naturally pessimistic. To carefully sift through all the evidence and arrive at an unemotional prediction is extremely difficult.

Gamblers often wishcast. "Red hasnt come up if seven spins, so its more likely to now." Part of this reasoning is due to misunderstanding or not knowing the rules of probability that govern simple games, but part is also due to the desire for the outcome. Wishcasting is also prevalent in environmental circles. So much so, that an "activist" who doesnt embellish her predictions is a oddity. Brokers, financial planners, stock pickers, and similar professionals are no less prone to wishcasting, as any study of the economy will show.

Wishcasting is somewhat different than the experimenter effect, although there is some overlap. The experimenter effect is when a scientist (or group of them), consciously or not, manipulates an experiment to unfairly demonstrate the effect they were looking for. A common example is a drug trial. One group of patients is given a new drug, the other an old one or a placebo. If the patients are evaluated by a physician who knows which patient got which drug, it is likely the effects of the new drug will be exaggerated. This phenomena is so well known that the government mandates blinding of medical trials. This is where the physician who evaluates the patients has no idea which treatment the patient has received.

Michael Crichton, physician and author, in testimony to congress, gave an example of this:

> It's 1991, I am flying home from Germany, sitting next to a man who is almost in tears, he is so upset. He's a physician involved in an FDA study of a new drug. It's a double-blind study involving four separate teamsone plans the study, another administers the drug to patients, a third assesses the effect on patients, and a fourth analyzes results. The teams do not know each other, and are prohibited from personal contact of any sort, on peril of contaminating the results. This man had been sitting in the Frankfurt airport, innocently chatting with another man, when they discovered to their mutual horror they are on two different teams studying the same drug. They were required to report their encounter to the FDA. And my companion was now waiting to see if the FDA would declare their multi-year, multi-million dollar study invalid because of this chance contact.

His point in this testimony was to show that researchers in other fields, such as global warming, are nowhere near as careful as their colleagues in medicine:

> [T]he protocols of climate science appear considerably more relaxed. In climate science, it's permissible for raw data to be "touched," or modified, by many hands. Gaps in temperature and proxy records are filled in. Suspect values are deleted because a scientist deems them erroneous. A researcher may elect to use parts of existing records, ignoring other parts. But the fact that the data has been modified in so many ways inevitably raises the question of whether the results of a given study are wholly or partially caused by the modifications themselves...
>
> ...[A]ny study where a single team plans the research, carries it out, supervises the analysis, and writes their own final report, carries a very high risk of undetected bias. That risk, for example, would automatically preclude the validity of the results of a similarly structured study that tested the efficacy of a drug.

Wishcasting meets the experimenter effect when the results from a non-blinded experiment are exaggerated to "raise awareness" of the potential horrors that await us if we do not heed the experimenters' advice. Sometimes this exaggeration is done on purpose, as with the weather forecaster who feels his viewers would be "better safe than sorry", and sometimes the overstatement is unconscious because the forecaster has not recognized his limitations. Scientists are particularly prone to this when announcing their

results to the public. They often feel they are special and able to avoid the frailties that plague the rest of us, but of course, they cannot; they are still human.

It is nearly impossible to disentangle experimenter effect from wishcasting in any situation, nor can we easily identify the constituent facts and their relevance used by a forecaster in producing his forecast. To do so essentially means producing a rival forecast and is a laborious process. What we can do (this is my line of country) is to check how good the actual performance of a forecast is. If the forecast routinely fails, we can say something has gone wrong. Just what requires more work: was it bad data, mistaken theory, wishcasting, or something else? If the forecast routinely fails, we are rational to suspect it will fail in the future, and that the theories said to underly the forecast might be false. If the forecast fails, we are also right to question the motives of the forecaster, because it is these motives that influence the presence or amount of wishcasting.

These cautions do not just apply to weather or climate forecasts, but in all areas where routine predictions are made. Could you be making more money in your stock portfolio or office football pool, for example? Generally, wishcasting takes places when forecasting complex systems, like the weather, climate, or any area involving human behavior. Its much less likely in simple situations, like how much this electron will move under a certain applied force, or what will happen when these two chemicals are mixed. But well save complexity for another book.

6. Homework

(1) We can easily figure out how much more Pepsi sold is twice as much as Coke at Times Square. This, from your high school algebra days, is found by solving the equations: Pepsi $>$ 2$\times$Coke, and Pepsi + Coke = 5000. Thus, Coke = 5000 - Pepsi, so (substituting this into the first equation), Pepsi $>$ 2$\times$ (5000 - Pepsi). Finally, 3$\times$Pepsi $>$ 10000, or Pepsi $>$ 3334. Use this technique to solve the amount of Pepsi at Gaylord. Then use `R` (the `pbinom` equation) to formally solve the probability of selling twice as much Pepsi at each location.

(2) If the probability of A (given some evidence) happening is 0.0000001, about how many chances for A to happen have there to be so we can be 50% or more certain A will occur? An example might be making a basket by tossing the basketball, behind your back, the length of the court. HINT: think binomial.

(3) What is the probability that I get at least one boule out of three within 5 cm of the cochonette? What is the probability that my teammate also gets at least one boule as close?

(4) A cop is about to shoot a bad guy. The chance that the cop hits the bad guy is about 30%. How many times should the cop fire so that he is at least 99.9% sure of hitting the bad guy (and thus making it more certain that he himself is not shot)? HINT: think binomial; what is n?

(5) Two groups of people exist. Group A is normal. The people in group B were exposed to deadly, hulk-inducing gamma radiation. This radiation affected their systolic blood pressure, the uncertainty of which is measured with a normal distribution with central parameter of 125.4 mmHg (millimeters of mercury) and some spread parameter. The central parameter for group A is 120 mmHg, about 5% lower than in B. It has the same spread parameter as B. Patients only come into blood-pressure physician Dr. Banner's office if they have blood pressures over 160 mmHg. About what percentage of Dr. Banner's patients will be from group B: 5%, 50%, 55%, or 95%?
(6) Uncle Ted's also sells the Super Freedom 4x4 Ground Round Pounder, 4 pounds of quality chuck, topped by a slab of premium yellow cheese (the kind with the colorful things in it!), a sliver of onion, and a cup of sugar-free ketchup, all crammed between a loaf of whiter-than-white bread. The probability of any customer buying this cardiologist's delight is 0.002. Can you tell which location, Times Square or Gaylord, is more likely to sell, on any given day, a Super Freedom burger?
(7) Two government-funded researchers, Drs. C and D, set out to discover how well Americans know their celebrities, for nothing is more important than celebrities. Dr. C asked 1 random person a day whether they recognized the haircut of a pop star from a set of photographs. Dr. D asked 3 people a day the same thing. Both found that Americans knew about 1/2 the haircuts shown to them. Dr. C counted each research day a success if the person he interviewed correctly identified the celebrity. Dr. D counted the day a success if all his interviewees correctly identified the celebrities. Can you say which researcher had more successful days?
(8) Dr W, a prominent psychiatrist, sees patients only on Wednesdays and Thursdays (the other days he sets aside to count his money). He sees about 30 on Wednesdays and about 5 on Thursdays. The chance that a patient has the Heebie-Jeebies is 10%. On which day will Dr. W more likely see a greater proportion of patients for that day with this dread disease?
(9) Dr. W's main interest is in the Screaming Willies. He developed a drug which he gave to his patients to improve their scores on a mental exam. He quantified the uncertainty in the improvements with a normal distribution. He also discovered that patients given a placebo instead of a drug also tended to improve their scores. He quantified his uncertainty in placebo-patient improvements with a normal distribution, too. Both normal distributions, for the drug- and the placebo-patients, had the same central parameter, but the placebo group had a spread parameter 3 times that of the drug group's. Dr. W wanted to hold a reunion party only for those patients who improved—regardless of their group; he couldn't bear to face those who got worse. He could only afford to have a party with 30 people, because the disco hall he had rented only had a small number of rental roller skates. So he sent out invitations to the top 30 patients that got better. About how many of the people that Dr. W invites will be from the *drug* (not the placebo) group: 1, 15, 20, or 30?

(10) What will be the approximate ratio of drug group to placebo group patients, just for those patients who did not improve nor got worse? (This is the number of drug group patients who did not improve nor got worse divided by the placebo group patients who did not improve nor got worse).
(11) Dr. W could not face his failures, but a lawyer hired by those whose scores got worse could (the lawyer was actually hired by other family members, because those who got worse were no longer smart enough to think of hiring a lawyer). The lawyer obviously wants to score big—I mean, wants justice to prevail—and so selects as his customers—*clients*—those patients who really got worse. About what proportion of the lawyer's clients will be from the drug group?
(12) A new casino has opened in the Greek Quarter of Detroit, called the Zaplutus. Gamblers suspect the high-stakes roulette table is rigged. This roulette table only has black and red slots, an even number of each, and to play it gamblers must bet at least $10,000 a roll. Plus, no spectators are allowed at the table: you must bet to watch the game. Two gamblers decide to bet. Gambler A bet black 5 times. The wheel came up red 4 times and black 1 time. Gambler B, who is richer, bet 20 times, all on black too. His wheel came up 13 times red and 7 times black. Can you tell which gambler should be more suspicious that the table is rigged toward coming up red more often?
(13) Think about an upcoming or recent election. Who do you want to win and who do you think will win? If it's possible (the right season and circumstances), carry out a survey first asking people who they want to win and then who they think will win. You will find the results surprising, particularly if the election is a major or contentious one.

CHAPTER 7

Reality

1. Kinds of data

Somewhere, sometime, somehow, somebody is going to ask you to create some kind of data set (that time is sooner than you think; see the homework). Here is an example of such a set, written as you might see it in a spreadsheet (a good, free open-source spreadsheet is `Open Office`, `www.openoffice.org`):

Q1	...	Sex	Income	Nodules	Ridiculous
rust	...	M	10	7	Y
taupe	...	F		3	N
⋮	⋮	⋮	⋮	⋮	⋮
ochre	...	F	12	2	Y

This data is part of a survey asking people their favorite colors (Q1), while recording their sex, annual income, the number of sub-occipital nodules on their brain, and whether or not the interviewee thought the subject ridiculous or not. There is a lot we can learn from this simple fragment.

The first is *always* use full, readable, English names for the variables. What about Q1, which was indeed the first question on the survey. Why not just call it "Q1"? "Q1" is a lot easier to type than "favorite color". Believe me, two weeks after you store this data, you will *not*, no matter how much you swear you will, remember that Q1 was favorite color. Neither will anybody else. And nobody will be able to guess that Q1 means favorite color.

Can you suggest a better name? How about "favcol", which has fewer letters than "favorite color", and therefore easier to type? What are you, lazy? You can't type a few extra letters to save yourself a lot of grief later on?

How about just "favorite color." Well, not so good either, because why? Because of that space between "favorite" and "color"; most software cannot handle spaces in names. Alternatives are to put underscore or period between words "favorite_color", or "favorite.color". Some people like to cram the words together camel style, like "favoriteColor" (the occasional bump of capital letters is supposed to look like a camel: I didn't name it). Whichever

style you choose, *be consistent!* In any case, nobody will have any trouble understanding that "favoriteColor" means "favorite color".

Notice, too, that the colors entered under "Q1" use the full English name for the color. Spaces are OK in the actual data, just not in variable names: for example, "burnt orange" is fine. Do *not* do what many sad people do and use a *code* for the colors. For example, 1=taupe, 2=envy green, 3=fuschia, etc. What are you trying to do with a code anyway? Hide your work from Nazi spies? Never use codes.

That goes for variables like "Sex", too. I cannot tell you how many times I have opened up a data set where I have seen Sex coded as "1" and "2", or "0" and "1". How can anybody remember which number was which sex? They cannot. And there is no reason to. With data like this, abbreviation is harmless. Nobody, except for the politically correct, will confuse the fact that "M" means male and "F" female. If you are worried about it, then type out the whole thing.

Similarly for "Ridiculous", where I have used the abbreviation "Y" for yes and "N" for no. Sometimes a "0" and "1" for "N" and "Y" are acceptable. For example, in the data set we'll use in a moment, "Vomiting" is coded that way. And, after all, 0/1 is the binary no/yes of computer language, so this is OK. But if there is the least chance of ambiguity for a data value, *type the whole answer out.* Do not be lazy, you will be saving yourself time later.

It should be obvious, but store numbers as numbers. Height, weight, income, age, etc., etc. Do not use any symbols with the numbers. Store a weight as "213" and not "213 lbs". If you are worried you will forget that weight is in pounds, name the variable `Weight.LBS` or something similar. *Never* put percentage signs (%s)next to percentages. *Never* use dollar signs next to money. Leave numeric data as numbers!

What if one of your interviewees refused to answer a question? This will often happen for questions like "Income". How should you code that? *Leave his answer blank!* For God's sake, whatever you do, do *not* think you are being clever and put in some mystery code that, to you, means "missing." I have seen countless times where somebody thought that putting in a "99" or a "999" for a missing income was a good idea. The computer does *not* know that 999 means "missing"; it thinks it is just what it looks like—the number nine-hundred and ninety-nine. So when you compute an average income, that 999 becomes part of the average. Also don't use a period, the full stop. That's a holdover from an ancient piece of software (that some people are still forced to use). Incidentally, we'll talk about people lying on surveys in Chapter 14.

There are times when an answer is purposely missing, and a blank should not be used. For example, if "Income" is less than 20000, then the interviewee gets an extra question that people who make more than 20000 do not get. Usually, this kind of rule can be handled trivially in the analysis, but if you want to show that somebody should not have answered and not that they did not answer, then use a code such as "PM" for "purposely missing".

Even better would be to write "purposely missing", so that somebody who is looking at your data three months down the road doesn't have to expend a great deal of energy on interpreting what "purposely missing" means.

Try to use a real database to store your data, and keep away from spreadsheets if you can. A real database can be coded so that all possible responses for a variable like "Race" are pre-coded, eliminating the chance of typos, which are certain to occur in spreadsheets. You will probably need help building a real database, but you will not be sorry if can find it.

Here's something you don't often get from those *other* textbooks, but which is a great truth. You will spend from 70 to 80% of your time, in *any* statistical analysis just getting the data into the form readable for you and your software. This may sound like the kind of thing you often hear from teachers, while you think to yourself, "Ho, ho, ho. He has to tell us things like that just to give us something to worry about. But it's a ridiculous exaggeration. I'll either (a) spend 10-15% of my time, or (b) have somebody do it for me." I am here to tell you that the answers to these are (a) there is no known way in the universe for this to be true, and (b) Ha ha ha!

2. Databases

The absolute best thing to do is to store you data in a database. I often use the free and open source *MySQL* (.com, of course). Knowing how to design, set up, and use such a database is beyond what most people want to do on their own. So most, at least for simple studies, opt for spreadsheets. These can be fine, though they are prone to error, usually typos. For instance, the codings "Y" and "Y " might look the same to you, but they are different inside a computer: one has a space, one doesn't. The computer thinks these are as different as "Q" and "W". This kind of typo is extraordinarily common because you cannot see blank spaces easily on a computer screen. To see if you have suffered from it, after you get your data into `R` type `levels(my_variable_name)` and each of the levels, like "Y" and "Y " will be displayed. If you see something like this, you'll have to go back to your spreadsheet and locate the offending entries and correct them.

A lot of overhead is built into spreadsheets. Most of it has to do with prettifying the rows and columns—bold headings, colored backgrounds, and so on. Absolutely none of this does anything for the statistical analysis, so we have to simplify the spreadsheet a bit.

The most common way to do this is to save the spreadsheet as a CSV file. CSV stands for Comma Separated Values. It means exactly what it says. The values from the spreadsheet are saved to an ordinary text file, and each column is separated by a comma. An example from one row from the dataset we'll be using is

```
0,0,0,0,39,"black","male","Y",17.1,80,102.4,0
```

Note the clever insertion of commas between each value.

What this means is that you *cannot* actually use commas in your data. For example, you cannot store an income value as "10,000"; instead, you should use "10000". Also note that there is no dollar sign.

Now, in some countries, where the tendrils of modern society have not yet reached, people unfortunately routinely use commas in place of decimal points. Thus, "3.42" written here is "3,42" written there. You obviously cannot save the later in a CSV file because the computer will think that comma in "3,42" is one of the commas that separates the values, which it does not. The way to overcome this without having to change the data is to change the delimiter to something other than a comma; perhaps a semicolon or a pound sign; any kind of symbol which you know won't be in the regular data. For example, if you used an @ symbol, your CSV file would look like

```
0@0@0@0@39@"black"@"male"@"Y"@17.1@80@102.4@0
```

The only trick will be figuring out how to do this. In Open Office, it's particularly easy: after opening up the spreadsheet and selecting "Save As", select the box "Edit Filter settings" and choose your own symbol instead of the default comma. A common mistake is to type an entry into, say, an `Opinion` variable, where a person's exact words are the answer and that answer contains a comma. Guard against using a comma in these words else the computer will think you have extra variables: the computer thinks there is a variable between each comma.

3. Summaries

It's finally time to play with real data. This is, in my experience, another panic point. But it need not be. Just take your time and follow each step. It is quite easy.

The first trick is to download the data onto your computer. Go to the book website and download the file `appendicitis.csv` and save it somewhere on your hard disk in a place where you can remember. The place where it is is called the `path`. That is, your hard drive has a sort of hierarchy, a map where the files are stored. In you are on a Windows machine, this is usually the `C:/` drive (yes, the slash is backwards on purpose, because `R` thinks like a Linux computer, or Apple, which has the slashes the other way). Create your own directory, say, `mydata` (do not put a space in the name of the folder), and put the appendicitis file there. So the path to the file is `C:/mydata/appendicitis.csv`. Easy, right? If you are on a Linux or Mac, it's the same idea. The path on a Mac is usually something like `/Users/YOURNAME/mydata/appendicitis.csv`. On a Linux box it might be `/home/YOURNAME/mydata/appendicitis.csv`. Simple!

Open `R`. Then type this exact command:

```
x = read.csv(url("http://wmbriggs.com/book/appendicitis.csv"))
```

There is a lot going on here, so let's go through it step by step. Ignore the `x =` bit for a moment and concentrate on the part that reads `read.csv(...)`. This built-in `R` function reads a CSV file. Well, what else would you have expected from its name? Inside that function is another one called `url()`, whose argument is the same thing you type into any web browser. The thing you type is called the URL, the Uniform Resource Locater, or web address. What we are doing is telling `R` to read a CSV file directly off the web. Pretty neat!

If you had saved the file directly to your hard drive, you would have loaded it like this

```
x = read.csv("C:/mydata/appendicitis.csv")
```

where you have to substitute the correct path, but otherwise is just as easy.

The last thing to know is that when the CSV file is read in it is stored in `R`'s memory in the object I called `x`. `R` calls these objects `data frames`. Why didn't they call them `data sets`? I have no idea. How did I know to use an x, why did I choose that name to store my data? No reason at all except habit. You can call the dataset anything you want. Call it *mydata* if you want. It just doesn't matter.

Now type just `x` and hit enter. You'll see all the data scroll by. Too much to look at, so let's summarize it:

```
summary(x)
```

This is data taken on patients admitted to an emergency room with right lower quadrant pain (in the area the appendix is located) in order to find a model to better predict appendicitis (Birkhahn et al., 2006). Each of the variables was thought to have some bearing on this question. We'll talk more about this data later. Right now, we're just playing around. When we run the command we get the summary statistics for each variable in `x`. What it shows is the mean, which is just the arithmetic average of the data, the median, which is the point at which 50% of the data values are larger and 50% smaller, the `1st Qu.`, which is the first quartile and is the point at which 25% of the data values are smaller, the `3rd Qu.` which is the third quartile and is the point at which 75% of the data values are smaller (and 25% are larger, right?). Also given in the `Min.` which is the minimum value and `Max` which is the maximum. Last is `NA's`, which are the number, if any, of missing values. These kinds of statistics only show for data coded as numbers, i.e. numerical data. For data that is textual, also called categorical or factorial data, the first few levels of categories are shown with a count of the number of rows (observations) that are in that category.

You will notice that variables like `Pregnancy` are not categorical, but are numerical, which is why we see the statistics and not a category count. `Pregnancy` is a 0/1 variable and is technically categorical; however, like I said above, it is obvious that "0" means "not pregnant", so there is no ambiguity. The advantage to storing data in this way is that the numerical

mean is then the proportion of people having `Pregnancy =1` (think about this!).

Let's just look at the variable `Age` for now. It turns out we can apply the `summary` function on individual variables, and not just on data frames. Inside the computer, the variable `age` is different than `Age` (why?). So try `summary(Age)`. What happens? You get the error message `Error in summary(Age) : object "Age" not found`. But it's certainly there!

You can read lots of different datasets into `R` at the same time, which is very convenient. I work on a lot of medical datasets and every one of them has the variable `Age`. How does `R` know which `Age` belongs to which dataset? By only recognizing one dataset at a time, through the mechanism of *attaching* the dataset directly to memory, to `R`'s internal search path. To attach a dataset, type

```
attach(x)
```

Yes, this is painful to remember, but necessary to keep different datasets separate. Anyway, try `summary(Age)` again (by using the up arrow on your keyboard to recall previously typed commands) and you'll see it works.

Incidentally, `summary` is one of those functions that you can *always* try on anything in `R`. You can't break anything, so there is no harm in giving it a go.

4. Plots

The number one, unalterable rule that you must obey when beginning work with a new dataset is **always look at the data first!** Too many people forget this rule to their ultimate embarrassment.

The `summary()` function is easy and gives you information on the distribution of your data in text. But it's usually easier to see what's going on with pictures. The visual equivalents of `summary` are `boxplot`, `hist`, and `table`. Let's do a boxplot first—it's easy, `boxplot(Age)`.

The y-axis are the values of `Age`. The center line on the boxplot is the median, the outer edges of the box are the first and third quartile, and the far ends of the lines are the 5% and 95% quantiles, defined in just the same way as the other quartiles. Boxplots will often also stick dots beyond the far ends for numbers that exceed that 99% quantile and numbers that are less than the 1% quantile.

Next up is `hist(Age)`, that tries to do exactly the same thing as `boxplot`, which is to give you a visual summary of the range and likelihood of various data values.

You can't do a boxplot on data like `Race`, because that variable is categorical. Instead, do a table by `table(Race)` to get a count of each category. This is OK, but just gives the counts when frequently you want the frequencies. To get that, you have to make a table of the table (yes, this is a pain): `prop.table(table(Race))`.

`plot` is another one of those commands, like `summary`, that you can always try on anything. It never hurts and you can't break anything.

I originally included these plots in the book so you could see them, but I decided against doing this to guard against your laziness in the homework. Do these commands yourself!

5. Extra: Advanced topics

`Temperature` is one of the variables. You can try the `summary` command on it and it works just fine. Sometimes you only want the mean and don't need all the other business, so you can use the function `mean(Temperature)`. Try it and you get `[1] NA`. What gives? Do a `summary(Temperature)` and you'll see that there are 7 missing values. The function `mean` is too stupid to give you a mean in the presence of missing values. In a way, this is a good thing, because it forces you to recall that you have an incomplete dataset, and that should give you pause. Why are the values missing? It could be important. You can get around the missing values by typing `mean(Temperature, na.rm=T)`, which says take the mean, and remove (`rm`) the missing (`na`) values. The `=T` means `TRUE` (you could also type the whole word out as `TRUE`; use capitals). The mean will then be computed. `R` is wonderful, but sometimes the way it handles missing values is a pain in the ass.

A back-of-the envelope drawing that you *can* make by hand is called a stem-and-leaf plot: it does not require you to first sort your data, but you do have to discover the minimum and maximum values. In `R` it is `stem(x)`.

Histograms and boxplots are very old, were wonderful in their day, and in some cases (discrete data) are just the thing, but we can do better with numbers that more are approximated as continuous (see Chapter 4), like `Age`. For those, use a *density estimate*, which is, in a sense, an automated superior histogram. To do this in `R` type `plot(density(Age))`.

You can assign the output of any function to a new variable, created by you. So, if you want to store the table for `Race`, type `fit = table(Race)`, where I chose the name `fit` for no good reason. All the table results are now in `fit`. To see it, just type `fit`. This makes getting proportions easier because you can now `prop.table(fit)`. You could also `plot(table(Sex))` or `plot(prop.table(table(Sex)))` or any categorical variable; try `plot(fit)`.

Also try `plot(x)` or `pairs(x, panel=panel.smooth)` and see what happens.

6. Homework

(1) If you were going to collect a person's college status (i.e. freshman, sophomore, etc.), what is the best coding? HINT: there are more than just the four standard levels of status.
(2) Go through this Chapter and do every single one of the examples. Do `not` be lazy about this. It only looks easy on paper if you've never done it before.

(3) Type this into R:
`source(url("http://wmbriggs.com/book/Rcode.R"))`
That will load the file `Rcode.R` from this book's website, and run it inside R. That file contains a number of functions that we will use in later chapters. Running this file produces no output, so don't look for any. It merely loads these functions into R's memory so that they can be accessed. Now download the file (by pasting the URL `http://wmbriggs.com/book/Rcode.R` into any web browser), save it in the same place your stored the data. Then type something like
`source("C:/mydata/Rcode.R")`
(if that is where you stored your data). Then you can use any word processing software to open `Rcode.R` and view it. Do *not* make any changes to the file unless you are comfortable programming R. I will never require you to look at any of these files.

CHAPTER 8

Estimating

1. Background

Let's go back to the petanque example, where we wanted to quantify our uncertainty in the distance x the boule landed from the cochonette. We approximated this uncertainty using a normal distribution with parameters $m = 0$ cm and $s = 10$ cm. With these parameters in hand, we could easily quantify uncertainty in questions like X = "The boule will land at least 17 cm away" with the formula $\Pr(\mathrm{X}|m = 0\,\mathrm{cm}, s = 10\,\mathrm{cm}, \mathrm{E}_N) = \Pr(x > 17\,\mathrm{cm}|m = 0\,\mathrm{cm}, s = 10\,\mathrm{cm}, \mathrm{E}_N)$. `R` even gave us the number with `1-pnorm(17,0,10)` (about 4.5%). But where did the values of $m = 0$ cm and $s = 10$ cm come from?

I made them up.

It was easy to compute the probability of statements like X when we knew the probability distribution quantifying its uncertainty and the *value* of that distribution's parameters. In the petanque example, this meant *knowing* that E_N was true and also *knowing* the values of m and s. Here, *knowing* means just what it says: knowing for certain. But most of the time we do *not* know E_N is true, nor do we know the values of m and s. In this Chapter, we will assume we do in fact know E_N is true. We won't question that assumption until a few Chapters down the road. But, even given E_N *is* true, we still have to discern the values of its parameters somehow.

So how do we learn what these values are? There are some situations where are able to deduce either some or all of the parameter's values, but these situations are shockingly few in number. Nearly all the time, we are forced to guess. Now, if we do guess—and there is nothing wrong with guessing when you do not know—it should be clear that we will not be certain that the values we guessed are *the* correct ones. That is to say, we will be uncertain, and when we are uncertain what do we do? We quantify our uncertainty using probability.

At least, that is what we do nowadays. But then-a-days, people did not quantify their uncertainty in the guesses they made. They just made the guesses, said some odd things, and then stopped. We will not stop. We will quantify our uncertainty in the parameters and then go back to what is of main interest, questions like what is the probability that X is true? X is called an *observable*, in the sense that it is a proposition about an observable number x, in this case an actual, measurable distance. We do *not* care about

the parameter values per se. We need to make a guess at them, yes, otherwise we could not get the probability of X. But the fact that a parameter has a particular value is usually not of great interest.

It isn't of tremendous interest nowadays, but again, then-a-days, it was the *only* interest. Like I said, people developed a method to guess the parameter values, made the guess, then stopped. This has led people to be far too certain of themselves, because it's easy to get confused about the values of the parameters and the values of the observables. And when I tell you that then-a-days was only as far away as yesterday, you might start to be concerned.

Nearly all of classical statistics, and most of Bayesian statistics is concerned with parameters. The advantage the latter method has over the former, is that Bayesian statistics acknowledges the uncertainty in the parameters guesses and quantifies that uncertainty using probability. Classical statistics—still the dominate method in use by non-statisticians[1]—makes some bizarre statements in order to avoid directly mentioning uncertainty. Since classical statistics is ubiquitous, you will have to learn these methods so you can understand the claims people (attempt to) make.

So we start with making guesses about parameters in both the old and new ways. After we finish with that, we will return to reality and talk about observables.

2. Parameters and Observables

Here is the situation: you have never heard of petanque before and do not know a boule from a bowl from a hole in the ground. You know that you have to quantify x, which is some kind of distance. You are assuming that E_N is true, and so you know you have to specify m and s before you can make a guess about any value of x.

Before we get too far, let's set up the problem. When we *know* the values of the parameters, like we have so far, we write them in Latin letters, like m and s for the Normal, or p for the binomial. We always write *unknown* and *unobservable* parameters as Greek letters, usually μ and σ for the normal and θ for the binomial. Here is the normal distribution (density function) written with unknown parameters:

$$x|\mu,\sigma,\mathrm{E}_N \sim \mathrm{N}(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{19}$$

where μ is the central parameter, and σ^2 is the spread parameter, and where the equation is written as a function of the two unknowns, $\mathrm{N}(\mu,\sigma)$. This emphasizes that we have a different uncertainty in x for every possible value

[1] I mean those people who were not formally trained in the mathematical subjects of probability and statistics. The vast numbers of people who compute statistics have not had this training beyond, say, a class given in a Psychology department by a professor who himself was not so trained, etc.

of μ and σ (it makes no difference if we talk of σ or σ^2, one is just the square root of the other).

You may have wondered what was meant by that phrase "unobservable parameters" last paragraph (if not, you should have wondered). Here is a key fact that you must always remember: not you, not me, not anybody, can ever measure the value of a parameter (of a probability distribution). They simply cannot be seen. We cannot even see the parameters when we know their values. Parameters do not exist in nature as physical, measurable entities. If you like, you can think of them as guides for helping us understand the uncertainty of observables. We *can*, for example, observe the distance the boule lands from the cochonette. We cannot, however, observe the m even if we know its value, and we cannot observe μ either. Observables, the reason for creating the probability distributions in the first place, must always be of primary interest for this reason.

So how do we learn about the parameters if we cannot observe them? Usually, we have some past data, past values of x, that we can use to tell us something about that distribution's parameters. The information we gather about the parameters then tell us something about data we have not yet seen, which is usually future data. For example, suppose we have gathered the results of hundreds, say 200, of past throws of boules. What can we say about this past data? We can calculate the arithmetic mean of it, the median, the various quantiles and so on. We can say this many throws were greater than 20 cm, this many less. We can calculate *any* function of the observed data we want (means and medians etc. are just functions of the data), and we can make all these calculations never knowing, or even needing to know, what the parameter values are. Let me be clear: we can make just about any statement we want about the past observed data and we *never* need to know the parameter values![2] What possible good are they if all we wanted to know was about the past data?

There is only one reason to learn anything about the parameters. This is to make statements about *future* data (or to make statements about data that we have not yet seen or *cannot* see. Though that data may be old; we just haven't seen it yet; say archaeological data; all that matters is that the data is *unknown* to you (and what does "unknown" mean?). That is it. Take your time to understand this. We have, in hand, a collection of data x_{old}, and we know we can compute any function (mean etc.) we want of it, but we know we will, at some time, see *new* data x_{new} (data we have not yet seen or might not ever see), and we want to *now* say something about this x_{new}. We want to quantify our uncertainty in x_{new}, and to do that we need a probability distribution, and a probability distribution needs parameters.

The main point again: we use old data and other evidence to make statements about data we have not yet seen.

[2]This is one of the most important sentences in the entire book.

3. Classical guess

We first need to find some way to map our evidence E and the past values of x into information about the parameters. There are lots of different ways to guess at parameter values, some easy and some hard, and these all fall into two broad classifications: yes, a classical and a modern.

We have past values of x and we want to know about future, or at least other, unknown values of x. Our evidence is E, which at least means that we know the probability distribution (Normal, say) of the observables. In this book we will also assume that E also means that knowledge of each individual observation is irrelevant to knowing what each other observation with be, but we must understand that this assumption does *not* always hold; it is just the dealing with violations of this assumption is complicated. We have to find a way to guess, or estimate, these unknown and unobservable parameters given E and the old data x_{old}.

The classical way to do this is to pick an ad hoc function of the old data and label it $f(x_{\text{old}}) = \widehat{\mu}$, where that "hat" indicates that the value of μ is only a guess. Most classical estimates have the goal that the estimate is "unbiased", or $E_x(\mu - \widehat{\mu}) = E_x(\mu - f(x_{\text{old}})) = 0$, meaning that the expected distance between the actual value of μ and the guess $\widehat{\mu}$ is 0. Sounds like a nice thing to have, unbiasedness, and it surely isn't a bad idea, but it turns out to cause a lot of problems, most of which I cannot tell you about without introducing a lot of math. However, this criterion is not compelling because of that expected value business. Expected value with respect to what? Well, with respect to an infinite number of future (not yet observed) data x...which is just the data that we are trying to quantify the uncertainty of.

Anyway, in `R`, to estimate the parameters of a normal distribution classically is easy, and you already know how to do it! If `x` is our old, previously observed data, $\texttt{x}_1$`, `$\texttt{x}_2$`, ..., `$\texttt{x}_n$, then

$$\widehat{\mu} = \texttt{mean(x)} \qquad \widehat{\sigma} = \texttt{sd(x)}$$

The mean you already know to calculate. It is often written $\bar{x}$, and called "x bar". When you see a data value with a bar over it, you know it is a mean. The observed variance of old data is $\sum_{i=1}^{i=n}(x_i - \bar{x})^2$, and the observed standard deviation of old data is the square root of that. Look at the formula and notice that the standard deviation is a measure of how far, on average, the old data values are away from the observed mean. The square is taken, $(x_i - \bar{x})^2$, so that data values that were lower than the observed mean are treated the same as data values that were higher. (If you have missing data in `x`, recall Chapter 7, where we had to modify the function like this `mean(x, na.rm=T`; same for the `sd` function).

We'll never calculate the observed standard deviation by hand. But it's pretty convenient to have the observed mean stand in for our guess of μ. Unfortunately, because $\widehat{\mu}$ = mean, a lot of people have taken to calling μ

(without the hat) *the* mean, which it most assuredly is not. μ is an unobservable parameter, while the mean is just the weighted sum of a bunch of data we have already observed. This is a subject that I'll return to later.

Quick reminder quiz. Suppose that we do know the value of the parameter exactly: what will be the value of the next observable? Right! You don't know!

4. Confidence intervals

OK, it might have been hard to understand all this so far, but it's about to get weird, so be steady. The value $\widehat{\mu}$ we got before was precise; it is a known, observed number (it *is* the mean). But do we really believe, given the data and other evidence, that the *exact*, all-time, incorruptible, *immutable* value of μ is, to as many decimal places as you like, equal to $\widehat{\mu}$? You may have guessed, by the subtle way I've asked that question, that the answer is "no." And you'd be right! Suppose $\widehat{\mu} = 5.41$. Maybe μ *is* 5.41, but it might also be, say, 5.40, or 5.39, or other values close by, mightn't it? This is a fancy way to state that we are uncertain what the value of μ is. How do we express this uncertainty? Use probability? No. It is *forbidden* to use probability to quantify the uncertainty of parameter values in classical statistics.

Instead, classical statisticians use something called a *confidence interval*, which is an interval on the order of $\widehat{\mu} \pm c(n)$, where $c(n)$ is some number that usually depends on the number n of your data points and on the old data itself. Bigger $c(n)$ lead to wider intervals; smaller $c(n)$ lead to narrower ones. So you might expect that when you say that "I think μ is 5.41 plus or minus 4" you have a better chance of being right then when you say "I think μ is 5.41 plus or minus 1", because the former interval allows you greater scope of covering the actual (unobservable) value of μ. And, classically, you'd be dead wrong.

Which is why confidence intervals are one of the screwiest things to come out of the classical tradition, in that they fail utterly to do what they set out to do. But their use is so ubiquitous (not to say iniquitous) that I'm afraid you are going to have to learn to interpret them. *And they are one of the most important things you* must *learn in this book!* because you will see confidence intervals everywhere, thus it is imperative you learn what they are and what they are not.

Part of the problem is that you simply cannot learn what a confidence interval is by reading most introductory statistics books. Take, for example, the very typical book *Statistics: Informed Decisions Using Data* by Sullivan (2007, pp. 448-449), often used in Stats 101 courses. He officially defines a confidence interval for an unknown parameter as "an interval of numbers" (p. 449), which is as pure a tautology as you're ever likely to meet, and being a tautology, it is therefore, of course, true, but of no help (it says the confidence *interval* is an *interval*). But a page earlier, we find Sullivan implying that smaller intervals give us less confidence in the value of the

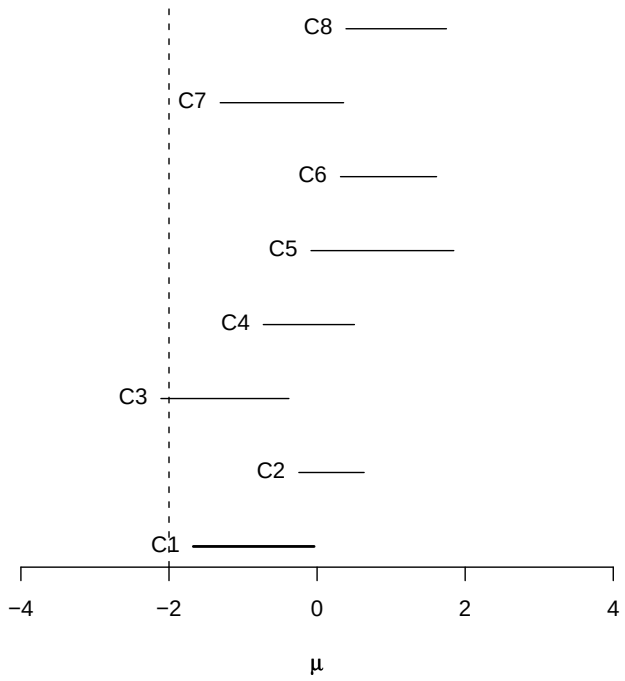
parameter than larger intervals. This implication is, as I said above, false, and is no part of the actual, mathematical definition of a confidence interval.

Maybe something like this is more accurate:

> [A] 95% level of confidence...implies that, if 100 different confidence intervals are constructed...we will expect 95 of the intervals to include the parameter and 5 to not include the parameter [p. 449].

Actually, we can expect nothing like this. And though this definition is closer to the truth, it is still false (to find out why, keep reading). Incidentally, classical theory lets you calculate confidence intervals at any level you want, but the only one you ever really see is the 95% interval, so that one is all I will talk about.

Here's the actual definition. Suppose you gather some data and construct a confidence interval using the formula $C_1 = \{\widehat{\mu} \pm c(n)\}$ (the actual formula is not of much interest to us; the software will give us the interval automatically). That is, C_1 is the interval calculated using the data we just collected. Now imagine (incidentally, this is all you can do) that you re-collect your data in exactly the same way, where every physical thing is exactly the same as it was when you collected it the first time. That is, the state of the universe has to be identical to where it was when you first collected your data. Except that it must be "randomly" different, or different in ways that you know nothing about. Very well, you now have a second data set equal in every way to the first, except that it is "randomly" different, whatever that means. You then construct a new confidence interval C_2 using the exact same formula on this second set of data (which is also the same size, n). Now do it all again and construct C_3, and again for C_4, and again and again an *infinite* number of times. When you are done, 95% of *those* intervals will cover the actual value of μ. *That*—and nothing else—is the true definition of the confidence interval.



This is shown in the picture for the eight confidence intervals for some imaginary scenario. The true value of μ is indicated by the dotted line. Some of the intervals "cover", i.e. contain, the true value of μ, and some do not. More than that, we cannot say. *Our* confidence interval, the bottom bold one, is the *only* confidence interval we'll actually see; the others are hypothesized entities that are conjured into existence if confidence intervals are properly interpreted.

I only showed the first 8 (out of an infinite number of) confidence intervals (that are said to exist for every problem you ever do). If you only repeat your experiment a finite number of times, and therefore only have a finite number of confidence intervals, say, 1,000,000, then it is false that we expect any number of them will cover the true value of μ: stopping constructing confidence intervals at any finite value invalidates the interpretation that 95% of intervals will cover the actual value of μ.

Yes, this is the actual definition, but saying it this way leaves a bad taste in one's mouth, especially because of that bit about "infinite" numbers of repetitions. Statisticians, feeling uneasy about infinities, and their physical impossibility, usually resort to the euphemism "long run" to describe the number of repetitions needed. They know very well that, mathematically,

long run equals infinite, but saying "in the long run" gives the comfortable impression that all you need is a lot, and not an infinite number, of repetitions.

By now you are thinking, "OK, I get it. So what? What you're saying is just a quibble. Who cares about infinities or long runs, anyway. Give me some information I can use! What do you do with *your* confidence interval, the one you just constructed? What does *it* mean?"

Nothing. Not a thing. It certainly does *not* mean that you are 95% sure that your interval contains the actual value of μ. That is, you *cannot*, under any circumstances, say that "There is a 95% chance that the true value of μ lies in the 95% confidence interval I have constructed." That statement, after all, is a direct probabilistic statement about the interval you have just created. Recall our key rule: it is forbidden in classical statistics to make direct probability statements about unobservable parameters. Memorize this. Your confidence interval only has an interpretation as part of an infinite set of other confidence intervals.

We have just hit upon the dirtiest open secret of classical statistics. There is *no* interpretation of *your* confidence interval other than this: the best you can say is that *your interval either contains the actual value of μ or it does not*, a statement which is a tautology, and, again therefore, true, but of no help (incidentally, Sullivan (2007) finally acknowledges this on p. 500). So what do you do with the interval you have just created? Why even bother, since it has no direct relation to the problem at hand? It's even worse. Pick any two different numbers, say, 12 and 42. It is a true statement to say that *this* interval either contains μ or it does not for *any* statistical problem done by anybody with any data any time whatsoever (make sure you understand that before reading further).

The guy that invented confidence intervals, Dzerzij (Jerzy) Neyman, a statistician, knew about the interpretational problems of confidence intervals, and was concerned. But he was even more concerned about something called inductive arguments. An example due to Stove Stove (1986): All the flames I have observed before have been hot (the premise); therefore, this flame will be hot (the conclusion). Neyman, and many other influential 20th century statisticians, rejected inductive arguments a basis for probability. They felt arguments like these were "groundless" or that inductive arguments were fallible because of the true statement that, for the flames example, there was nothing in the universe *guaranteeing* that this flame will be hot[3]. Inductive arguments are needed to make direct probabilistic statements about things like confidence intervals. If you reject them, then you cannot use probability. So Neyman, and those who followed him (which was nearly everybody), tried to take refuge in arguments like this: "Well, you cannot say that there is a 95% chance that the actual value of the parameter

[3]To which you can argue; Ok, if you doubt it, stick your hand into this flame (Briggs, 2006).

is in *your* interval; but if statisticians everywhere were to use confidence intervals, then *in the long run*, 95% of *their* intervals will contain their actual values." Thirty-two extra credit points to those who can show the obvious flaw in this argument (see the homework).

The flaw in that argument was so obvious that it was evident to Neyman himself. And so, with nowhere else to turn, Neyman recommended a dodge and said this: " The statistician...may be recommended...to state that the value of the parameter μ is within (the just calculated interval)" merely by an act of will (Neyman (1937), quoted in Franklin (2001a)).

What you would like to be able to say is that "I have 95% (or whatever) confidence that *this* interval covers the true value of μ." But you can *never* do this in classical statistics.

In `R`, to get the confidence interval of a normal distribution classically is a little more work than just getting the estimates, but it isn't really that hard. This is for the appendicitis data, the `White.Blood.Count` (don't forget to read the data in and `attach` it):

```
confint(glm(White.Blood.Count~1))
```

The function `confint` calculates 95% confidence intervals. The inside function `glm`, with that funny argument `~1`, basically says, "The uncertainty in the variable should be quantified by a normal distribution." Just take my word for it now; we'll see this function later and this notation will become clear then. Anyway, after you run the command you will see something like this:

```
    2.5 %    97.5 %
9.991874 10.818126
```

Ignore the word `(Intercept)`, it is actually `White.Blood.Count` (this is because this function works for any variable name you care to enter). The `2.5 %` and `97.5 %` are like the quantiles; subtract 2.5 from 97.5 and get the length of the interval, which is 97.5%-2.5% = 95%.

We could use another `R` function and compute the confidence interval for $\widehat{\sigma}$, but it is not of great interest because later, we'll see how to do all these things more or less automatically. Besides, we want to concentrate on what these intervals mean. If you've already forgotten, then go back and read this section from the beginning. One thing that is certain is that confidence intervals say *nothing* about the observables, the data x. If they say anything, they say something about the unobservable parameters. But what? The interval we computed for white blood count was about $[10, 11]$. This is an interval about estimated central parameter $\widehat{\mu}$ and *not* about the mean. We `know` the mean (it is...? find it in `R`). The confidence interval is an attempt to put a measure of precision on the guess $\widehat{\mu}$. It says nothing about the mean, and nothing about actual values of white blood count. Never forget this.

5. Bayesian way

The idea behind modern statistics is that you quantify any and all uncertainty you have in anything using probability. We've already seen how to quantify uncertainty using probability for observables; that is, for actual data. That turns out to be done the same way classically and Bayesianly. This is what we did the first few Chapters, was it not? We wrote down some probability distribution, with known parameters, and made probability statements about observable data. Classical and Bayesian statistics begin to diverge when we start to talk about unknown parameters and how to make guesses about these parameters.

We made guesses classically by specifying some ad hoc function of the data, giving us $\widehat{\theta}$; afterwards, we created a confidence interval for this guess. I stressed, heavily, that this confidence interval is not designed to express any actual uncertainty in θ, because that goes against the classical philosophy: which is that you *cannot* directly express uncertainty in unobservable parameters using probability.

In Bayesian statistics, you *can*, and *must*, express uncertainty in unobservable parameters using probability. How this works might sound complicated, and some of it is, but once you get how it works for, say, normal distributions, you will then know how it works for every other statistics problem in the world. This is not so for classical statistics, where you have to memorize a new set of ad hoc functions for every problem. In this way, Bayesian statistics is a vast simplification; however, before you can reach this simplification plateau, you initially have to climb up a steeper hill than you do classically. However, the good news is that there is only *one* hill to climb.

Let's recall the normal probability distribution (density function):

$$p(x|\mu,\sigma,\mathrm{E}_N) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

written here as a function of x, or $p(x|\mu,\sigma,\mathrm{E}_N)$ (we could have use N() as before; the actual letter does not matter). Do you remember probability rule number 4, or Bayes's rule? If not, go back and re-read Chapter 2. Pay special attention to equation (6). I'll wait here until you're done.

Back? OK, let's write equation (6) using different letters, so that

$$\Pr(\mathrm{B}|\mathrm{AE}) = \frac{\Pr(\mathrm{A}|\mathrm{BE})\Pr(\mathrm{B}|\mathrm{E})}{\Pr(\mathrm{A}|\mathrm{E})}$$

becomes

$$p(\mu,\sigma|x,\mathrm{E}_N) = \frac{p(x|\mu,\sigma,\mathrm{E}_N)p(\mu,\sigma|\mathrm{E}_N)}{p(x|\mathrm{E}_N)}, \tag{20}$$

where B is now (μ,σ) and A is x. Remember, (μ,σ) is shorthand for the statement "The value of the central parameter is μ *and* the value of the spread parameter is σ", and x is shorthand for the statement X = "The

value of the observed data is x." We already know how to write $p(x|\mu, \sigma, \mathrm{E}_N)$ mathematically. Our goal is to discover how to write the left-hand side, which is the probability distribution of (μ, σ) given the data and E_N. This quantifies our uncertainty in (μ, σ) given what we learned in the data (and considering the evidence E_N). In order to calculate the left-hand side, we then also need to know $p(\mu, \sigma|\mathrm{E}_N)$. We also need $p(x|\mathrm{E}_N)$, but once we know $p(\mu, \sigma|\mathrm{E}_N)$, it automatically pops out because of some math that need not concern us here.

What is $p(\mu, \sigma|\mathrm{E}_N)$? Well, it quantifies our uncertainty in (μ, σ) before seeing any data, that is, it is only conditional on E_N. $p(\mu, \sigma|\mathrm{E}_N)$ is a probability distribution that you have to specify before you can get to the probability $p(\mu, \sigma|x, \mathrm{E}_N)$. It also has an official name, which is *the prior*, because it's what you know about (μ, σ) *prior* to adding in information in the data. Not surprisingly, then, $p(\mu, \sigma|x, \mathrm{E}_N)$ is called *the posterior*, which is the probability distribution expressing everything we know, all our uncertainty, about (μ, σ) *after* having seen some data x.

How about the value of $p(\mu, \sigma|\mathrm{E}_N)$? Well, it's turns out to be a complicated situation, but the gist of it is that $p(\mu, \sigma|\mathrm{E}_N)$ explains the probability of each possible value of (μ, σ), and since we initially know very little of (μ, σ), every possible value of (μ, σ) is more or less equally probable. This situation is called *assigning a flat prior*, the "flat" describing the shape of the probability distribution picture (i.e., a flat line)[4](for a discussion of priors, see Jeffreys, 1998). Once you have the prior, and $p(x|\mu, \sigma, \mathrm{E}_N)$, you can then calculate the posterior using equation (20). Technically, since we are saying (μ, σ) has a certain probability distribution, this is also information that we should keep note of, but we'll append this on E_N so that it now means "The uncertainty in the observable is quantified by a normal distribution *and* the prior on the parameters is 'flat'." If we need to be careful about this, and sometimes we do (not in this book), we can expand the notation to indicate the exact kind of prior we use.

Now here is another little secret: for very simple situations, *the Bayesian results are the same as the classical results!* No new calculations have to be learned or done!

After we take some old data, we can calculate our full uncertainty in (μ, σ) by drawing pictures of the probability distributions (we'll do this later). If we are forced to pick just one "best" value, we would pick the arithmetic mean and standard deviation, exactly like in classical statistics. If we wanted to express our uncertainty a little more fully than just using one number (for each parameter), we could give the best number and an interval, some plus/minus bound on how certain that best value actually matches the

[4]There is more than one prior that you can use besides this "flat" one, but the differences it makes in the posteriors is minimal. Another problem is that the parameters are usually assumed to be continuous numbers, and if you recall the discussion from Chapter 4, you know these can be a problem. We will ignore all these difficulties in this book.

true value of (μ, σ). Here is the best part: the confidence interval, which was meaningless before, *is* this interval, and is now called a *credible interval*. It has the natural interpretation that there *is a 95% chance that the true value of the parameter lies in this interval*. Isn't that wild?

Before you start thinking, "Hey, if the results are the same, why did you go on and on and on about how confidence intervals are meaningless? All you did was to give them a new name! Big deal. You are wasting my time and trying to confuse me." Hold on a minute, though. The Bayesian results are the same as the classical ones, but *only* for *simple* situations. The good news for you is, that in this book, you hardly move beyond these very simple situations. Once you do move into the great statistical beyond, like using Binomial instead of normal distributions, the Bayesian methods really come into their own, and then you *cannot* assume the classical computations give you the correct answer. I'll talk about these techniques as we move along.

6. Homework

(1) What, if anything, is wrong with this sentence, "The mean of this normal distribution is 4, and the standard deviation is 2."
(2) Look around you and directly measure an observable. Make at least eight measures of each observable. Then, using `R`, calculate the classical estimates $\widehat{\mu}$ and $\widehat{\sigma}$. Store values in `R` like this example: `x = c(3.4,5.2,6.9,1.2)`.
(3) I played petanque and measured the distance several times, assumed a normal distribution to quantify uncertainty, and computed the mean, which was -1.8 cm. The classical 95% confidence interval was -6.4 cm to 2.8 cm. Which of the following statements which are true: (a) There is a 95% chance that μ is in the interval; (b) There is at least a 95% chance that $\widehat{\mu}$ is in the interval; (c) If I were to play petanque 100 times and each time I calculate a mean and construct a confidence interval, then about 95 of those intervals will contain μ; (d) Either μ is in the confidence interval or it isn't; (e) If I had made more throws in my game, i.e. had a larger n (sample size), then I'd be more certain that μ was in my constructed confidence interval.
(4) You read in a newspaper a story which reports that famous scientists discovered that, "Using classical statistical methods, we are now 95% confident that the mean age at which people develop conniption fits is 48 to 54 years." There are two things wrong with this statement: what are they? Try to answer *succinctly*.
(5) I played petanque and measured the distance several times, assumed a normal distribution to quantify uncertainty, and computed the mean, which was -1.5 cm. The Bayesian 95% credible interval was -2.8 cm to 1.4 cm. Which of the following statements which are true: (a) There is a 95% chance that μ is in the interval; (b) There is at least a 95% chance that $\widehat{\mu}$ is in the interval; (c) If I were to play petanque 100 times and each time I calculate a mean and construct a confidence interval, then about 95 of those intervals will contain μ; (d) Either μ is in the confidence interval or it isn't; (e) If I had made more throws in my game, i.e. had a larger

n (sample size), then I'd be more certain that μ was in my constructed confidence interval.

(6) EXTRA: What is the flaw in Neyman's argument that, sure, your individual confidence interval has no probabilistic interpretation, but in the long run, all confidence intervals created for all problems will cover their true values 95% of the time?

CHAPTER 9

Estimating and Observables

1. Binomial estimation

In the 2007-2008 season, the Central Michigan football team won 7 out of 12 regular season games. How many games will they win in the 2008-2009 season? In Chapter 4, we learned to quantify the probability in this number using a binomial distribution, but we assumed we knew p, the probability of winning any single game. If we do not know p, we can use the old data from last season to help us make a guess about its value. It helps to think of this old data as a string of wins and losses. So that, for the old x, we saw $x_1 = 0, x_2 = 1, \ldots, x_{12} = 1$, which we can summarize by $k = \sum_i x_i$, where $k = 7$ is the total number of wins in $n = 12$ games.

Here's the binomial distribution written with an unknown parameter

$$\Pr(k|\theta, n, \mathrm{E}_B) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{21}$$

where θ is the success parameter and k the number of successes we observed out of n chances.

How do we estimate θ? Two ways again, a classical and a modern. The classical consists of picking some function of the observed data and calling it $\widehat{\theta}$, and then forming a confidence interval. In `R` we can get both at once with this function

binom.test(7,12)

where you will see, among other things (ignore those other things for now),

```
95 percent confidence interval:
 0.2766697 0.8483478
sample estimates:
probability of success
             0.5833333
```

This means that $\widehat{\theta} = 0.58 = 7/12$ so again, the estimate is just the arithmetic mean. The 95% confidence interval is 0.28 to 0.84. Easy. This confidence interval has the same interpretation as the one for the $\widehat{\mu}$, which means you *cannot* say there is a 95% chance that θ is in this interval. You can *only* say, "either θ is in this interval or it is not.

Here is Bayes's theorem again, written as functions like we did for the normal distribution

$$p(\theta|k,n,\mathrm{E}_B) = \frac{p(k|\theta,n,\mathrm{E}_B)p(\theta|\mathrm{E}_B)}{p(k|n,\mathrm{E}_B)} \tag{22}$$

We know $p(k|n,\theta,\mathrm{E}_B)$ (this is the binomial distribution), but we need to specify $p(\theta|\mathrm{E}_B)$, which describes what we know about the success parameter *before* we see any data, given only E_B ($p(k|n,\mathrm{E}_B)$ will pop out using the same mathematics that gave us $p(x|\mathrm{E}_N)$ in equation (20)). We know that θ can be any number between 0 and 1: we also know that it cannot be exactly 0 or 1 (see the homework). Since it can be any number between 0 and 1, and we have no a priori knowledge which number is more likely than any other, it may be best to suppose that each possible value is equally likely. This is the flat prior again[1]. Again, technically E_B should be modified to contain this information. After we take the data, we can plot $p(\theta|k,n,\mathrm{E}_B)$ and see the entire uncertainty in θ, or we can pick a "best" value, which is (roughly) $\widehat{\theta} = 0.58 = 7/12$, or we can say that there is a 95% chance that θ is in the (approximate) interval 0.28 to 0.84. I say "roughly" and "approximate" here, because the classical approximation to the exact Bayesian solution isn't wonderful for the binomial distribution when the sample size is small. The homework will show you how to compute the precise answers using `R`.

2. Back to observables

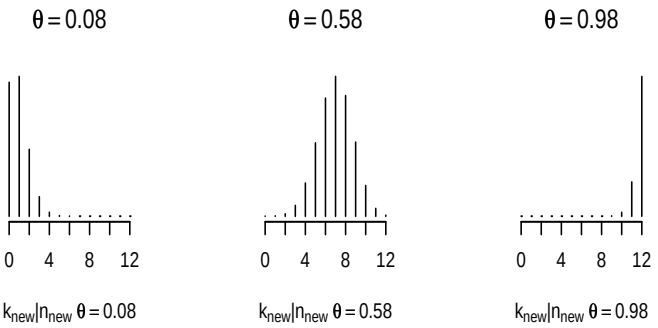
In our hot little hands, we now have an estimate of θ which equals about 0.58. Does this answer the question we started with? That question was How many games will CMU win in the 2008-2009 season? Knowing that θ equals something like 0.58 does not answer this. Knowing that there is a 95% chance that θ is some number between 0.28 to 0.84 also does not answer the question. This question is *not* about the unobservable parameter θ, but about the future (in the sense of not yet seen) observable data. Now what? This is one of the key sections in this entire book, so take a steady pace here.

Suppose θ *was exactly equal* to 0.58. Then how many games will CMU win? We obviously don't know the exact number even if we knew θ, but we could calculate the probability of winning 0 through 12 games using the binomial distribution, just as we did in Chapters 3 and 4. We could even draw the picture of the entire probability distribution given that θ was exactly equal to 0.58. But θ might not be 0.58, right? There is some uncertainty in its value, which is quantified by $p(\theta|k_{\text{old}},n_{\text{old}},\mathrm{E}_B)$, where now I have put the subscript "old" on the old data values to make it explicit that we are talking about the uncertainty in θ given previously observed data. The parameter *might* equal, say, 0.08, and it also *might* equal 0.98, or any other

[1]Like before, there are more choices for this prior distribution, but given even a modest sample size, the differences in the distribution of future observables due to them is negligible

value between 0 and 1. In each of these cases, given that θ exactly equalled these numbers, we could draw a probability distribution for future games won, or k_{new} given $n_{\text{new}} = 12$ (12 games next season) and given the value of θ. Make sure you understand this before reading further.

Let us draw the probability distribution expressing our uncertainty in k_{new} given $n_{\text{new}} = 12$ (and E_B) for three different possible values of θ.



If θ does equal 0.08, we can see that the most likely number of games next season is 1. But if θ equals 0.58, the most likely number of games won is 7; while if θ equals 0.98, then CMU will most likely win all their games.

This means that the picture on the far left describes our uncertainty in k_{new} *if* $\theta = 0.08$. What is the probability that $\theta = 0.08$? We can get it from equation (22), from $p(\theta|k_{\text{old}} = 7, n_{\text{old}} = 12, \text{E}_B)$. The chance of $\theta = 0.08$ is about 1 in 100 million (we'll learn how the computer does these calculations in the homework). Not very big! This means that we are very very unlikely to have our uncertainty quantified by the picture on the left. What is the chance that $\theta = 0.98$? About 3 in a trillion! Even less likely. How about 0.58? About 3 in 10,000. Still not too likely, but far more likely than either of those other values. We don't really need to know what the exact value of θ is anyway.

This is because we could go through the same exercise for all the other values that θ could take, each time drawing a picture of the probability distribution of k_{new}. Each one of these would have a certain probability of being *the correct* probability distribution for the future data, given that its value of θ was the correct value. But since we don't know the actual value of θ, but we *do* know the chance that θ takes any value, we can take a weighted sum of these individual probability distributions to produce one overall probability distribution that completely specifies our uncertainty in k_{new} given all the possible values of θ. This will leave us with

$$\Pr\left(k_{\text{new}}|n_{\text{new}}, k_{\text{old}}, n_{\text{old}}, \text{E}_B\right). \tag{23}$$

Stare at equation (23) for two minutes without blinking. This, in words, is the probability distribution that tells us everything we need to know about future observables k_{new} given that we know there will be n_{new} chances for

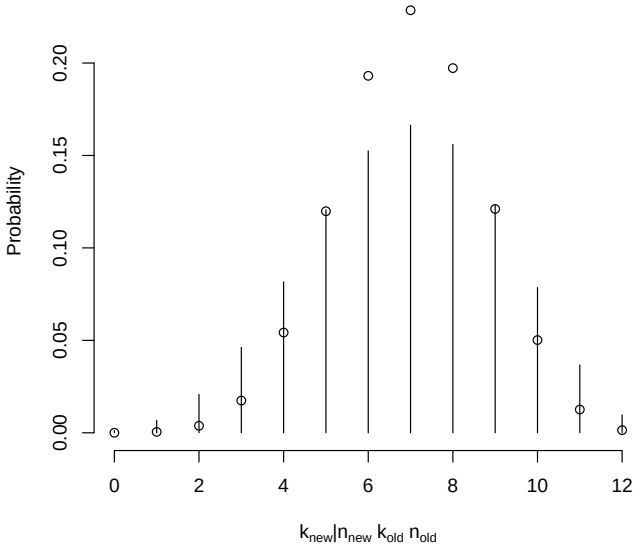


FIGURE 1. The distribution of future numbers of successes considering all possible values of θ (solid line), and for just one value, $\theta = 0.58$ (circles).

success this year, also given that we have seen the past observables k_{old} and n_{old}, and assuming E_B is true. Think about this. You do not know what future values of k will be, do you? You do know what the past values are, right? So this is the way to describe your uncertainty in what you do not know given what you do know, taking full account of the uncertainty in θ, which is not of real interest anyway.

The way to get to this equation uses math that is beyond what we can do in this class, but that is unimportant, because the software can handle it for you. This picture (Fig. 1) shows you what happens. The solid lines are the probability distribution in equation (23). The circles plotted over it are the probability distribution of a regular binomial assuming θ exactly equals 0.58. The key thing to notice is that the circles distribution, which assumes $\theta \equiv 0.58$ is too tight, too certain. It says the center values of 6 to 8 games won are more certain than is warranted (their probability is higher than the actual distribution). It agrees, coincidentally only, with the probability that the future number of wins will be 5 or 9, but then gives too little probability for wins less than 5 or greater than 9.

The actual distribution of future observable data (23) will *always* be wider, more diffuse and spread out, less certain, than any distribution with

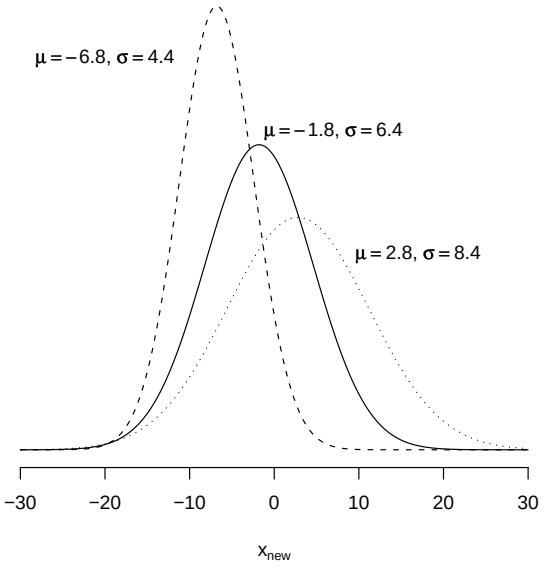


FIGURE 2. Three possible distributions of future observables, given three different sets of parameters.

a fixed θ. This means we *must* account for uncertainty in the parameter. If we do not, we will be too certain. And if all we do is focus on the parameter, using classical or Bayesian estimates, and we do not think about the future observables, we will be far, far more certain than we should be. Unfortunately, most statistical do stop short at parameters and so too many people are too certain about too many things.

3. Even more observables

Let's return to the petanque example and see if we can do the same thing for the normal distribution that we just did for the binomial. The classical guess of the central parameter was $\widehat{\mu} = -1.8\,cm$, which matches the best guess Bayesian estimate. The confidence/credible interval was -6.8 cm to 2.8 cm. In modern statistics, we can say that there is a 95% chance that μ is in this interval. We also have a guess for σ, and a corresponding interval, but I didn't show it; the software will calculate it. We do have to think about σ as well as μ, however—because *both* parameters are necessary to fully specify the normal distribution.

As in the binomial example, we do not know what the exact value of (μ, σ) is. But we have the posterior probability distribution $p(\mu, \sigma | x_{\text{old}}, \mathrm{E}_N)$ to help us make a guess. For every particular possible value of (μ, σ), we can

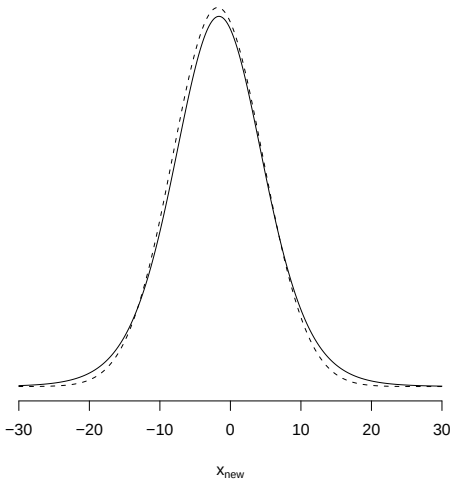


FIGURE 3. Distributions of future observables considering all possible values of (μ, σ), and another given a fixed given a fixed value of $(\mu = -1.8\,cm, \sigma = 6.4\,cm)$.

draw a picture of the probability distribution for future x given that that particular value is *the* exact value.

Figure 2 shows the probability densities for x_{new} for three possible values of (μ, σ). If $(\mu = -6.8\,cm, \sigma = 4.4\,cm)$, the most likely values of x_{new} are around 10 cm, with most probability given to values from -20 cm to 0 cm. On the other hand, if $(\mu = 2.8\,cm, \sigma = 8.4\,cm)$, the most likely values of new x are a little larger than 0 cm, but with most probability for values between -20 cm and 30 cm. If $(\mu = -1.8\,cm, \sigma = 6.4\,cm)$, future values of x are intermediate of the other two guesses. These three pictures were drawn (using the Advanced code from Chapter 5) *assuming* that the values of (μ, σ) are *the* correct ones. Of course, they *might* be the right values, but we do not know that. Instead, each of these three guesses, and every other possible combination of (μ, σ), has a certain probability, given x_{old}, of being true.

Given the old data, we can calculate the probability that (μ, σ) equals each of these guesses (and equals every other possible combination of values). We can then weight each of the new x distributions according to these probabilities and draw a picture of the distributions of new values given old ones (and the evidence E_N) like we just did for the binomial distribution. This is

$$p\left(x_{\text{new}}|x_{\text{old}}, \text{E}_N\right). \tag{24}$$

Here is a picture (Fig. 3 of this distribution (generated by the computer, of course)

The solid line is equation (24), and dashed is a normal distribution with $(\mu = -1.8\,cm, \sigma = 6.4\,cm)$. The two distributions do not look *very* different, but they certainly are, especially for very large or very small values of x_{new}. The dashed line is too narrow, giving too much probability for too narrow a range of x_{new}. In fact, for distribution (24), values greater than 10 cm are from the true distribution are twice as likely as the normal distribution where we plugged in a single guess of (μ, σ); values greater than 20 cm are six times as likely. The same thing is repeated for values less than -10 cm, or less than -20 cm, and so on. Go back and read Chapter 6 6 to re-familiarize yourself with the fact that very small changes in the central or spread parameter can cause large changes in the probability of extreme numbers.

The point again, like in the binomial example, is that using the plug-in normal distribution, the one where you assume you know the exact value of (μ, σ), leads you to be far more certain than you really should be. You need to take full account of the uncertainty in your guesses of (μ, σ), only then will you be able to full quantify the uncertainty in the future values x_{new}.

4. Summary

Since this Chapter is so important, let's sum up. We want to express uncertainty in an unknown observable. We do that using a probability distribution. Probability distributions have parameters, variable numbers that are needed to fully specify the distribution. The parameters are unobservable, but we can express our uncertainty in them using previously observed data (and other evidence). Almost all statistics methods, classical and Bayesian, stop after they have said something about the unobservable parameters (given some data and evidence). However, nearly all the time we are not interested in the parameters (exceptions will be noted in Chapter 15), but we are interested in unknown observables. So we have to express our uncertainty in the future observables taking account the uncertainty we have in the parameter values. Doing this will give us the true picture of uncertainty, and we will fool ourselves far less often than using older statistical methods.

5. Homework

(1) When can't the binomial success parameter exactly equal 0 or 1? When can it?
(2) For the normal problem, write out a definition of $p(\mu, \sigma^2 | x_{\text{old}}, \text{E}_N)$ and $p(\mu, \sigma^2 | \text{E}_N)$. Explain what the difference between these two things are.
(3) Write out the difference between $p(\theta | k_{\text{old}}, n_{\text{old}}, \text{E}_B)$ and $p(\theta | \text{E}_B)$.
(4) See the last homework question in Chapter 7. Type this in `R`:
`source(url("http://wmbriggs.com/book/Rcode.R"))`
Recall that this loads the file `Rcode.R` into `R`'s memory. You will now have available two functions. The first is `newdbinom(x, n_new, k_old,`

`n_old)` which operates just like `dbinom(x, n_old, p)` did, except it gives you the probability of seeing `x` *new* successes out of `n_new` chances given you saw `k_old` successes out of `n_old` chances. The second function is `newpnorm(x, x_old)` which operates something like `pnorm(x, central, spread)`, and gives you the probability of seeing values less than or equal to `x` given the old data `x_old`. These two functions first estimate the posterior probabilities of θ and (μ, σ) (but you don't see them) and then uses these to calculate the probability distributions of future observables, which are of main interest.

(5) Find something, some observable, the uncertainty of which can be modeled by a binomial distribution. Count the number of times that thing could have been a success, and count how many times it was a success. An example might be "People you meet who watch wrestling on TV." Any observable that you think has a constant probability of a success will do. State what the n_{old} and k_{old} were. Find the classical estimate and confidence interval of $\widehat{\theta}$. Now assume that the thing you picked will have 3 more opportunities to be a success. That is $n_{\text{new}} = 3$. What is the probability that $k_{\text{new}} = 0, 1, 2, 3$? For the CMU football example, you would type `newdbinom(0:3, 3, 7, 12)` (of course, in the CMU example, we expect $n_{\text{new}} = 12$, but we are supposing we are only looking at the first three games of the season). Use the estimate $\widehat{\theta}$—-assume that this is the exact probability of a success—in the function `dbinom(0:3,n_old,`$\widehat{\theta}$`)` and compare this to the values given in `newdbinom`. Comment on the differences.

(6) Find some observable the uncertainty of which can be approximated by a normal distribution. An example might be "Number of pairs of shoes." Any observable that you think using a normal distribution to approximately quantifying the uncertainty of will do. In `R`, store that data as `x = c(ob1, ob2, ...)` where `ob1` is the first observation, and so on. Find the classical estimate and confidence interval of $\widehat{\mu}$. Find the classical estimate *widehat*σ. What is the probability that a new observation is less than $\widehat{\mu}$? Find this by typing `newpnorm( mean(x), x)`. Also type `newpbnorm(mean(x) - sd(x), x)` and `newpnorm(mean(x) + sd(x), x)`, which gives you the probability of seeing new values less than $\widehat{\mu} - \widehat{\sigma}$ and less than $\widehat{\mu} + \widehat{\sigma}$. Compare these to the probabilities of seeing data using the classical estimates as plug ins. For example `pnorm(mean(x) - sd(x), mean(x), sd(x)`. Comment on the differences.

CHAPTER 10

Testing

1. First Look

Here are two experiments:

(1) Uncle Ted wanted to test two promotional advertising campaigns at two different restaurants at his chain of Kill 'em and Grill 'em Venison Burger joints. For the Detroit restaurant, he gave out free plastic antler-hat sets for the kids (call this campaign A), and for Chicago he gave out, to the adults, one round of .30-06 Springfield Ammunition, 180 Gr SP (Per 20) CB (call this campaign B). He measured the number of venison sausage sandwiches sold over the course of a week in both places and wondered which advertising campaign was more effective. Before the campaign, both stores sold about the same amount of sandwiches.

(2) The Army decided to test the hand-eye coordination of right-handed recruits who had undergone a new version of super symmetric physical training by having them shoot aliens in the classic video game Space Invaders. Twenty recruits first shot using their left hand, then shot using their right hand. The number of aliens blasted using each hand was counted. The Army wanted to know whether more aliens would be blasted using their "best", i.e. right, hand: it was hoped the super-symmetric training would make the hands equally proficient.

Both of these situations are common, but slightly different. Uncle Ted wants to know the difference between two advertising campaigns, where each campaign presumably does not affect the another. The people in Chicago do not hear about the antlers, and the folks in Detroit do not hear about the ammo. The Army wants to know the difference in aliens blown away by each recruit's hand, but where we can guess that, for each recruit, the total shot using either hand may be related; that is, some recruits may naturally be better at video games than others, and would waste more aliens, so we should somehow take this into account.

Start with the ad campaign. How can we tell if the effectiveness of the two ads is different? Obviously, it has something to do with the number of sausages sold. Suppose Detroit sold 1000 and Chicago 1005 over the first week of the campaigns. Does that mean that the Chicago ad campaign did a better job? In a way, yes. We already knew that before the campaigns

the number of sausage sales in both cities was "about the same." Is 1000 and 1005 about the same? Maybe; probably. What seems clear is that this one data point, this one week of sales, is not enough information to be able to tell if the results from the two campaigns will continue to be different. Uncle Ted is going to have to suck it up, give the advertising agency more money, and continue the ad campaigns for at least a few more weeks so he can collect more data. Suppose he does, and here are the results:

	Week 1	Week 2	...	Week 20
Campaign A	302	355	...	402
Campaign B	280	426	...	418

The full dataset is on the class website. Incidentally, it is actually stored, and should be stored, like this:

Campaign	Sales	Week
A	302	1
A	355	2
⋮	⋮	⋮
B	208	1
⋮	⋮	⋮
B	418	20

This is the more general way, and the way that makes it easiest to use in software. Suppose that the number of weeks the campaigns lasted at each store was different. Then storing data the first way stinks, because you have an uneven number of columns between the two campaigns. Storing data the later way means the weeks, even the number of campaigns (Uncle Ted may expand to more cities, and have campaigns C, D, etc.) can be different. Plus we can store more information by just adding another column, as we did a few weeks back when we looked at how to store data.

What is the *first* thing you always do once you have collected some data? You *look* at it. You certainly do *not* just run some pre-packaged statistical procedure. Failing to look at their data first is *the* biggest mistake people make. Do not be lazy. Look first. Always.

How do we look at this data? We already know that the two campaigns cannot influence one another, in the sense that knowledge of what happens in one campaign is irrelevant to knowing what happens in the other. So we can look at each campaigns' data somewhat independently. Looking at data is somewhat of an art; there are few hard rules that say, "For situation 1, use a boxplot; for 2 use a scatter; and so on." No, we often plot and re-plot and re-re-plot and re-re-...—you get the idea—until we find a way that displays the data nicely.

In this case, we already have some experience with data like this: namely, simple summaries, boxplots, histograms, and density estimates. Let's first read this data into R and see what we can come up with.

```
x = read.csv(url("http://wmbriggs.com/book/advertising.csv"))
summary(x)
```

If you want to create your own data set, open a spread sheet (like OpenOffice.org) and save it as a CSV file, like was discussed in Chapter 7. Store the file some place on your hard disk that you can remember, and read it in like this:

```
x = read.csv("C:/mydata/MyFile.csv")
```

The summary information of the advertising.csv data is somewhat useful, but it doesn't break the means etc. down by campaigns. We only see the overall mean, median, etc., and count of the number of campaign As and Bs we had. We have to do something else.

```
attach(x)
boxplot(Sales~Campaign)
```

Recall that the attach(x) function makes the variable names of the data x "visible" to R. If you forget this step—and you will, you will—you'll see this error[1]

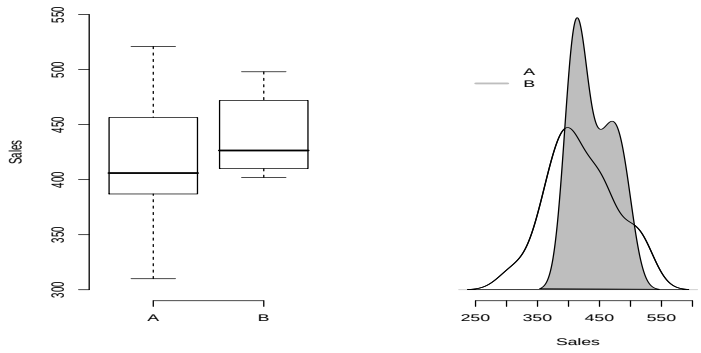
```
Error in eval(expr, envir, enclos):object "Sales" not found
```

You can always forgo the attach(x) command if you tell the boxplot() function which data you had in mind.

```
boxplot(Sales~Campaign,data=x)
```

but it's a pain to type all that extra stuff.

Anyway, here's what you get:



Pretty nifty, eh? We can directly compare, at least roughly, the distribution of the observables (sales) side by side. There is at least some different between the two campaigns.

ADVANCED: The other, more complicated plot, is done like this (see the Advanced Section in Chapter 7)

[1]For goodness' sake: do not panic when this happens.

```
plot(density(Sales[Campaign=="A"]))
lines(density(Sales[Campaign=="B"]),lty=2,col=2)
```

This creates the two density estimates of the observables, and plots them together. The first line plots the density of sales figures only at those times when `Campaign=="A"` (those brackets tell `R` to look for a subset, the statement tells it which one). Recall the density estimates are fancy, souped-up histograms. Don't worry about trying to remember how this is accomplished. You can always look it up. Sometimes it comes in handy. (You can try to stick histograms side by side, too, but that's ugly and makes it hard to compare what is happening.)

What can we notice about the boxplots? The number of weekly sales is on the vertical axis, the two campaigns on the horizontal. The medians are in different places, meaning, of course, that the number of sales was different over the weeks in the two cities. The "spread," the distance between the two extreme quantiles, of the two distributions isn't quite the same. There's no need to be precise about this; we're just getting a qualitative feel. If the two distributions are wildly different, say, the spread is nowhere near the same, or one or both distributions show asymmetry in the quantiles, then you will need to do something you won't learn in this book.

You might also need to do a time series plot of the data, since the campaigns' effectiveness may weaken, strengthen, or do anything in between these, over time. If the data do look like there's some kind of "signal", or changes through time, you *cannot* use the techniques you'll learn here. Although, of course, many people do; at the very least, it will only be an approximation. This is fine if you know what you're doing, but if you don't, then you are misleading yourself, others, and humanity in general. (To get a time series plot, type `plot(Sales[Campaign=="A"],type='l')` and the same for B.)

2. Classical 1

We now move into the quaint, and very confusing, realm of *Hypothesis Testing.* That's the classical term, anyway, for the statistical methods of deciding when things are different. A slightly more modern term is *decision analysis*, which is more descriptive, but is a field that does more things than you normally get in a book like this.

Here is the main question again. We have several weeks of sales in two cities where two different ad campaigns have run. What is the probability that the sales of sausages is different due to the ad campaigns? Sadly—very sadly—it turns out that *we cannot answer that question directly in classical statistics: in hypothesis testing we can never say anything directly about the observables*, neither can we make direct statements about parameters. Yes, differences in the observables is we want to know, but the best we can do classically is to answer a proxy question.

The proxy question always goes like this (pay attention here, this is hard, hard stuff to understand): Assume that the uncertainty in the observables, in the different categories like sales campaigns, can be quantified by probability distributions. This, as we already know, is a reasonable thing to do. These probability distributions, as we also know, will have certain parameters: like μ and σ^2 for the normal, or θ for the binomial.

We then assume that at least some of these parameters will have different values in the different categories: e.g. a μ_A for ad campaign A and μ_B for campaign B; while it may be that $\sigma_A^2 = \sigma_B^2$ or $\sigma_A^2 \neq \sigma_B^2$for both campaigns. That is, the uncertainty in the sales for different campaigns will quantified by normals distributions with the same spread parameter (usually), but with different central parameters.

Last Chapter, we learned how to estimate central parameters. This means that $\widehat{\mu}_A$ is the mean of sales in campaign (or city) A. This turns out to be (glance at the boxplot; we'll learn how to do this officially in a moment) $\widehat{\mu}_A = \overline{\text{Sales}}_A = 421$ and $\widehat{\mu}_B = \overline{\text{Sales}}_B = 440$. Sit down for this next question. *What is the probability that* $\overline{\text{Sales}}_A = \overline{\text{Sales}}_B$? This is *not* a trick question!

The probability is *0*; that is, it is false that the two means are equal. One mean is 421 and the other is 440. These numbers are obviously not equal. Do not laugh.

Now a different question. What is the probability that $\mu_A = \mu_B$ (we already *know* that $\widehat{\mu}_A \neq \widehat{\mu}_B$)?. It turns out that this question is *forbidden* in classical statistics. The reason it is forbidden is that, if you remember, and you should, you are not allowed to make probability statements about unobservable parameters. Asking the question is ruled out. Instead, classically, we turn the question around and ask something else.

Let's be sure where we are first. We have two sets of data, from two campaigns, the uncertainty of both sets quantified by normal distributions, each with its own central parameters, but both share the same spread parameter. We want to use the data we have to ask questions about the differences in the ad campaigns, such as are sales A greater than sales B? Or are sales A at least thirty-percent larger than B, and so on. If the sales campaigns were ceased tomorrow for all time, our work would be done, would it not? We could sum up the sales in city A and those in B and easily say whether sales of A were larger than B, or that they were at least thirty-percent larger, or *whatever* question about the observable sales we wanted to ask. We *do not* need any fancy statistics, or hypothesis testing, *unless* we are asking questions about *data not yet seen.* That is, data we might see if we extended the ad campaigns into the future. Make sure this is stuck firmly to your sticking place before reading further.

Thus, assuming we will see future data, the following procedure is used classically:

(1) First, *assume* $\mu_A = \mu_B$ (along with assuming normal distributions etc.).
(2) Calculate a *statistic* $t(x)$, which is an ad hoc function of the data (you look these up in books, or trust the software to do them for you. Incidentally, this is usually called "differences in the means problem", which makes no sense because we already know whether the means are different).
(3) Then calculate this magic number:

$$\text{p-value} = \Pr(T(x) > |t(x)| \ | \mu_A = \mu_B, \sigma_A = \sigma_B, \mathrm{E}_N) \tag{25}$$

which is read, "The probability of seeing the statistic $T(x)$ as large or larger (in absolute value) than the statistic $t(x)$ I did see *given* $\mu_A = \mu_B$, *if* I repeated the experiment/trial/campaign an *infinite* number of additional times." That is, if you were to endlessly repeat the ad campaign (for 20 weeks each time, matching the sample size we had before), and each of these endless times you calculated a $t(x)$, then the *p-value* measures the chance that these other statistics exceed (in absolute value) the one ($t(x)$) you actually got using the observed data.
(4) You *must* memorize this: the *p-value* is *not* the probability that $\mu_A = \mu_B$, nor that $\mu_A > \mu_B$, nor is it any other direct statement about the unobservable parameters. It is a probability statement about a function of the observed data assuming something about the unobservable parameters.
(5) A ridiculous[2] tradition has developed that if your p-value ≤ 0.05 then you are allowed to say that your results are *statistically significant*. If your p-value is larger than this publishable level, then you are cast out beyond the gate where there is weeping and gnashing of teeth. If you somehow do get a p-value larger than 0.05, do not despair; see Chapter 14 for why.

There are no rules (except habit) that specify which statistic $t(x)$ to use in any situation except this: you *must* be able to figure out the probability distribution of this statistic given $\mu_A = \mu_B$. This distribution is needed to calculate the p-value (this is how you calculate the probability that another statistics is larger than your statistic). For any given problem, there are always lots of choices of statistics, and you are free to pick the one that gives the best results. Cheating? See Chapter 14.

Stating that $\mu_A = \mu_B$, that is, that the two central parameters are the same, is called stating the *null hypothesis* (you are hypothesizing that the central parameters are equal). If your p-value ≤ 0.05, then you are said to *reject* the null hypothesis; that is, to *indirectly* conclude that $\mu_A \neq \mu_B$. If your p-value > 0.05 then you are said to *fail to reject* the null hypothesis,

[2]And it is ridiculous. I recently had a client who was near tears because her p-value was—and I am not kidding—0.052. "Isn't there anything we can do to make it significant?"

not that you accept it, because, after all, how can you know if $\mu_A = \mu_B$? You cannot!

What people say is that if $\Pr(T(x) > |t(x)| \ |\mu_A = \mu_B, \sigma_A = \sigma_B, \mathrm{E}_N)$ is very low, then it is unlikely that the given information, i.e. $\mu_A = \mu_B, \sigma_A = \sigma_B, \mathrm{E}_N$, is suspect. Actually, they usually just say that $\mu_A = \mu_B$ is suspect, else it would not be so improbable to get a larger statistic. Unfortunately, this ignores the fact that you might get just a small a p-value if, say, $\mu_A = 2 * \mu_B$, or $\mu_A < \mu_B$, or whatever (Berger and Selke, 1987). You can get a *smaller* p-value, given some different, alternate hypothesis about the parameter values!

All right, that's a lot of thinking for you to do. The key points are that you cannot—you can never—make any probability statements about unobservable parameters classically. So you are forced into making weird indirect statements. You also do not make direct statements about actual observables—which are our main interest—only about the parameters. Incidentally, the statistics $t(x)$ that you calculate are decided upon by tradition, mostly.

Here is a key fact: *nobody ever remembers the definition of a p-value.* Journal editors never remember. Even statisticians typically won't remember. Everybody will be tempted beyond their capacity to resist to say, when their p-value is less than the magic number, that the probability that $\mu_A \neq \mu_B$ is high, despite the fact that this kind of statement is heresy in classical statistics. Even worse, they might say that $x \neq y$, or some other false proposition about the observables, is highly probable. Too, if your p-value is greater than the publishable limit, you will ache to (fasely) declare "It is highly probable that $\mu_A = \mu_B$." Or, worse, that it is highly probable that future $x = y$. You, and everybody else, will simply not be able to help it. You will surrender to the seductive call of the p-value, I promise you.

Now, unlike confidence intervals, which actually have no useful meaning, p-values do. The probability statements you make, given all the assumptions are true, are perfectly accurate. There is nothing, then, directly wrong with a p-value, as there was with confidence intervals. But they are evil nonetheless, because they do not answer the questions you need answered, so I discourage their use.

We are about to meet the most-used $t(x)$ in problems like the ad campaign. It is called the "t-statistic." In `R` (once the data is read in and attached)

```
t.test(Sales~Campaign)
```

One version in math:

$$t(x) = \frac{\overline{\text{Sales}}_A - \overline{\text{Sales}}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \tag{26}$$

where s_A and n_A are the calculated standard deviations and number of data points for campaign A, etc. You can see that this is just a simple function

of the difference in observed means. If $t(x)$ is "big" (in absolute value), then you get a low p-value. If $t(x)$ is small, then you get a high p-value.

In `R`, you will see something like this

```
data:  Sales by Campaign
t = -1.3285, df = 30.94, p-value = 0.1937
alternative hypothesis: true difference in
                              means is not equal to 0
95 percent confidence interval:
 -48.17209  10.17209
sample estimates:
mean in group A mean in group B
         420.75          439.75
```

The t-statistic is there, the p-value, calculated automatically, is also. The "`df`" means "degrees of freedom", it is holdover terminology from long ago (borrowed from mechanics); think of it as a function of the sample size. It is needed to figure out the probability distribution for t-statistics. The official "alternate hypothesis" is given, but using incorrect terminology. We already know the means are different; the true alternate hypothesis should read $\mu_A \neq \mu_B$. The observed means of both groups is at the bottom, but for no good reason, the actual differences in means is left for you to do by hand ($420.75 - 439.75 = -19$). Ignore the confidence intervals business for now.

The good news is that the `R` function `t.test` does *not* assume that $\sigma_A = \sigma_B$ (there are actually more than one statistic that goes by the name "t-statistic", but the differences are minor). This means the p-value is actually $\Pr(T(x) > |t(x)| \ |\mu_A = \mu_B, \mathrm{E}_N)$ (there is no dependence on the spread parameters).

The p-value is a non-publishable 0.19, and it means what? Classically, you *fail* to reject the "null". You do *not* accept it. What is the probability that $\mu_A < \mu_B$? What is the probability that future A<B? We *cannot* answer these questions classically.

That's more or less it. Every single other hypothesis testing problem fits into the sales campaign paradigm.

3. Classical 2

Let's enlist in the Army problem[3]. The main difference between it and the ad campaign problem is the nature of the data. The ad campaigns do not influence one another, but the left and right hand of each recruit is obviously related. The differences *between* the recruits are like differences between the cities, in that there is no known relation between them. Well, known to us, anyway. There is always the possibility that certain batches or subgroups of recruits will perform differently due to some reason that was not measured, such as recruits coming from a deprived upbringing, or regions with little access to technology (like Ohio). We might expect that these unfortunate

[3]Five extra credit points for those who laugh at this joke.

souls would score more poorly than their more advanced cousins. But, like we said, we do not know this information for this particular problem.

What's usually done in these situations is to create a measure that removes the dependency in some way, and the most usual way to do this is to take a difference *within* the groups, which are the recruits and their hands. So, instead of keeping track of the left-hand and the right-hand score for each recruit, we create the difference measure

$$x_i = \text{left-hand score}_i - \text{right-hand score}_i$$

for each recruit $i = 1 \ldots n$. All we have left is n numbers $x_1 \ldots x_n$. Recall what the Army had hoped: that the difference in scores between hands would be small or non-existent. If so, this would mean the x_i would be near 0. (We do not have to actually compute these x_i, the software will do it for us.)

What are the classical steps? First assume that our uncertainty in x_i is quantified by a normal distribution with central and spread parameters. Then assume that $\mu = 0$ (why?). Second, calculate some statistic, which here is—surprise!—also called a t-test. Supposing the `left` and `right` hand data is coded by its natural name, then in `R` we get it like this

```
t.test(left,right,paired=T)
```

where the only change from the other t-test is the `paired=TRUE` (or `paried=T`), to indicate the hands are part of a pair. The third step? Right, calculate the magic number. For us, suppose it is 0.051. What does that mean?

4. Modern

Similar to classical estimation, for these simple problems, the modern default Bayesian calculations turn out to be the same as the classical calculations, but again with different interpretations, though it's a little trickier in parts. Let's go back to the ad campaign `t.test` computer output and look at the part that said

```
95 percent confidence interval:
 -48.17209  10.17209
```

In Bayesian statistics, confidence intervals become credible intervals, but what is this confidence/credible interval for? We already know how to estimate μ_A and μ_B, and how to express our uncertainty in these parameters given the past data. It turns out that we can also directly estimate the "joint parameter" $\mu_A - \mu_B$ (or any other function of the two), and we can express our uncertainty in it given the past data. The best estimate of $\mu_A - \mu_B$ is the observed difference in means $\overline{\text{Sales}}_A - \overline{\text{Sales}}_B = 420.75 - 439.75 = -19$. The confidence/credible interval for this difference is the one given by the function `t.test`.

The credible interval states that there is a 95% chance that the difference $\mu_A - \mu_B$ lies in the stated interval. This is, of course, a part, but only a small part, of what you really want to know. Recall, that this is a probability

statement about unobservable parameters. We'll talk about how to answer important questions about observables later.

Another common thing to want to know is the probability that the difference $\mu_A - \mu_B$ is actually greater (or lesser) than 0 (or anything other number that might be of interest). This—we all remember by now—is not a question that can be answered, or even asked, in classical statistics. Here is a rough way to find this probability:

$$\Pr(\mu_A - \mu_B > 0 | x, \mathrm{E}_N) \approx 1 - \frac{\text{p-value}}{2}. \tag{27}$$

Be careful! This approximation is extremely crude once you venture even a little beyond the simple t-test scenario where all the assumptions are met. In our example, $\Pr(\mu_A - \mu_B > 0 | x, \mathrm{E}_N) \approx 1 - 0.19/2 = 0.91$. Since $\Pr(\mu_A - \mu_B > 0 | x, \mathrm{E}_N) + \Pr(\mu_A - \mu_B < 0 | x, \mathrm{E}_N) = 1$, then $\Pr(\mu_A - \mu_B < 0 | x, \mathrm{E}_N) \approx 0.09$. We will see how to tighten up this guess in a couple of Chapters, when we learn a unifying way to treat modern problems like this.

T-tests are only used in simple situations. Strike that. I should say that t-tests *should* be used only in simple situations. What really happens is that people apply them indiscriminately, strewing p-values like confetti. That's bad enough, but then they persist in calling t-tests as "differences between means" tests, which is wrong. Technically, and we might as well be technical since this is a technical concept, the classical t-test is not even a "differences between central parameters" test, since it is always *assumed* that the differences in central parameters is null (or 0).

Later, we are going to learn that even if we have a thrillingly small p-value, and even if $\Pr(\mu_A - \mu_B > 0 | x, \mathrm{E}_N)$ is near 1, that it still does not necessarily say that the probability os *new* observable As and Bs will be different. By concentrating solely on parameters, we will end up being too sure of ourselves.

5. Homework

(1) What is the first thing you always do when you start a data analysis?
(2) In classical hypothesis testing, what kind of probability statements about differences in the observables (for example, differences between means) can you make?
(3) In classical hypothesis testing, what kind of probability statements about differences in the unobservable parameters can you make ?
(4) In classical hypothesis testing, what *can* you say about differences in the observables or parameters?
(5) In classical hypothesis testing, what is the so-called "null hypothesis" and the "alternative hypothesis" in the advertising campaign example? Be careful about your wording!
(6) In classical hypothesis testing, what is the so-called "null hypothesis" and the "alternative hypothesis" in the army example? Be careful about your wording!

(7) Think of a situation like Uncle Ted's advertising campaign. Come up with a plausible scenario, with real-life data, using an observable the uncertainty of which can be approximately quantified by a normal distribution, and which you can measure in at least two different groups. For example, you might ask if somebody is an underclassman (freshman or sophomore), or upperclassman (junior or above) and see how many credit hours they are signed up for, or how many times they touch a mirror during the day[4]; or if they are male or female and how many pairs of shoes they own. In any case, two different groups of people, measuring the same thing in both groups. Store the data like the ad campaign *advertising.csv* and read it into `R`. Then create the per-group comparison boxplot and do the classical t-test. Also compute the Bayesian probabilities (using the p-value approximation). Show the data, the plot, and the results from the test.
(8) Keep this data at hand. It might become the basis of your book project, described in the Preface.

[4]This scenario came from a homework I gave, where a fraternity member asked males and females how many times they touched a mirror. The student was surprised to find he was the only one with this particular proclivity.

CHAPTER 11

More Testing

1. Proportions

Here is another experiment:

- You stand, at a specific time, at a certain intersection and note who comes by a fixed point, either a male or female. You mark down whether these people have some sort of device that helps prevent them from thinking or being alone with their thoughts; namely, earphones from an iPod, cell-phone, or similar device. It is your hypothesis that just as many males being affixed to such a device as females.

These kinds of experiments are easier to understand than the previous one, because there's only one thing to think about. So let's call it a "success"[1] if a person is attached to a thinking suppression device, or TSD. We will obviously know, after you've taken your survey, for each sex, the number of successes and failures. But before you start your survey you did not know how many men and women would wear a TSD. Since you did not know, you chose to quantify your uncertainty with a probability distribution. Does this situation remind you of any particular probability distribution?—if it doesn't by now, you are in deep kimchi—for it should bring to mind the binomial distribution. Which we all remember has only one unknown parameter, θ, which is the success parameter. In this experiment, we have two of these parameters, one for males (M) and one for females (F).

Your hypothesis is that the probability of success is the same for both males and females, or that $\theta_M = \theta_F$. Suppose, as you sat at your corner, you saw $k_{\text{old, M}} = 14$ out of $n_{\text{old, M}} = 20$, and $k_{\text{old, F}} = 12$ out of $n_{\text{old, F}} = 15$. What is the probability that the mean (or rate) of males is the same as the mean as females? You know by now that this is not a trick question. The mean of males is $14/20 = 0.7 = \widehat{\theta}_M$ and the mean of females is $12/15 = 0.8 = \widehat{\theta}_F$. You can state authoritatively that the probability these means are equal is 0, that is, it is false that they are equal. Are we done?

Maybe. If all you were interested in where those $20 + 15 = 35$ people, then you are done. You can say with certainty that a greater proportion of women wore a TSD. There is really nothing more to say. Pause and reflect on this statement: it is a key point in this book.

[1] For Steve Jobs, anyway.

You would not be finished if you were curious about the proportion of women and men who would wear a TSD the next day, or on any other day—where the future men and women who walk by would be the "same" as the old men and women in some sense. What does that mean? Suppose on the first day that a bus pulled up to your intersection and disgorged its passengers. The bus transported a group of college students from a big hockey game. It was these students you counted. Tomorrow, no hockey, and all you expect to see, say, are businessmen and women. Would this change the characteristics of the people who walk by, would they be the "same" as the college students, in the sense that the probability they wear a TSD is the same? Maybe. This is a tricky area. It might be that the people tomorrow *are* different from the people today, such that there is expected to be a difference in the probability they wear a TSD. But it might not. If you do not take additional steps to measure this difference, you will never be able to say. At the very least, the characteristics of the people you measure become part of your list of premises, you background evidence E. To be explicit, the proposition E at least contains this clause "The people we measure look like those who pass by this corner at the time and place we measured them." This is a topic that requires deep thought. See Chapter 14 for more details.

The main point is that the results are for quantifying uncertainty in *future* data. For instance, if tomorrow you expect to see 10 men and 10 women, how many of those men would wear a TSD, how many women? You don't know, are uncertain, etc. Do you expect more women than men? That is the real question, but we'll put that one off for a moment and first ask an indirect question about the success parameters. We'll do this in the classical and modern way.

The steps to test classically are the same (they are always the same) as they were when your uncertainty was represented by a normal distribution. These steps are: First, assume that your uncertainty in the data is quantified by a binomial distribution for both sexes, and that $\theta_M = \theta_F$ (what is this odd proposition called?). Next, a statistic is calculated. For shorthand, let $\widehat{\theta}_M = x_{\text{old, M}}/n_{\text{old, M}}$, and similarly for females. Then one of the most-used classical statistics is

$$z(x) = \frac{\widehat{\theta}_M - \widehat{\theta}_F}{\sqrt{\frac{\widehat{\theta}_M(1-\widehat{\theta}_M)}{n_M} + \frac{\widehat{\theta}_F(1-\widehat{\theta}_F)}{n_F}}} \tag{28}$$

which should remind you quite a bit of the t-statistic. It's a difference in the observed means, dividing by a function of the observed standard deviations. What's next? Yes, the p-value, which here is

$$\text{p-value} = \Pr(Z(x) > |z(x)| \mid x, \theta_M = \theta_F, \text{E}_B)$$

Don't forget that E_B tacitly includes the evidence that future populations will "look like" your previous population.

To get this in `R`

```
prop.test(c(k_Males, k_Females), c(n_Males, n_Females))
```

It's a little awkward because you have to type in four numbers, two total successes and two "n"s, and you have to remember to concatenate the numerators and denominators with the `c` function. The output looks like this

```
        2-sample test for equality of proportions
                          with continuity correction

data:  c(14, 12) out of c(20, 15)
X-squared = 0.0779, df = 1, p-value = 0.7802
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.4434840  0.2434840
sample estimates:
prop 1 prop 2
   0.7    0.8
```

You'll immediately see, if you are observant, that there is something called a `X-squared`, which, mathematically, is a χ^2 (pronounced *chi*-squared) which is a *different* statistic than the $z(x)$ shown above. Well, close enough. Remember that there are always many, many choices of statistics you can use in any given problem. I showed you the $z(x)$ statistic, and `R` gave you a χ^2, but it could have given you "Fisher's exact", or a slew of others. The idea behind them is the same.

There are other peculiarities to notice. The classical "null" hypothesis is $\theta_M = \theta_F$, which means the "alternate" is `two.sided`, or $\theta_M \neq \theta_F$. There is, like in the `t.test`, a `df`, which is still "degrees of freedom", and is necessary to compute the mathematical χ^2 distribution. The proportions, or `prop`s, are there, and again you are left to compute their difference by hand $0.7 - 0.8 = -0.1$. Ignore the confidence interval for now.

The p-value is a disappointing 0.14. Which means what? You *fail* to reject the "null". You do *not* accept it. What is the probability that $\theta_M < \theta_F$? We cannot answer this question classically.

One last thing. There are some words in the output saying something about a "`continuity correction`". What is that about? Remember that we (the computer, actually) chose the χ^2 statistic because we knew how to mathematically compute its probability distribution *given* $\theta_M = \theta_F$. It turns out that we don't know how to exactly compute the χ^2 distribution unless the sample is "large" (meaning going towards infinity). What we can calculate, when the sample size is not infinite, is an approximation to this distribution. To make the approximation better, the data you observed is inflated by a small amount. This doesn't change the observed means (which are proportions here), but it does effect the p-value. Since you are stuck with the sample size you have, there isn't much you can do about this (except to choose another, different statistic, which will give you a different, possibly

lower and therefore more appealing, p-value, but that would be cheating, wouldn't it?).

2. Power & Real Significance

Up to this point, we have been talking about using the hackneyed 0.05 for our criterion of classical significance. If the p-value of a classical test is less than this number the results are said to be "statistically significant." That 0.05 number is called the *level* of the classical test. We have also been computing 95% confidence intervals, and it is no coincidence that the two numbers are related (100% - 5% = 95%, right?). It turns out that you can mathematically show that the decisions you make using classical confidence intervals and or classical tests would be the same. For example, when your uncertainty about two groups A and B is quantified by normal distributions, and if $\Pr(T(x) > |t(x)| \ |\mu_A = \mu_B, \mathrm{E}_N)$ is less than 0.05, this is the *same* as saying that the 95% confidence interval of the estimate for $\mu_A - \mu_B$ does not contain 0. It is not really necessary to remember this, as the software you use will usually give you both numbers. However, as I stress repeatedly, if the p-value is 0.05 it does *not* mean that $\mu_A \neq \mu_B$. You might decide to act as if $\mu_A \neq \mu_B$, but there is no way to know classically, given the data you observed, the probability this decision is the right one.

The modern version of statistics directly tells you, given the old data and some evidence E, the probability that $\mu_A < \mu_B$, or $\mu_A > \mu_B$, or $\mu_A < 2\mu_B$, or whatever function $f(\mu_A, \mu_B)$ that might be of interest to you. Modern methods gives you this direct probability and *you* make the decision whether this probability is high enough or low enough to make and act on the decision, say, that $\mu_A < 2\mu_B$ (this still does not directly tell you anything about future observable data).

Let's splay this open a little more. Suppose that you have not yet taken any data. What is the probability that $\mu_A < \mu_B$ given E_N (recall the E_N has tacit information on the prior distributions of (μ, σ) for groups A and B)? If you have no information except for E_N, then it should be intuitively obvious that it is just as likely that $\mu_A < \mu_B$ or $\mu_A > \mu_B$. Do not read more until you understand this.

Incidentally, classical statistics can say nothing in this case. Some data has to be taken before you can say anything about the parameters. As always, classical statistics cannot make a probability statement about $\mu_A < \mu_B$ or $\mu_A > \mu_B$ or any other question you might have about the parameters.

Now take one data point. In modern statistics, you can calculate the probability, given *this* data point and E_N, of $\mu_A < \mu_B$. It will probably be the case that the probability that $\mu_A < \mu_B$ is not much different than 0.5. After all, you have only seen one additional piece of evidence (additional to E_N, of course), so you cannot expect the probabilities to change too much.

Classical statistics can still say nothing in this case. It can't really say anything until at least four different data points are take, assuming at least

two observations come from each group. Well, that's hard luck for classical statistics, but the idea behind this is similar in spirit—enough data has to be taken to make some kind of sense of the guesses we make.

If you only take one or two pieces of data, or only a few, it should be intuitive to you that you learn nothing or little about the parameters given this data. If you take a lot of data, you learn a lot about the parameters. This is the idea of *power*. More data means more certainty in the guesses you make, it implies that your posterior probability of the parameters is very tight, meaning that a lot of probability is given to a very narrow interval, meaning you are more certain what the actual values are.

In modern statistics, power is built right in. The posterior distribution of the parameters tells you how certain you should be about their values. Less data means wider, more spread out probability, less certainty about the parameter values; more data means narrow, tighter probability, more certainty about the parameter values.

Classical statistics does not have this advantage. Recall that the p-value is used to indirectly infer whether the so-called null hypothesis is true or false, usually phrased as $\mu_A = \mu_B$ (for uncertainty quantified by normals) and $\theta_A = \theta_B$ (for uncertainty quantified by binomials). If the p-value is less than the magic number, then you announce "μ_A does *not* equal μ_B" or "θ_A does *not* equal θ_B" even though, obviously, you cannot be certain that these statements are true, you just act like they are. If the p-value is big, then you say "I fail to reject the idea that μ_A does *not* equal μ_B" or "I fail to reject the idea that θ_A does *not* equal θ_B", which is just confusing. What is the probability that you are right or wrong in both cases given the data you just observed? Can't say. But you can say what this probability of being right and wrong would be *if* you were to repeat your experiment an infinite number of times. If you say that $\mu_A \neq \mu_B$ then if $\mu_A = \mu_B$, you will be incorrect 5% of the time (in this infinite series), which is the test level. If you fail to say $\mu_A \neq \mu_B$, then if $\mu_A \neq \mu_B$ you will be incorrect some about of time that we can call β (by tradition). Then $1 - \beta$ is called the *power* of the classical (infinite series of) tests. Confused? Don't worry, most people are, and most people never remember, which is why so many people misinterpret the p-value and classical power.

In modern statistics, you say "Given just the data I observed, and my evidence E, the probability that $\mu_A < \mu_B$ is this-and-such" or "Given just the data I observed, and my evidence E, the probability that $\theta_A < \theta_B$ is this-and-such." We can replace these inequalities by any function of the parameters we want, say, $\mu_A < 2\mu_B$. You as the statistician then give your customer that actual probability of whatever hypothesis is interesting and let him decide whether this probability is high enough or low enough to make the decision that $\mu_A < \mu_B$ or $\theta_A < \theta_B$. This is a tremendous advantage over the old way of making the judgment for him.

3. Back to observables: normal

Eventually, if you take enough data, you'll be certain enough about the parameter values; that is, you'll be so sure what the values of the parameters are that it will cause no grief for you to say you know their values exactly (technically, the proposition "The central parameter μ equals 28.2" is a contingent statement, which based on our evidence can never have probability 0 or 1, but it can have probability $1-\epsilon$, where ϵ is a number as small as you want as long as it is greater than 0).

Suppose Uncle Ted ran his ad campaigns A and B so many times that the posterior probabilities of the central parameters are nearly certain to be just one value (for each campaign). For example, $\mu_A \approx m_A$ and $\mu_B \approx m_B$, where we switch to Latin letters to indicate that we know the values precisely. Now let $m_A = 420$ and $m_B = 420.01$ (recall this is the central parameter for the number of sandwiches sold in a week). What is the probability that $m_A = m_B$? Again, this is not a trick question. It is no different than asking, what is the probability that $7 = 104$? The probability is 0 in both cases. The values are certainly not equal. Congratulations! You have just *proven* that, given our evidence E_N, campaign B is better than A. There is no ambiguity: B is better than A and that is that.

But is m_A enough smaller than m_B to make any difference in sausage sales? Well, yes. As long as $s_A = s_B$, if $m_B > m_A$, then campaign B is better because (we recall from Chapter 4) it has a higher probability of larger numbers of sausage sales. In reality, however, the difference is miniscule and not in the least interesting. The difference between the two campaigns is real, it is certain, and you could announce to the world that you have proved a "statistically significant" difference, and that giving out bullets is better than offering free antler sets. You can then go further and say that this implies "Americans are becoming more bloodthirsty as a new study shows that they prefer bullets over antlers." Well, enough of that for now. See Chapter 14 for more.

Why so much interest in the parameters? The *real* question is: Given the difference in the parameters, how different are actual measurable observable sausage sales? Because we *know* the exact values of the parameters *does not mean that we know the values of the observables*. Obviously, we do not. If we knew the values of the observables, we would not have needed probability to begin with!

Let's work through an example. Suppose, after a long period of time, we conclude that $\mu_A \approx m_A = 420$ and $\mu_B \approx m_B = 421$, and similarly for the spread parameters, say $\sigma_A = \sigma_B = s_A = s_B = 46$. What is the probability, given our data and E_N, that $\mu_A = \mu_B$? It is 0 precisely: we are certain that the parameters are different. But what is the probability that next week's sausage sales under campaign A are less than the sales under campaign B? It is *not* 0. In fact (calculations show) it is **only 50.4%**! This result should

shock you. If it does not, then go back to the beginning of this Section and re-read it. (We will learn how to do this calculation in a couple of Chapters.)

This is a key point! The measly difference in central parameters which we knew with absolute certainty, i.e. saying $m_A < m_B$ with boastful confidence, only translated into very little certainty that the $\text{Sales}_A < \text{Sales}_B$.

What if the difference in central parameters was an astounding 100, that is $m_A = 420$ and $m_B = 520$, then what is the probability of the sales in A being less than B? Only 86%![2] That means there is still a respectable 14% chance that sales in A will beat sales in B. Differences, even huge differences, in parameters do not necessarily translate into great confidence that future observables will be different.

It's even worse than it seems. For we rarely or *never really do know the exact values of the parameters*. The best we can do is to present the posterior distributions of the parameters. Thus, instead of stating that $m_A < m_B$ is certainly true, the best we can say $\mu_A < \mu_B$ has a only a chance at being true. Since this is the case, we have to carry this uncertainty through to the uncertainty we have in the future observables.

Using the actual data (last Chapter), we estimated that the probability that $\mu_A < \mu_B$ was about 0.9. The software also lets us estimate the posterior distributions of σ_A and σ_B. But what is the probability that next week's sales under campaign A is less than sales under campaign B *given* the past data and E_N? Only **60%**! (We'll learn how to do these calculations later.)

Think about this. If you knew *nothing* (collected no data) about this situation except that there were two campaigns and that one of them will do better than the other, then the probability, under this evidence, of $\text{Sales}_A < \text{Sales}_B$ is 50%. Adding the evidence of 20 weeks of actual data only improved the sharpness of this guess marginally: that is, we moved from 50% certainty to only 60% certainty. We were 90% certain there was a difference in central parameters, but only 60% sure that there would be a difference in actual observables.

To emphasize again: *certainty in the unobservable parameters does not directly translate to certainty in the observable data.* This is important to imprint on as many neurons as possible because nearly all of statistics, classical and modern, states results about certainty or uncertainty in *parameters*. You can be as sure of the values of parameters as you like, but this does not mean that you are as sure of reality.

We will certainly come back to these topics.

4. Back to observables: binomial

The same story of the difference between parameters and observables repeats itself for the binomial success parameters. Thus, let us say, in the TSD experiment, that we have enough data so we can confidently say $\theta_M \approx$

[2]You should be gasping, particularly if you have ever practiced classical statistics before.

$p_M = 0.7$ and $\theta_F \approx p_F = 0.8$. What is the probability that $p_M < p_F$? Yes, it is 1, they are *certainly* different.

Suppose, then, you set out the next day with p_M and p_F in hand and a female approaches. What is the probability that she wears a TSD? Is it 100% because we can say that there is a 100% chance that $p_M < p_F$? Obviously not! It is 80%.

Now suppose that the next day we will see 10 men and 10 women. Given our prior observations and E_B, what is the probability that more women than men in that future group of 20 will wear a TSD? Is it 100% because it is 100% certain that $p_M < p_F$? No! It is (calculations show) **only 60%**!. And there is an 18% chance the number of women wearing a TSD will equal the number of men wearing one, leaving, of course, a 22% chance that more men than women wear TSDs. We'll learn how to do these calculations later.

Again, let's imagine that we have taken no data. What would we guess the probability of more women than men wearing a TSD to be (given E_B) ? 50%. So even if we knew the *exact* values of the parameters, we only improve our knowledge to a 60% chance. Not a very large move.

Wait, it's still worse (in the sense of less sure) than this. We are *not* certain that $\theta_M < \theta_F$, and we have to take into account our uncertainty in these measures when we consider the observables. Doing that (via calculations we will learn) says the probability of more women than men has only a 57% chance (there is a 16% chance of an equal number of men and women; and a 27% chance men outnumber women). Again, just like when we quantified our uncertainty with normal distributions, if we started off knowing nothing except that we would have two groups and that one of these groups would be larger than the other, the probability of seeing more women than men affixed to TSDs would be 50%. The data marginally sharpened this guess to 57%. Even if we absolutely knew the values of the parameters, the probability is just 60% that more women than men are hooked to the TSD. And that is as sure as we can *ever* be.

5. Homework

(1) Perform the thinking suppression device survey. Use your own real-life data and measure at least two different groups (men and women, those over 40 and those under, etc.). Then do the classical z-test (or χ^2) test. What is the approximate probability that one success parameter is less than the other?
(2) Keep this data at hand. It might become the basis of your book project, described in the Preface.

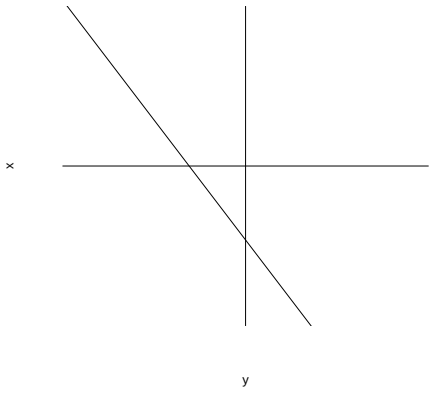
CHAPTER 12

Regression Modelling

1. Uncle Ted

Here is a simple picture and a simple formula, both of which, I hope to God, you remember from high school.

$$y = a + bx \tag{29}$$



Yes, $y = a + bx$ is the equation for a *straight line*, the simplest geometric figure after the dot. It explains what happens to y when we change x: a is called the intercept (it is the point, when $x = 0$, that intercepts the y axis), and b is the slope. If $b > 0$ the slope of the line is positive and y increases as x does. Another way to say that is that y is (linearly, positively) proportional to x, or $y \propto x$, or y is positively "correlated" with x. If $b < 0$ the slope of the line is negative and y decreases as x increases, or y is proportional to $-x$, or y is negatively correlated with x.

Suppose we have some observable data x and we want it to help us predict the observable data y. The idea behind statistical regression is to *model* the relationship between y and x as some kind of probabilistic straight line. Sound complicated? It isn't; in fact, you already know how to do it.

Let's recall Uncle Ted's advertising campaign experiment. Two campaigns, A and B, ran for 20 weeks and sales of sausage sandwiches were tallied. I want you now to take out a piece of paper, a mere scrap will do. Draw a horizontal line and under it write an A and to the right of that write

a B. I am not going to do this for you, so get the paper and get to work. Now, to the left of the horizontal line draw a vertical one, and label it (to the left) Sales. The horizontal line is your x-axis and your vertical the y-axis. Place a dot over the spot marked A at the point 421, and another dot over the B at the point 440. There is no need to be meticulous about this. As long as the point for B is a little higher than the point for A, then you are in business. The picture you have drawn is finished. What it says is that, for campaign A, sales are 421, and for campaign B, sales are 440. Believe it or not, that is a version of a straight line. Here is the equation for it.

$$\text{Sales} = 421 + 19 \times \text{I(Campaign B)} \tag{30}$$

The only thing that is new is the bit about I(Campaign B). The I() is called an *indicator function*, and it is equal to 1 when its argument is true, else it is equal to 0. No, no; it's not difficult. In this case, when we are considering campaign A, then Sales $= 421 + 19 \times$ I(Campaign B) $= 421 + 19 \times 0 = 421$ because we are *not* in Campaign B, therefore the indicator is 0. But if we are in Campaign B, then Sales $= 421 + 19 \times$ I(Campaign B) $= 421 + 19 \times 1 = 440$. Simple.

The thing to notice is that the "19" is the *difference* in sales between the two campaigns (B minus A). The 421 is still called the "intercept", out of habit. If you like, the 19 is a modification of the intercept for campaign B.

Does equation (30) perfectly represent the data we saw over the 20 weeks? Obviously not. The sales at A were not always exactly 421 and at B they were not always $421 + 19 = 440$. Sometimes sales at A were a little higher than 421, sometimes a little lower. Draw in some more dots above A to indicate this: some dots higher than 421 and some lower. Then do the same for B: some dots higher than 440, some lower. This picture *does* realistically represent the sales. Now go back to Chapter 10 and look at the boxplot we drew (`boxplot(Sales~Campaign)`) and compare it to the one you just sketched. Surprise. The only thing we have to do is to modify equation (30) so that it represents the fact that the sales were not always constantly 421 and 440. Here's how to do it:

$$\text{Sales}_i = 421 + 19 \times \text{I(Campaign B}_i) + \epsilon_i. \tag{31}$$

Two things have changed. The easy one first: a subscript i has been added to Sales to indicate the ith data point, where we remember that $i = 1, 2, \ldots, 40$ (20 weeks over two cities; look at the CSV file to convince yourself). The hard thing is ϵ_i, a Greek letter, and what do Greek letters mean? That's right, you remembered! It means ϵ_i is *an unobservable parameter*, and that there is a different one for each data point, which is why it has the subscript i. It represents the departure of each observation from the point 421 (if we

are in A; 440 if we are in B)[1]. This is easy to see. Suppose we are in campaign A, then $\text{Sales}_i - 421 = \epsilon_i$.

This parameter is a little screwy, in that it is *almost* observable. If we *knew* that the sales in A were always 421 plus some "departure", then each ϵ_i would be exactly observable, and each would equal $\text{Sales}_i - 421 = \epsilon_i$ (for campaign A; similarly for B). But, we don't know that the sales in A are always 421, do we? In Chapter 10, we quantified our uncertainty in sales using a normal distribution, which had two parameters for each campaign, e.g. μ_A and σ_A. We still don't know what future sales under the different campaigns will be, and so we still have to quantify their uncertainty using a probability distribution. Well, we can suppose that sales under each campaign are some central value plus or minus some other value that can change from week to week. Let's rewrite equation (31) for campaign A with this in mind:

$$\text{Sales}_i = \mu_A + \epsilon_i, \tag{32}$$

where μ_A is the central parameter and ϵ_i the unknown plus or minus difference from that central value. What can we logically say about ϵ_i? We don't know each value, but we can suppose that we could see just as many positive values of ϵ_i as negative; that is, the departures from the central value for each observation are equally likely to be greater than 0 or less than 0. Now, because ϵ_i is semi-observable, it turns out we can semi-describe our uncertainty in it using probability in classical statistics. We have already deduced (or mandated) that whatever probability distribution we use must be symmetric about 0: points below 0 are equally likely as points above. Well, many distributions meet this criterion, so we still have to choose among these. The most popular choice is—can you guess?—the normal distribution. This means we describe our uncertainty in ϵ_i with a normal distribution with (a deduced) central parameter 0 and a spread parameter as σ^2. This means that Sales (at A) is some number μ_A plus some number ϵ, thus our uncertainty in Sales at A is quantified by a normal distribution with parameters μ_A and σ^2. In other words, it is just as it was in Chapter 10.

What about campaign B? In Chapter 10, we quantified our uncertainty in future values with a normal distribution with parameters μ_B and σ_B. Let's rewrite equation (31) for campaign B with this in mind (the indicator function equals 1):

$$\text{Sales}_i = \mu_A + ? \times 1 + \epsilon_i,$$

where we have to solve for "?" and where we still need to fix it so that when I(Campaign B) = 0 the equation still works for campaign A. We want, when

[1] ϵ is sometimes called *random error*. What do people mean by *error*? Do they mean the sales would have been 421 in every case had not something gone horribly wrong? Actually, to call it error is to idealize the parameter μ_A as *the* Platonic, incorruptible Sales Of A, which again proves that there is too much attention paid to parameters. In some cases you are interested in the measurement error of some apparatus, and calling ϵ "error" makes sense; else, it is just silly.

we are in campaign B, the right hand side to equal $\mu_B + \epsilon$, so what does "?" equal?

$$\text{Sales}_i = \mu_A + (\mu_B - \mu_A) + \epsilon_i = \mu_B + \epsilon_i. \tag{33}$$

We can write $(\mu_B - \mu_A)$ as μ_{B-A}. This parameter is still the difference in central parameters for the two campaigns, just like we did using the classical t-test in Chapter 10. In other words, we went to all the trouble of re-writing what we already knew how to do, so that it fit into the framework of what is called *regression*. Before we learn how to do this practically on a computer, let's learn a little more about regression and see why it's a more general framework than the old testing procedures we learned.

2. White blood

We will use the appendicitis data set, so it might help to re-read Chapter 7, and read the data into `R` to get ready.

One of the variables in that data set is `White.Blood.Count`, and since we used an English name for this variable, we might guess it means white blood count (it does). We have two groups of people, those who have appendicitis and those who do not: in the data set `Appendicitis` this is coded as `N` or `Y`. White blood cells are used to fight off infection, so it is guessed that in patients with appendicitis, which is an infection, white blood count would be higher than in patients who did not have appendicitis.

Do we know, for future patients with right lower quadrant pain, the value of their white blood count? We do not, we are uncertain. And how do we quantify our uncertainty? Using probability. What distribution best represents our uncertainty in white blood count? Well, it's not the binomial, so we'll go with the normal. Let's write out our uncertainty in the form of a regression model

$$\text{White.Blood.Count}_i = \mu_N + \mu_{Y-N} \times \text{I}(\text{Appendicitis} = \text{Y}_i) + \epsilon_i, \tag{34}$$

where μ_N is the central parameter for patients who do not have appendicitis and $\mu_{Y-N} = \mu_Y - \mu_N$ is the difference in central parameters for those with and without appendicitis. Except for swapping A with N and B with Y, this equation is no different than the one for sales in difference campaigns. Now let's add a twist.

Older people might produce less white blood than younger people, regardless whether they have appendicitis or not. After all, they're old. Running the `summary(Age)` shows we have people as young as 3 and as old as 93 in our dataset. If mostly young people have appendicitis and old people do not, then it could look like there was a difference in white blood count because of appendicitis just because the ages were skewed in our dataset. We want to *control* for age so that this is not a problem. For shorthand, write $\text{I}(\text{Appendicitis} = \text{Y}) = \text{I}(\text{Y})$. Then the way the model is modified is

$$\text{White.Blood.Count}_i = \mu_N + \mu_{Y-N}\text{I}(\text{Y}_i) + \beta\text{Age}_i + \epsilon_i. \tag{35}$$

Take your time with this. Suppose that $\text{Age}_i = 0$ for some person (yes, it is impossible). Then μ_A is the central parameter for a normal distribution which describes our uncertainty in people who do not have appendicitis and who are 0 years old. Then $\mu_N + \mu_{Y-N}$ is the same thing but for people with appendicitis. Now let $\text{Age}_i = 1$. Then $\mu_A + \beta$ is the central parameter for a normal distribution which describes our uncertainty in people who do not have appendicitis and who are 1 year old. If $\text{Age}_i = 40$, then $\mu_A + 40\beta$ is the central parameter for a normal distribution which describes our uncertainty in people who do not have appendicitis and who are 40 years old. Do you see? In order to express our uncertainty in white blood count we have to specify two things: (1) whether or not the person has appendicitis, and (2) their age. If we do not have both of these, we cannot express our uncertainty. You have to plug these values into the equation, else it is meaningless. Remember this key point, it is often forgotten.

Because the central parameter is $\mu_A + \beta\text{Age}$ (for people without appendicitis), we have controlled for age, because we allow the uncertainty in white blood count to change with age. It changes in a *linear* fashion, too. To see that, suppose we look at people without appendicitis

$$\text{White.Blood.Count}_i = \mu_N + \beta\text{Age}_i + \epsilon_i. \tag{36}$$

Doesn't that remind you of the equation for a straight line? Except for the ϵ, it *is* that equation. Thus, we can say that we have *linearly modeled* white blood count, controlling for appendicitis and age.

Suppose we also wanted to control for each patient's weight? (We don't have this variable in the dataset, so we'll just suppose.) Then

$$\text{White.Blood.Count}_i = \mu_N + \mu_{Y-N}\text{I}(\text{Y}_i) + \beta_a\text{Age}_i + \beta_w\text{Weight}_i + \epsilon_i,$$

where I have modified the subscripts on the "βs" to indicate age and weight. Incidentally, these Greek letters are also called *coefficients* (of the regression line) or just "βs" (the appendicitis indicator, Age, etc. are *independent variables*—don't ask why). I will call the Greek letters either parameters of coefficients. What does μ_N mean in this case? Well, set Age and Weight equal to 0 for people without appendicitis, and μ_A is the central parameter of the normal distribution describing our uncertainty in white blood count for patients who have these characteristics (no appendicitis, age and weight equal to 0).

Stop! Right now, you should be thinking, "This guy is nuts. I can buy an age equalling 0 because maybe that means babies less than a year old. But weight equalling 0? Nonsense!" I agree with you. It *is* nonsense. What it means is that these kinds of models, used everywhere, are limited in scope, and not always applicable (yet they are still used). We'll talk more about this later.

What if we further wanted to control for, say, blood pressure? Right, just add a $\beta_{sbp}SBP$ for systolic blood pressure and a $\beta_{dbp}DBP$ for diastolic

blood pressure. And so on for as many variables as you think need to be controlled for.

Almost done. Let's ignore Age and Weight and suppose we instead wanted to control for `Sex`, because men and women might naturally (appendicitis or not) have different levels of white blood count. Sex is a categorical variable, with two levels, M and F (actually, in the database they are spelled out, but we'll use the shorthand here). How is this added? Like this:

$$\text{White.Blood.Count}_i = \mu_N + \mu_{Y-N}\text{I}(\text{Y}_i) + \beta_s\text{I}(\text{M}_i) + \epsilon_i,$$

where we need the indicator function again, to tell us if the patient is M or F. What does μ_N mean now? Well, we have to say whether or not the person has appendicitis and what sex that person is. We always have to pick values for the variables we put in the model! Always, always, always! Pick N and F. Then μ_N is the central parameter for a normal distribution which describes our uncertainty in people who do not have appendicitis and who are female. Technically, it would be better if we wrote it as μ_{NF}, but this notation isn't usual. What if the patient is a male? Then $\text{I}(\text{M}) = 1$ and so $\mu_N + \beta_s$ is the central parameter for a normal distribution which describes our uncertainty in people who do not have appendicitis and who are male. If they have appendicitis and are male, then $\mu_N + \mu_{Y-N} + \beta_s$ is the central parameter for a normal distribution which describes our uncertainty in people who have appendicitis and who are male. Get it? If not, stay here until you do.

Almost done (this time I mean it). Ok, ignore sex and add back age, which we already know how to do. This time, let's imagine that, indeed, older age means less white blood, as before, but now imagine that the rate at which white blood drops off is different for those with appendicitis and those without. Say that those with appendicitis do naturally have less white blood when they age, but because they have appendicitis they have more than older people without appendicitis. This means that we would like one straight line to indicate the relationship of age and white blood for those with appendicitis and another line for those without. Here's how:

$$\text{White.Blood.Count}_i = \mu_N + \mu_{Y-N}\text{I}(\text{Y}_i) + \mu_a\text{I}(\text{Y}_i)\text{Age}_i + \beta_a\text{Age}_i + \epsilon_i,$$

which looks pretty complicated, but do not despair! When confronted with a beast like this, take it one step at a time. Remember that we *always* have to supply a value for each of the variables we put into the model. Let's do that and see what we get. Pick a person without appendicitis, so that $\text{I}(\text{Y}_i) = 0$ and with age equal to 0. Then all we are left with is μ_N, which must be the central parameter for the normal distribution which describes our uncertainty in white blood count for people without appendicitis and who are 0 years old.

Let age be 20. Then we have $\mu_N + 20\beta_a$, which, to be verbose, is the central parameter for the normal distribution which describes our uncertainty in white blood count for people without appendicitis and who are 20 years

old. The β_a is the "b" of the regression line, describing the rate of change of the central parameter with age.

Incidentally, since you might have forgotten, and since all normal distributions have two parameters, the other parameter besides the central is the spread σ^2, which is assumed to be always the same regardless of the values we pick for the variables, and all levels of those variables. Not always a great assumption, incidentally, but it will have to be good enough for this book.

Now let the person have appendicitis and an age of 0. All that is left, after multiplying by Age $= 0$, is $\mu_A + \mu_{Y-N}$, which is the central parameter etc. If you like, $\mu_A + \mu_{Y-N}$ is the "a" of the regression model, a different intercept for people who have appendicitis. Age can now be 20, what do we get? $(\mu_N + \mu_{Y-N}) + 20\mu_a + 20\beta_a$. That looks a little confusing, so let's rewrite the model for people with and without appendicitis.

$$\mu_N + \beta_a \text{Age}$$

or

$$(\mu_N + \mu_{Y-N}) + (\mu_a + \beta_a)\text{Age}$$

where these are written in the form of straight lines. They are, of course, the central parameters of the normal distribution describing our uncertainty in white blood count, given the patient has or does not have appendicitis and for a known age.

That's it, you have just got a graduate education in writing linear regression models (for a truly graduate education, see Bernardo and Smith, 2000). No kidding (most other courses spend all the time learning how to calculate various guesses of the βs by hand and so scrimp on understanding the models actually are). All other regression models are like this. The only thing that changes is the names of the variables and how many variables you put on the right hand side.

Never forgot what you are doing, however. You are trying to describe your uncertainty in an observable (the left hand side) like white blood count. You do this by assuming that the uncertainty in this observable is quantified by a normal distribution with spread parameter σ^2 and with a central parameter that depends on *specific* values of the variables that are on the right hand side. You are saying you know the values of variables on the right hand side, or you are assuming that you know. But you obviously do not know what the value of the observable on the left hand side will be for new data.

3. Practicals

We want to quantify our uncertainty in white blood count controlling for appendicitis and age, and we assume that the relationship of white blood and age is different for those with and without appendicitis. This implies, as we have seen, the model

$$\text{White.Blood.Count}_i = \mu_N + \mu_{Y-N}\text{I}(\text{Y}_i) + \mu_a\text{I}(\text{Y}_i)\text{Age}_i + \beta_a\text{Age}_i + \epsilon_i.$$

What is the value of μ_N? Of β_a etc.? You don't know.

Here is Uncle Ted's problem again

$$\text{Sales}_i = \mu_A + \mu_{B-A} \times \text{I}(\text{Campaign B}_i) + \epsilon_i.$$

What is the value of μ_N and μ_{B-A}? Again, you don't know, you are not certain. This means we have to find a way to guess the values of these parameters, and then quantify our uncertainty in these guesses. And we have to do this in the classical and modern way, never forgetting that our goal is to learn something about the observable itself and not just the parameters.

In order to build a linear regression model, you'd like to have the observable and variables at least look like there is some kind of straight line relationship between the two, right? So *always* start with a picture! This is because of the well known wisdom: *never assume*. There are at least 1432 ways to plot data in `R`. Here are two. Assuming you have read Uncle Ted's data into `R` and called it `x` (see Chapter 7), then then all you have to do is

```
plot(x)
```

which will give you the boxplots we have seen before. `R` is smart enough to know that if you only have an observable (like Sales) and a categorical variable (like Campaign), to automatically give you a boxplot. You could also directly get the boxplot in the way you learned before.

Let's also read in the appendicitis data and called it `d` (for, I suppose, "data"). You can try `plot(d)`, but it's a busy figure. It is a plot of each variable in the dataset by each other variable. The way to read these is simple. In each row, the variable that is named takes the y-axis, and in each column, the variable that is named takes the x-axis. These "scatterplot matrices" are sometimes just the thing, but not in this case because we have so many categorical variables. For example, the first plot on the upper left is `Belly.Button.Pain` by `Vomiting`. Both of these variables were coded as 0/1 (no/yes). The plot, then, is just a dot for every possible combination of `Belly.Button.Pain` and `Vomiting` seen in the database (which are 0/0, 0/1, 1/0, 1/1). Not very interesting to look at. The good stuff is `White.Blood.Count` by `Age` or `Temperature`, etc. To get those separately—first `attach(d)`—just ask for them, i.e. `plot(Age, White.Blood.Count)`, etc. Especially do that one (by Age). Does it look like a straight line could be run through these data points? (I'd give you the picture in the book, but then you would be tempted to not produce it yourself, and we can't have that.) Not really, but it doesn't *not* look like it either. By that I mean, it doesn't look like any other kind of line would fit these points. What we can take from that is Age is probably not that useful in helping describe our uncertainty in white blood. Also try `plot(Appendicitis, White.Blood.Count)`. A boxplot, since `R` realizes `Appendicitis` is categorical. Looks like higher white blood counts for people with appendicitis, just as we suspected. (You're really going to have to start up `R` to follow along here. Since this is the homework, you might as well.)

We could go on and on and on, producing new and better plots of the data and do a tremendous job of exploratory analysis, which is one of the fairest things you can do with data, but the techniques to learn are enormous in number. The datasets you will play with in this book are also "clean", by which I mean they will roughly match linear regression assumptions; that is, you won't see any screwy data here, nor learn how to deal with it. All these things are very important, but this book is about understanding the results of analyses, which is much more important because most of you will never or rarely do your own analyses. You will almost certainly, however, be confronted with the analyses of others, and thus it is crucial that you be able to comprehend the claims that are made, and how they are typically overstated or too certain. Sadly, then, we have to leave off here on the critical topic of exploratory data analysis.

Time to get estimates of the coefficients. We learned how to do this in Uncle Ted's case back in Chapter 10 using the `t.test` function in `R`. The more general approach, classically anyway, is the `glm` function (for "generalized linear model"). This is easy in `R`:

```
fit = glm(Sales ~ Campaign)
```

where the symbol "$\sim$" in computerese means equals; but since Sales doesn't really *equal* Campaign, this symbol really means, as it did in Chapter 4, "the uncertainty in Sales is quantified as a (linear) function of (Campaign)". To see the results, type

```
summary(fit)
```

Store the results of the model in the object `fit` so you can play with them easier later. How did I decide upon the name `fit` to store the results? The same way you decided on the name of your dog. It just doesn't matter what you call it. You'll get something this:

```
Call:
glm(formula = Sales ~ Campaign)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-110.75   -30.25   -13.75    31.75   100.25

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   420.75      10.11  41.604   <2e-16 ***
CampaignB      19.00      14.30   1.328    0.192
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for
    gaussian family taken to be 2045.566)

    Null deviance: 81342  on 39  degrees of freedom
```

```
Residual deviance: 77732  on 38  degrees of freedom
AIC: 422.4
```

There's a lot of detail here, so let's take our time to get it right. The first is `Call`, which merely echoes the same the code you typed to get the results. The next is `Residuals`, and you can safely ignore it. You can see that `R` calls the μs and βs etc. `Coefficients`, but it does not print out any Greek letters. How could it? The Greek letters we used were arbitrary anyway. Most software, not just `R`, calls μ_A, for no good reason in parentheses, the `(Intercept)`. It is up to *you* to remember the mathematical formula and that `(Intercept)` means μ_A.

Next up is `CampaignB`, note especially the `B`. This is `R`'s way of telling you that this is the coefficient for the indicator function where `Campaign` takes the value of `B`. `A` is, of course, missing (why?). If you ran a third campaign, `C`, then you would see an additional line called `CampaignC`. The only thing to keep in mind is that `R` assumes alphabetical order for the levels of categorical variables, and it sticks the lowest in the `(Intercept)`, which is why `CampaignA = (Intercept)`. Of course, `CampaignB` is *not* μ_B, it is μ_{B-A}. And if we had a campaign C, then `CampaignC` would represent μ_{C-A}, and so on.

The value of $\widehat{\mu}_A$ is under the `Estimate` column, and is equal to 420.75, just like in Chapter 10. $\widehat{\mu}_{B-A}$ is 19, again like in Chapter 10. The next two columns, `Std. Error`, `t value` can be ignored (except to note that this is the same t-statistic as before, now called a `t value`; we don't really care about the exact numerical value of this thing). These two columns' main purpose is to help calculate the fourth, `Pr(>|t|)`, which is our old pal, the p-value. What are p-values? They are the probability of seeing a statistic as large or larger than the one we got given some "null" hypothesis about the parameters is true. The statistic is the `t value`, so what is the "null" hypothesis? For the second row, it is that $\mu_{B-A} = \mu_B - \mu_A = 0$, or $\mu_A = \mu_B$, same as it was in Chapter 10.

What about the first row? It is that $\mu_A = 0$. It is *always* that the parameter of that row equals 0! So what does $\mu_A = 0$ mean? Well, that's the central parameter for the normal distribution describing our uncertainty in Sales under Campaign A. A central parameter of 0 means sales are as likely to be greater than 0 as they are less than 0. Yes, Sales less than 0. Does this make any sense? The answer is no. Given that we are talking about Sales, what is the probability that $\mu_A > 0$. You'd be tempted to say it is 1, and you'd be almost right, but we have here a problem first noted in Chapter 4: normal distributions should not be used to quantify uncertainty for real-life observables. In this case, it is impossible that Sales are less than 0, yet there is a non-zero probability that they would be if uncertainty in Sales is quantified by a normal distribution. The situation is already absurd, but let's, like everybody else, draw a veil, and silently carry on. Thus, given that $\mu_A = 0$, the probability of seeing a `t value` larger than 41.604 is $< 2e - 16$

(of course, given that we run the ad campaigns an infinite number of times). Fascinating? The answer is again no.

In case you weren't already enthralled by p-values, R helps you out by flagging with asterisks (the `Signif. codes:`) the publishable ones. Ignore the `Dispersion parameter` and `Deviance` lines for now.

On to appendicitis! Here's how to get our model:

`fit = glm(White.Blood.Count ~ Appendicitis*Age)`

where the only thing that is different is the `Appendicitis*Age` which mathematically translates to Appendicitis + Age + Appendicitis × Age. If we did not want the "interaction" term (Appendicitis × Age), then we would have put just Appendicitis + Age. The `summary`, stripped of all the extraneous matter, gives us

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.588673   0.443045  21.643  < 2e-16
AppendicitisY      4.933527   0.824903   5.981 4.79e-09
Age               -0.010289   0.010406  -0.989    0.323
AppendicitisY:Age -0.003703   0.022129  -0.167    0.867
```

This table looks more complicated, but follows the same rules as the previous one. The `(Intercept)` means μ_N, and $\widehat{\mu}_N = 9.6$; μ_{Y-N} is `AppendicitisY` (the indicator function) with classical estimate 4.9; β_a is `Age`; and μ_a is `AppendicitisY:Age`. The last column is the classical p-value for the "null" hypothesis that each of those parameters is equal to 0. The only trick is to recognize that the interaction term is written `AppendicitisY:Age`. Incidentally, you could have written the right hand side of the `glm` in R as `Appendicitis + Age + Appendicitis:Age` instead of `Appendicitis * Age`.

One last note about classical regression. To get the classical confidence interval on each of the estimates, type `confint(fit)`. What do confidence intervals mean?

4. Modern

In one sense, we are done, because it turns out that the classical confidence intervals can once again be interpreted as modern credible intervals. This means, after typing `confint(fit)`, we see

```
                      2.5 %      97.5 %
(Intercept)        8.72032036 10.45702473
AppendicitisY      3.31674790  6.55030629
Age               -0.03068433  0.01010720
AppendicitisY:Age -0.04707424  0.03966906
```

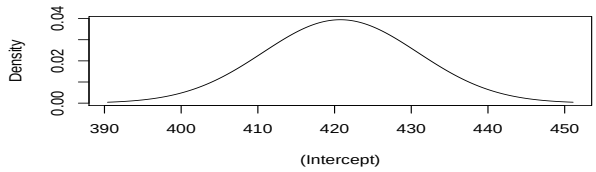
and so, given the data and E_N, there is about a 97.5%-2.5% = 95% chance that μ_N is in this interval, and so on for the other intervals. If somebody (and there unfortunately is always somebody) insisted on having just one "best" number, then, while reminding them that this is not wise because one number does not well summarize the uncertainty you have in the parameters,

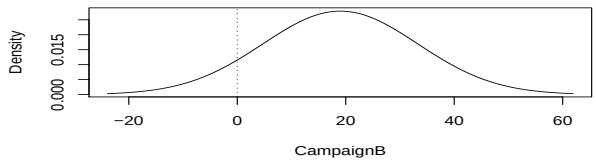
nor in any sense actual observables, you would give them the `Estimate`. But you should remind them that this is an estimate for an unobservable parameter which gives only weak knowledge of actual observable data.

A few Chapters back, I promised that we would learn how to visualize the uncertainty in the parameters, and since honor is important except to those who wish to get away with something, it's time to keep that promise. Type (first noted in Chapter 7) `source(url("http://wmbriggs.com/book/Rcode.R"))` to load up the software for this book into `R`'s memory. Go back and refit Uncle Ted's example, then type

```
glm.posterior(fit)
```

This will give you the picture of the posterior probability distributions for μ_A and μ_{B-A} given the old data and E_N. It does not give you a picture of the posterior probability distribution for σ, but do not forget that it exists and we need to know it when we talk about observables later. You have to hit "Enter" to get each new picture, which are shown here.





A dotted vertical line is shown at 0 so you can eyeball how much area is above and below it. Even better, the probability that each of these posterior parameters is less than 0 is also given by `R`; you should see this:

```
Probability parameter  (Intercept) < 0 | x, E_N = 0
Probability parameter  CampaignB < 0 | x, E_N = 0.0920139
```

This says that the posterior probability that μ_A is 0, but it actually means "close enough to 0 for anybody", because recall these parameters are based on normal distributions, and there is always some probability of being less than any number, not matter how small (the probability or the number). The posterior probability that μ_{B-A} is less than 0 is 0.09 (round the numbers!), which means that the probability it is greater than 0 is $1 - 0.09 = 0.91$.

It is up to you to do this for the appendicitis example. Incidentally, you can get both pictures of Uncle Ted's example, or all four pictures of the white blood example on the same graph in `R`, you just have to be a little clever. For the former, type `par(mfrow=c(2,1))`, which says "create a plot matrix

with 2 rows and 1 column"; for the white blood type `par(mfrow=c(2,2))`, which says "create a plot matrix with 2 rows and 2 columns." This command is ugly, unintuitive and unmemorable, but we are stuck with it. Type the `glm.posterior` *after* the `par` function (which stands for "set a graphics parameter").

5. Back to observables

Recall, what you might have forgotten, that we want to quantify our uncertainty in actual sales or white blood count and *not* the parameters. This is why we created the regression model in the first place. In the former example, we had probative information in the form of campaigns, in the latter, the probative information was whether or not each patient had appendicitis and their age. In order to quantify our uncertainty in the observable *we have to specify values for each of the probative variables*. These are the values at which you are interested in seeing what happens to the uncertainty in the observable. In the Uncle Ted example, this means specifying a campaign (A or B). In each case, we can quantify the uncertainty in future sales given this specified campaign value (and given the information from the old data and E_N). We can then ask very important questions like "What is the probability that future sales in campaign A are less than campaign B given the old data etc.?"

We first need to ask for the probability distribution that quantifies our uncertainty in the future observable if `Campaign = A`. To do this in `R`

```
s1 = obs.glm(fit,data.frame(Campaign="A"))
```

We will store the results in `s1`—for "scenario 1"—which contains information on the probability distribution describing our uncertainty in *future* sales under campaign A. You have to create the scenarios! Just like you had to supply a value for each variable on the right hand side of the regression model. The new function is `obs.glm`, which tells us it is concerned with `obs`ervables from the `glm` regression model. If you can remember as far back as Chapter 5, you might recall that `R` calls datasets `data.frame`s. Well, one term is as good as the other, I suppose. We have to tell `obs.glm` what the future new data will be, so we pass it a new dataset, or `data.frame`. The only thing you have to do is to remember to include a value for every probative variable (all the variables on the right hand side); since this model only had `Campaign`, that's all we have to provide. Spelling and capitalization counts! Typing `Campaign="a"` will give you an error because there was only a `Campaign="A"` before. Similarly, Typing `campaign="A"` will give you an error because there is no variable called `campaign` with a small `c`. Now do a scenario for Campaign B

```
s2 = obs.glm(fit,data.frame(Campaign="B"))
```

The last step is to answer the question and draw a picture (which I'll show you only for the appendicitis example), which we do with

```
obs.glm.prob(s1, s2)
```

and you should see

```
Posterior probability that s1 < s2 =  0.616904
```

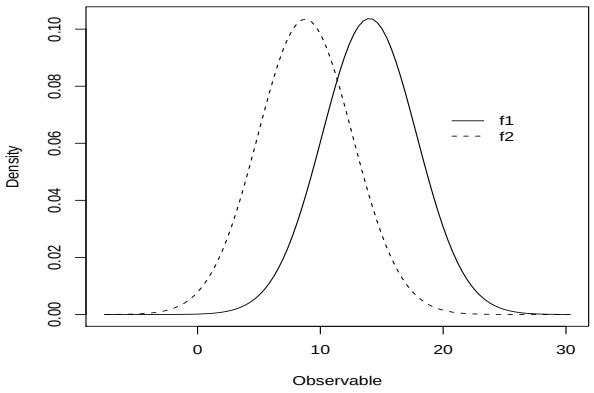
The function `obs.glm.prob` gives you the posterior probability of the difference between two scenarios, where it is up to you to create the probability distributions of the scenarios.

Last Section, we learned that, given the old data etc., the probability of $\mu_{B-A} > 0 = 1 - 0.09 = 0.91$ (this means a 91% chance that the central parameter of sales under B is larger than the central parameter of sales under A). However, this only translated to a 62% chance that the *sales* under B is larger than the *sales* of A. *We are less certain in the observables than in the parameter!* This is how we did that calculation.

Now to the appendicitis example (first re-run the `fit = glm(...)` for this example), were we must remember to specify values for *both* `Appendicitis` *and* `Age`. Scenario 1 might be

```
s1 = obs.glm(fit,data.frame(Appendicitis="N", Age=31))
```

where I picked `Age = 31` for no good reason, other than this was the median observed value of age and thought it would be interesting. Scenario 2 is the same except for `Appendicitis="N"`. What is the probability that, give the old data etc., a white blood count from somebody not having appendicitis will be less than the white blood count from somebody having appendicitis? This is had by typing `obs.glm.prob(s1, s2)` and we see `Posterior probability that s1 < s2 = 0.1531845`. If you did the homework as you were supposed to you would have discovered that the probability, given the old data etc., of $\mu_{Y-N} > 0$ was almost 1 (it was about $1 - 10^{-10}$). But the probability the *observables* differ is 85%, which is certainly respectable and, if you had to bet, perhaps even convincing, but it is *not* 1. Running the `obs.glm.prob` function also gives you this picture



This shows the probability distributions of the observables under both scenarios. As expected, the distribution under scenario 2 gives higher probability for large amounts of the observable (white blood count); the distribution under scenario 1 gives higher probability for smaller amounts of the observable. You can see that scenario 2 even gives, what is perhaps not a negligible amount of, probability for white blood counts *less than 0!* This is the problem using normal distributions I warned you about. This is real data, folks, and real studies and papers were published with it using just these kinds of distributions. This happens *constantly*.

The reason is happens so often—and you're probably sick to death of hearing this by now—is that people stop at saying something about the parameters and do not carry the analysis through to the observables. If more analysis were to take the extra step to the observables, people would realize the huge mistakes they are making. Normal distributions are quite simply overused. Ah well, close enough.

See that point where the two distributions overlap? At a white blood count of about 12. This is the point at which, if you learned a 31 year-old patient had a white blood count larger than 12, you would guess he had appendicitis. If he had a count less than 12, then you would guess no appendicitis. This is because more probability is allocated to those counts under both scenarios. Next Chapter, we'll quantify this decision process better.

Before we move on, let me answer a question that has probably been bugging you. Let me create two more scenarios, both of which are for people who have appendicitis. The first scenario is for people 80 years old; the second is for those who are 20. What is the probability that, given the old data etc., 80 year-olds with appendicitis have a lower white blood count than 20 year-olds? Using the code above (*do this!*), we discover it is 56%. For people without appendicitis, it is nearly identical: 55%. Consider, if you knew nothing about appendicitis or white blood count or age's effects on them, and somebody asked you, "Given I grab two people, one 80 and one

20, what is the probability that the white blood count for the 80 year-old is less than the 20 year-old's?" Conditional on that evidence, it is 50%. Thus, adding the evidence that age and appendicitis are linearly related to white blood, and the old data (and E_N), the probability they are different is about 55%. Which is to say, *Age is not that important in quantifying our uncertainty in white blood count.* All that new evidence about age only changed the probability five points. Whether that five points is important or not depends, of course, on how you use the information, but it is likely that, in most applications, the difference is trivial. The question that was bugging you can now be voiced: If Age is not important, why have it in the model? Why indeed? We'll have to talk about this later when we learn how to cheat with these models.

6. Homework

(1) Follow the examples above for Uncle Ted and Appendicitis. Type every command you see in the book. Make sure you understand what you are doing.
(2) Now do the same thing for the data you collected in Chapter 8 homework. The data the uncertainty of which can be described by a normal distribution.
(3) Keep this data at hand. It might become the basis of your book project, described in the Preface.

CHAPTER 13

Logistic Regression & Observables

1. Logistic Regression

When we sat on the corner in Chapter 11 we saw 14 out of 20 men and 12 out of 15 women wear a thinking suppression device[1]. Let's write this data in a new way, one which is easier for the computer to read:

```
TSD,  Sex
  1,    M
  1,    M
 ...
  0,    F
```

This is in the form of a CSV file `tsd.csv`, which you can download at the book website. Each 1 is a "success" and each 0 a "failure." In Chapter 11, we handled this kind of data using classical testing, but here we want to use regression. From what we learned last Chapter, for this situation we might try a model like this:

$$\text{TSD}_i = \mu_F + \mu_{M-F}\text{I}(\text{Sex}_i) + \epsilon_i,$$

where μ_F represents the central parameter...wait! Didn't we use binomial, and not normal, distributions to express our uncertainty in the number of men and women wearing TSDs? We did. Look at the equation again. The left hand side must be either a 0 or 1 because TSD can be only 0 or 1, but it's hard to imagine values of μ_F, μ_{M-F}, and ϵ_i that could make that happen in this equation. It would too easy for values of TSD to be larger than 1 or smaller than 0, or some number in between, no matter what μ_F etc. are. Something has to change. Either we change the way we write the equation on the right hand side, or we change the way we write the response on the left hand side. I won't leave you in suspense. It's the latter.

Suppose the probability that TSD_i is a success is θ_i, where the i subscript allows the probability of success to change due to the changing values of the variables on the right hand side. This is effectively what we did in Chapter 11, where we had two different probabilities of success, one for men and one for women. It will be the same here: either θ_i would equal μ_F or $\mu_F + \mu_{M-F}$ (convince yourself of this first!).

[1] Obviously, this is historical data; today it would likely be closer to 20 out of 20 and 15 out of 15

This works in this simple case, but, like before, we will want to add explanatory variables that help us explain TSD successes. Age again might be one of these variables; so might income. If we are to use variables like this, the right hand side of the model, as before, must be allowed to take any possible value (I mean, once the equation is solved by plugging in values for all the variables). What are all the possible values? All numbers. Which are numbers going from negative infinity to positive infinity (See Chapter 4). This means we have to transform the left hand side so that it can, at least theoretically, go from negative to positive infinity, too. Remember odds? Odds were a one-to-one transformation of probability: you can either speak of odds or probability and mean the same thing. If the (unknown) probability of success is θ, then odds $= \theta/(1-\theta)$. Plug in a $\theta = 0.99$, which gives odds $= 99$ (probabilities close to 1 give larger odds, up to infinity for θ going to 1); $\theta = 0.1$ gives odds $= 0.11$; $\theta = 0$ gives odds $= 0$. Well, we're half way there. We have transformed the *probability* of the left hand side so that it can be any number from 0 to infinity; we still need to do something about 0 to negative infinity, since odds only go from 0 to infinity. The solution is to take the *logarithm* of the odds; this works because logs of numbers from 0 to 1 are negative, getting smaller as the number goes towards 0. A lot of work, but it leads to this model

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mu_F + \mu_{M-F}\mathrm{I}(\mathrm{Sex}_i), \tag{37}$$

This is our *logistic regression* model, and we use it whenever the left hand side is a 0/1 variable. The right hand side is the same as it was last Chapter. You need learn no new rules or techniques about how to interpret the coefficients, except for the change in wording due to the slight modification of the left hand side, so that now μ_F is the parameter describing our uncertainty in the log odds of a success for females. μ_{M-F} is still the difference between females and males in the parameter describing etc.

Since we already know all about how to work with these models, let's go straight to the computer! Type this in `R` (don't forget to `attach` the data first)

```
fit = glm(TSD ~ Sex, family=binomial)
```

This is exactly like the old regression code except for the addition of `family=binomial`, which tells `R` that the response is 0/1, that we want to model our uncertainty in the observable using a binomial distribution. Though you didn't know it, last Chapter you were actually typing `family = gaussian` (which means `family = normal`); you didn't see it because that option was the default, but it was always there. What's next? This: `summary(fit)`. The output looks just as it did before:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3863     0.6455   2.148   0.0317
SexM         -0.5390     0.8092  -0.666   0.5053
```

Same classical interpretation as before: $\widehat{\mu}_F = 1.39$ etc.; the p-value conditional on the "null" hypothesis of $\mu_F = 0$ is 0.03, etc. The modern details for the posterior distribution of the parameters are the same as before too:

```
glm.posterior(fit)
```

which shows

```
Probability parameter  (Intercept) < 0 | x, E_N = 0.0159
Probability parameter  SexM < 0 | x, E_N = 0.7473
```

where you have to still remember that `SexM` is the parameter μ_{M-F} etc. Also don't forget that this is all still in terms of "log odds" of a success, which makes the coefficients somewhat difficult to interpret.

This leaves us with future observables, which are not "log odds", but actual yes/no (or 1/0) observations. First thing to do is to back transform our model equation so that it at least looks more like our observable and not something weird like log odds. Get ready for the most difficult math in the book—you don't really need to understand how to derive the equation, but you do need to understand what it means! What we are going to do is to work with equation (37) by first taking the exponential of both sides. You will certainly remember from high school that taking the exponential of a natural logarithm cancels it, i.e. $\exp(\log(x)) = x$. This gives us

$$\begin{aligned}
\frac{\theta_i}{1-\theta_i} &= \exp\left(\mu_F + \mu_{M-F}\mathrm{I}(\mathrm{Sex}_i)\right) \\
\theta_i &= (1-\theta_i)\exp\left(\mu_F + \mu_{M-F}\mathrm{I}(\mathrm{Sex}_i)\right) \\
&= \frac{\exp\left(\mu_F + \mu_{M-F}\mathrm{I}(\mathrm{Sex}_i)\right)}{1+\exp\left(\mu_F + \mu_{M-F}\mathrm{I}(\mathrm{Sex}_i)\right)} \\
&= \frac{1}{1+\exp\left(-\left(\mu_F + \mu_{M-F}\mathrm{I}(\mathrm{Sex}_i)\right)\right)}
\end{aligned}$$

where in the last step I divided the numerator and denominator by the quantity $\exp\left(\mu_F + \mu_{M-F}\mathrm{I}(\mathrm{Sex}_i)\right)$. We now know what the probability of success is given we plug in a Sex. We still do not know the exact value of each θ_i because we do not know the exact values of the parameters, but given the old data and E_B and via some hidden math not different in spirit from what we did in Chapter 8, we can calculate our uncertainty in the parameters and in future values of TSD. The function `glm.posterior` has already given us some idea of the parameters.

Now we can ask questions about our future observables through the creation of scenarios. A good one might be, if tomorrow we saw 10 more men and women, what is the probability, given the old data etc., that more women than men will wear a TSD? Or if tomorrow we saw 1 man and 1 woman, what is the probability that she wears a TSD and he doesn't? The point is that you have to specify two things in these future scenarios: whether the people are men or women and how many people there will be. In regular regression we did not have to specify a future sample size, but you will remember that the binomial needs to have a sample size in order to

calculate probabilities. Here is one possible scenario (you still have to create all scenarios)

```
s1 = obs.glm(fit,data.frame(Sex="M", n=1))
```

and we create another with `Sex="M"` (note that we do not have to tell the function that we are working with a binomial distribution, it can figure it out for itself). Then, as in last Chapter, to compare these two scenarios use

```
obs.glm.prob(s1, s2)
```

where you should see

```
Posterior probability that s1 < s2 =  0.241
Posterior probability that s1 > s2 =  0.151
Posterior probability that s1 = s2 =  0.608
```

This means that the posterior probability that the woman we see wears a TSD and the man does not is about 24%. The probability that the man would wears one and the woman does not is 15%, while the probability they both would or both would not is 61%. With this function we get three different probabilities for logistic regression models, whereas in the normal we only saw two, because when we use the binomial there is a substantial probability that two future observable numbers of successes can be equal.

Recall that when we just ran `glm.posterior` the posterior probability that the *parameter* $\mu_{M-F} = \mu_M - \mu_F < 0$ or $\mu_M < \mu_F$ was 0.7473. But we see here (running `obs.glm.prob`) that the probability of the future *observable* woman wears a TSD when a future observable man does not is just 0.24. Once again, we are far less certain of the observable than we are of the parameter.

What if we expected 10 new men and women? Then

```
Posterior probability that s1 < s2 =  0.572
Posterior probability that s1 > s2 =  0.273
Posterior probability that s1 = s2 =  0.155
```

The probability that more women than men would wear a TSD is about 57% (compare this to the figure we got in Chapter 11). In the future, if you expect a lot more men and women (try an `n=1000` in the `newdata`), the chance that the exact same number of men and women wear a TSD shrinks (see the homework), and the probability that the women outnumber the men is about 75%. And that is the *highest*, given the old data etc., that it will *ever* be. Let's think about this. Suppose you knew nothing about the past data and somebody said that a large number of men and women would walk by tomorrow either wearing a TSD or not. This is all you know. Given that information, what is the probability that the number of women wearing TSDs would outnumber the men? Right, logically it is 50%. So the effect of the old data was to increase our certainty (for large future numbers of observables) in the question to 75%. Is that a big change? Well, that depends on the application. See the homework for another example.

Let's return to the appendicitis example and reveal its true purpose. Before, we tried to quantify our uncertainty in white blood count given the patient had appendicitis or not and knowing their age. But the true thing of interest is whether the patient has appendicitis, which is unknown, even after the white blood is counted. What we want is to use the white blood and age to help us explain the presence or absence of appendicitis. The mathematical model is

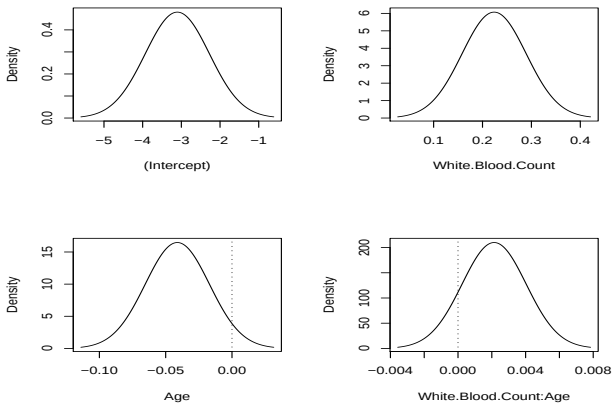
$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_w \text{WBC}_i + \beta_a \text{Age}_i + \beta_{wa}\text{WBC}_i \times \text{Age}_i, \tag{38}$$

where I have used the shorthand `WBC = White.Blood.Count`. The coefficient β_0 is the intercept and is not of direct interest (see homework). The parameter β_w describes how the log odds of appendicitis changes for every unit change in `WBC`: every increase in `WBC` by one increases the log odds of appendicitis by β_w. Same thing for β_a with respect to `Age`. The parameter β_{wa} is a little more complicated; it is the *interaction* between white blood count and age. As that interaction increases by 1 unit (the unit is count per μl-year), the log odds increases by β_{wa}.

Here's how to get our model:

```
fit = glm(Appendicitis ~ White.Blood.Count*Age, family=binomial)
```

Running `glm.posterior(fit)` gives




```
Pr parameter  (Intercept) < 0 | x, E_N = 0.9999
Pr parameter  White.Blood.Count < 0 | x, E_N = 0.0003
Pr parameter  Age < 0 | x, E_N = 0.9556096
Pr parameter  White.Blood.Count:Age < 0 | x, E_N = 0.1304
```

Here we learn interesting facts like the posterior probability that $\beta_0 < 0$ is almost certain, the posterior probability that β_w greater than 0 is $1 - 0.0003 = 0.9997$, which indicates that white blood count has something to say about appendicitis, and so on for other juicy tidbits about the parameters. The posterior distributions of the parameters are pictured. These are all hard to

think about because they all describe what happens to the log odds. So let's go to something important, the observables.

Pick two scenarios, both for a patient 31 years old, which is the median age. Our main interest is the effect of white blood count. Scenario 1 is for people with a white blood count of 7, and the other of 13, which are roughly the first and third quartiles of the data (run a `summary(White.Blood.Count)`). We still have to pick a future sample size. Start with $n = 1$, so for scenario 1 we have

```
s1=obs.glm(fit,data.frame(White.Blood.Count=7,Age=31,n=1))
```

and we create another with `White.Blood.Count=13`. Then

```
obs.glm.prob(s1, s2)
```

```
Posterior probability that s1 < s2 =  0.321
Posterior probability that s1 > s2 =  0.057
Posterior probability that s1 = s2 =  0.621
```

Suppose two 31-year old people walk into the Emergency room. One has a white blood count of 13, the other of 7. The output shows that the probability the person with a white blood count of 13 has appendicitis and the person with a count of 7 does not is 32% (remember: your results might differ slightly from this). The probability that both people either both have or both do not have appendicitis is about 62%. The probability that the person with a white blood count of 7 has appendicitis and the person with a count of 13 does not is about 6%. White blood count is important: higher counts give higher probability of having appendicitis.

Now suppose the future $n = 10$ (which means 10 new people in each group: low and high white blood count and aged 31). Run `w=obs.glm.prob(s1,s2)`. This stores the output in `w`, which will be a new *dataframe* with two variables, `s1d` and `s2d`, the probability distributions for the two scenarios (`obs.glm.prob` automatically calculates these for you; we'll use these in a moment). These two probability distributions show the probability of all possible events from 0 to n, which is $n+1$ possible events, for both scenarios. Here are the results:

```
Posterior probability that s1 < s2 =  0.887
Posterior probability that s1 > s2 =  0.037
Posterior probability that s1 = s2 =  0.076
```

The `Posterior probability that s1 < s2 = 0.887` is getting larger. The proposition "`s1 < s2`" means that "More people in the s2 group have appendicitis than in the s1 group." This proposition, conditional on all the evidence, shows that there is a 89% chance that more of the 10 people in the high WBC group will have appendicitis than in the low WBC group. There is still about a 4% that more people in the low WBC group will have appendicitis than in the high WBC group. Last, pick a future $n = 1000$ and we get

```
Posterior probability that s1 < s2 =  1
Posterior probability that s1 > s2 =  1e-10
Posterior probability that s1 = s2 =  1e-11
```

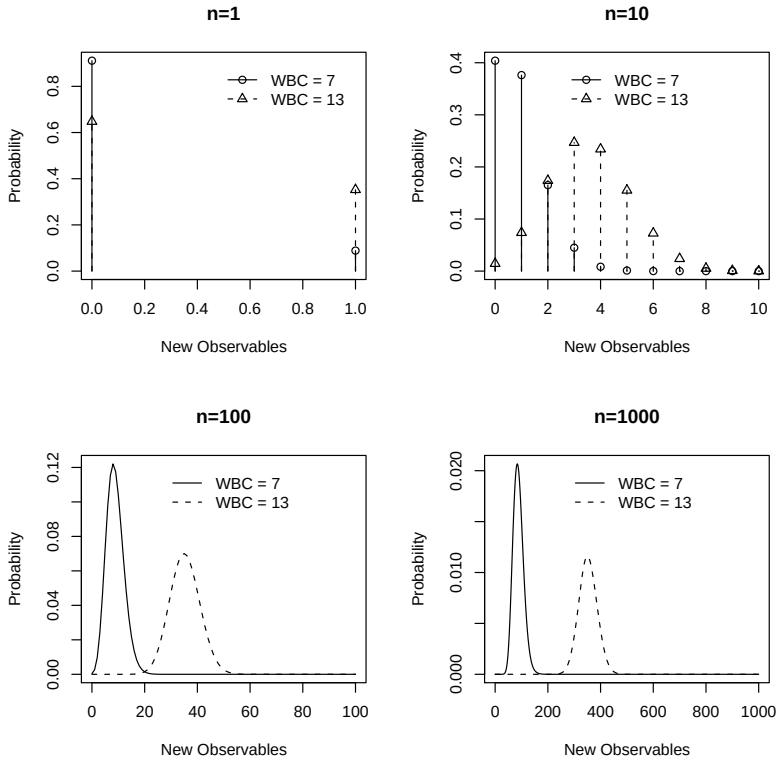


FIGURE 1. Probability distributions for scenario 1 (low WBC group) and scenario 2 (high WBC group) for four different future sample sizes. When $n \geq 100$, the software automatically switches from the discrete labels to lines to make the plot easier to read.

This means, in 2000 new people (1000 per white blood count group), that more people with white blood counts of 13 have appendicitis than the people with counts of 7 is near 100%. The probability that both groups of people either have or do not have appendicitis is now very small, almost 0% (why can't it be eactly 0%?). The probability that the people with white blood counts of 7 have appendicitis and those with 13 do not is also about 0%. Increasing the future n changes these numbers very little.

We can learn more about what is going on by examining the pictures produced by `obs.glm.prob` (Fig. 1), which shows the posterior probability distributions for both scenarios for four different future sample sizes, $n = 1, 10, 100, 1000$ (we didn't show $n = 100$ above, but it's produced in exactly the same way). Don't forget this shows the probability of there being k successes out of n, for $k = 0, 1, \ldots, n$.

Suppose $n = 1$, and pick a WBC group, high or low. What is everything that can happen for that group? Right, either the person has appendicitis or they do not. From the picture, it looks like the low WBC person has about a 10% chance of having appendicitis, while the high WBC person has just under a 40% chance. The probabilities spit out by `obs.glm.prob` merely summarize this picture. Now let's increase the number of people to $n = 10$ per group. The probability distribution of everything that can happen is pictured on the upper right panel of Fig. 1. We can start to see the divergence of the probabilities, but there is still a lot of overlap. For example, the probability that just 2 out of 10 people have appendicitis is about the same regardless if the people have high or low WBCs.

By the time we get to $n = 100$, we can see that there is a clear difference between the two distributions: most probability for scenario 1 is for low numbers of people with appendicitis, while scenario 2 gives most probability to a little less than half the group having appendicitis.

You will have noticed that `obs.glm.prob` also gives more information than we have so far discussed. It also shows you the most likely value for each scenario, plus the probability of future observables being greater than the most likely value for scenario 1. Sometimes these numbers are helpful diagnostics. In the two scenarios with $n = 1000$ 31-year olds, but with low and high WBCs, the most likely number of patients out of 1000 with appendicitis is 84 for the low group and 347 for the high. There is about a 57% chance of 84 or more patients having appendicitis in the low group, and a 52% chance of 347 or more patients with appendicitis in the high group.

Finally, let's change the scenarios, one for a white blood count of 4 and one of 25—both are near the ends of the observed white blood counts—for $n = 1$ new people. Then we get this result:

```
Posterior probability that s1 < s2 =  0.903
Posterior probability that s1 > s2 =  0.002
Posterior probability that s1 = s2 =  0.094
```

This means there is a 90% chance that the person with really high WBC has appendicitis and the person with the really low does not. There is still about a 9% that either both people have or do not have appendicitis (you can look at the picture of this on your own). With these levels of WBCs, by the time we get to $n = 10$, there is nearly 100% chance that more people with really high WBC have appendicitis and those with really low do not. These analyses clearly show that WBC is important in predicting appendicitis. It is not perfectly predictive—in the $n = 1000$ people who WBC = 13, only roughly 400 people will have appendicitis—but we certainly need to examine it when diagnosing a patient.

OK, let's try two new scenarios, both with the median observed white blood count (9.5), but with ages 48 and 22 (about the 3rd and 1st quartiles of the old data), and with $n = 1000$.

```
Posterior probability that s1 < s2 =  0.960
Posterior probability that s1 > s2 =  0.038
```

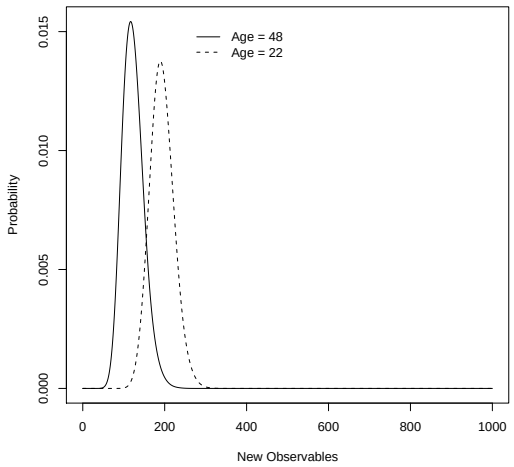


FIGURE 2. Probability distributions for scenario 1 (Age = 48) and scenario 2 (Age = 22) both with WBC = 9.5 a future sample size of $n = 1000$.

```
Posterior probability that s1 = s2 =  0.002
```

There is 96% chance the more younger people will have appendicitis (at this level of WBC). Age therefore seems to be important. Figure 2 backs this up to some extent: we expect around 200 or so out of the 1000 young people to have appendicitis and we also expect about 120 or so out of the 1000 older people to have appendicitis. This is a difference, but not nearly as dramatic as for WBC. Thus, Age seems less important to explain our uncertainty in appendicitis. You should also look at scenarios with low age and high white blood count versus high age and low white blood count, and so on.

By now you're likely thinking something like this: "Briggs, this is too much! You are being a nuisance. I don't understand why we need all this 'scenario' nonsense. Why can't you just give me a simple way to decide whether or not variables like white blood count are important or not? Classical statistics did this for me, at least, by saying that if the p-value was low, then I could write a paper saying that I had proved high white blood counts are indicative of appendicitis. All this modern stuff is just too confusing. I want relief!" To which I reply, sorry, pal. I have no comfort to offer. Understanding uncertainty and selecting the best models is hard work and there are rarely easy answers, and when easy answers are offered, they are usually too certain. We are asking complicated questions and the we must expect that the answers are just as complicated (e.g. Wasserman, 2000; West, 1986).

It is true, however, that classical statistics was designed to do the thinking for you. P-value less than 0.05? Success! P-value greater than 0.05? Success, too; at least we could say "no effect." Either way, you'd be done.

The decision was out of your hands. It was objective! That was a benefit of focusing solely on parameters, a focus which unfortunately is echoed in most Bayesian statistical methods. Parameters are lone creatures and it is easy to say with what probability one is smaller or larger than 0 and therefore "significant." How gratifying!

But let us not be complacent. The new methods we have just learned (for binomial and normal distributions), while they certainly give a clearer picture of the uncertainty than the old ways, are not panaceas. For example, the best question to ask about future data is *not* always "What is the probability s2 > s1?" (This is why we also show the full pictures of the probability distributions.) It may well be that this probability is greater than 50%, but that it still does not make sense to opt for acting like scenario 2 will happen. For example, acting like this scenario might happen might cost us a lot of money, more than we're willing to pay. This brings up the topic of *decision analysis*, a subject whose details unfortunately take us beyond this book.

So far we have ignored one of the most important provisos. The statements we make, like "The probability that more people, out of the next 10 in each group, who have high WBC will have appendicitis is 89%" is conditional on evidence like the past data and the uncertainty in it can be quantified with a binomial, but that statement is also conditional on the new people "looking like" the old people. That old data was collected on a group of people who had a list of certain characteristics, taken at a certain place and time. Are future people, the people we are making predictions about, just like the old people? Probably not in every way. This means that that 89% is too certain, and should, through a mathematical adjustment, actually be closer to 50%. We'll talk more about this in Chapter 15, where you might be surprised to learn that we often (or nearly always) cannot formally calculate this adjustment, and is yet another example of why too many people are too certain about too many things.

We have seen time and again that certainty, even absolute certainty, in the parameters does not translate into the same level of certainty in the observables. You can be as sure as you like about the value of a certain parameter, but it does not mean that it makes a meaningful difference in the observable. In this way, classical statistics, but also much of Bayesian statistics, gives one a inflated sense of surety. People come away from an analysis too confident. Instead, the modern approach forces you to focus on reality, which is never as simple as you would like it to be. Even after we have our model in hand, we are full of uncertainty. But that is as it should be, because that is the way it is. Still, it would be desirable to have some mechanism with which to judge the overall efficacy of our model. That mechanism is called skill.

2. All models are not wrong

The statements of uncertainty we made above were based on probability models. Were these the right or wrong models? There is a saying often heard in statistics, attributed to George Box, which goes "All models are wrong, but some are useful." That beloved statement is false. Box actually said, "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." That statement, with its richer and more complicated language, is just as false.

The odd thing is that word *wrong*. What do people mean when they say a model is "wrong"? When a patient is disgorged from the ambulance (complaining of right lower quadrant pain) we do not know whether or not he has appendicitis. We can certainly guess. Let's do so. If I guess right for the first patient then whatever "model" I used to create the guess is right and not wrong. Suppose I guess correctly a second time. My model is still right and not wrong. In plain English, my model has been right. If a statistician insisted that it was "wrong", he would be using that word in a way that does not make sense.

Would Box say that my model is "wrong"? He would, because he would suspect that I couldn't keep my batting average up. Box would argue, rightly, that future values of appendicitis are contingent, their value is conditional on the universe being in a certain way, and that the probability of all contingent events is between 0 and 1. The future values are contingent conditional on the information that people who complain of right lower quadrant pain have appendicitis or not, and by whatever past observations we have. So if by "wrong" Box means that it is *impossible* to correctly guess every *future* value of appendicitis conditional only on our past data, then it is strictly true that every model is wrong, because our model would spit out nothing but probabilities which would never equal 0 or 1. But this is a trivial truth because its says that "all models are wrong" is equivalent to "observables are contingent." This is, or should be, no surprise to statisticians whose livelihoods depend on creating probability models for observables.

How about our old friend, E = "Toss a die of six sides, just one of which is labelled 6" therefore A = "We see a 6." The proposition P = "The probability of A given E is 1/6" is true. That is, the probability of P given any tautology is 1. Our model here is P, and it is true, which is to say it is not wrong. There are, of course, many more examples like this. For example, suppose the proposition we made about people with high WBCs was labelled A (for ease of notation), then let

$$\mathrm{M} = \text{“}\Pr(\mathrm{A}|\text{Old data, T, } \mathrm{E}_B) = 0.89\text{”}$$

The probability, given any tautology T, that M is true is 1. That is, the statements we make conditional on assuming a model is true are true themselves. This does *not* say that we have chosen wisely with E_B. That is,

$$0 < \Pr(\mathrm{A}\&\mathrm{E}_B|\text{Old data, T}) < 1$$

It is not true that "A *and* (the model) E_B" is true, because that model was not deduced; the choice of E_B is contingent (or rather, the premises that lead to E_B are themselves contingent; this is so even if $\Pr(\mathrm{A}—\mathrm{E}_B) = 1$; but we are starting to go too far).

The revelation that we cannot guess perfectly contingent values every time almost makes Box's original statement true. It seems that all models *are* wrong in this sense. What about my correct guesses of the first two patients? My model for those two was not wrong. Are you ready for a *big secret*? It is *always* possible to find a model that fits past data perfectly. Even worse, it is always possible to find an infinite *number* of models that fit past data perfectly. It is even trivial exercise to find such models (I talk about this in Chapter 14). These models, since they fit the past observed data perfectly, are not wrong in the ordinary English sense of the word since; after all, they have no error. There is always a suspicion that these models are not really perfect because of our doubt that they perfectly predict *observables that are not yet seen*. It is believed that future data is contingent, and thus whatever model we have cannot keep up its performance. If the future data *is* contingent, and we will use our model built from old data to predict it, then the model must be wrong in the strict sense that it cannot always predict perfectly.

Very well. This much is true, but it isn't really what classical statisticians had in mind when they said a model was "wrong." Box is implying something about vague "randomness", which carries a lot of mystical baggage (you can often read statements, for example, of data *being* normal). In fact, we now know that the only reason we use a probability distribution is to quantify our uncertainty in observables. Statements made conditional on our old data and model plus something like E_N or E_B are *true statements*. They are probabilities (strictly between 0 and 1) that a particular observable will equal a certain value, but these probabilities are not wrong. It is the case that two different models (for the same old and future data) may be of different utility, however, but neither one, conditional on the old data etc., is wrong. One model might be more useful than another. Usefulness is measured by skill.

3. Skill

Here are two models that quantify the uncertainty in white blood count

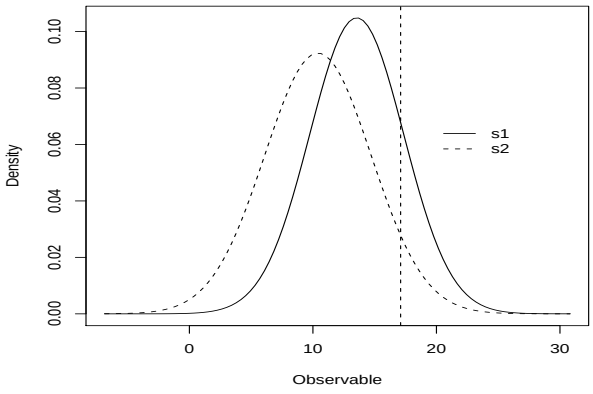
$$\begin{aligned} \mathrm{M}_1: \ \mathrm{WBC}_i &= \mu_N + \mu_{Y-N}\mathrm{I}(\mathrm{Y}_i) + \beta_a \mathrm{Age}_i + \epsilon_i. \\ \mathrm{M}_2: \ \mathrm{WBC}_i &= \beta_0 + \epsilon_i. \end{aligned}$$

M_1 we met before, where, you will recall, μ_N is the central parameter of a normal distribution describing our uncertainty in WBC for people without appendicitis and who are 0 years old, etc. M_2 we also met, but you might not remember it because here it is wearing a different set of clothes. To re-acquaint yourself, ponder the role of β_0. It is a central parameter for a normal distribution; the spread parameter is still hiding in ϵ. What variables do we

have to plug in? Well, what variables are on the right hand side? None. This means that β_0 is the central parameter of a normal distribution describing our uncertainty in WBC for people. Which people? All of them—with the huge proviso, ever present in any dataset, *those people that are* like *the people we measured* (we skip over this proviso until Chapter 14). In other words, M_2 means simply that we are quantifying our uncertainty in WBC using a normal distribution, not conditional on any other variable. M_1, in a sense, is many different distributions, one for each possible combination of values of the variables on the right hand side.

M_2 is a sort of minimal or "null model", in the sense that all it does is to quantify our uncertainty in WBC unconditional on everything except knowledge that WBC is contingent. Consider you are in the emergency department when a patient is wheeled in complaining of right lower quadrant pain. What is his white blood count? You don't know, you are not certain. And how do you quantify your uncertainty? Using probability, and the distribution you use is the normal (if you are normal). All that is another way of saying you use M_2. Or if you knew their age and if they had appendicitis, you could use M_1; but you might still use M_2 and it might be better in the sense that it better quantifies your uncertainty in WBC.

What do we mean by *better*? Well, let's take the third patient in the database (`d[3,]`). That person is 80 and has appendicitis. If M_1 is scenario 1 and M_2 scenario 2, then we can picture our uncertainty in this patient's white blood assuming all we knew was his age and appendicitis status using the same methods as before. We get this picture



A vertical line is drawn at this gentleman's actual white blood count. You can see that scenario/model 1 gives more of its probability to white blood counts near the actual value. In this sense, M_1 is better than M_2. Now, for *every* patient in the database, regardless of their age and appendicitis status, we will get the same picture for M_2 every time: this *is* the distribution quantifying uncertainty in WBC (given the old data etc.). M_1 is better than

M_2 if it shifts and narrows more of its probability towards the actual value of white blood count for every patient.

All we need now is a measure of the "amount" of probability each model puts near each observation, which is another way of saying that we measure the distance the probability distribution was away from the actual white blood count. The smaller this measure is the better. For example, if your model said the probability of appendicitis was 0 but the patient turned out to have the disease, your model was as far from the truth as could be. If my model said the probability was 0.5, then I would be much closer, and so on. We can compare this closeness across all the data we have and see which model put more probability near the eventual white blood counts. This measure is called a *score*. If $\text{score}_1 < \text{score}_2$, then we can say that M_1 is better than M_2. Usually, the two scores are put together in such a way that they are "normalized", which makes for easier comparison with other models. Doing this creates a *skill score* (Murphy and Winkler, 1987; Briggs and Ruppert, 2005; Briggs and Zaretzki, 2008)

$$\text{skill score} = \frac{\text{score}_2 - \text{score}_1}{\text{score}_2} = 1 - \frac{\text{score}_1}{\text{score}_2}. \tag{39}$$

It's written this way because it's expected that $\text{score}_2 > \text{score}_1$ (after all M_2 is not a very sophisticated model). M_1 has *skill* with respect to M_2 if the skill score is greater than 0 (to a maximum of 1). If the skill score is less than 0, then M_2 is better than M_1 .

Why pick M_2? Because it is a minimal model, a "null" model. Any other model we use will be conditional on more information than just the old white blood counts, so any other model should be able to beat M_1 if this additional information is truly probative. Skill with respect to M_2 represents an absolute minimum criterion that must be met if the model (M_1) is to be of any use. Beyond this, it is often useful to compare the skill of models that have just one additional variable, say M_2 with an M_3 which is the same as M_2 but without controlling for age. An example of this in a moment.

That's all there is to skill scores. The exact score used is called the rank probability score Gneiting and Raftery (2007), and its math is beyond this book. The score has some very nice properties and is often used. However, other scores are possible, particularly those built for user-specific purposes. For example, it might be that being "far off" (in terms of probability) when the white blood is low does not mean anything, and it is far more important to be close when white blood count is high (Briggs and Ruppert, 2005; Briggs and Zaretzki, 2008). In cases like this, custom scores can be built and used. This, as mentioned many Chapters ago, is part of decision analysis. We won't be doing anything that complex here.

Here's how to get the scores and skill score. First fit two models; you already know how to fit M_1. M_2 is got by

```
fit2 = glm(White.Blood.Count ~ 1)
```

where instead of any variables, a `1` is placed (this tells `R` that you only want β_0). To get the skill (assuming you stored M_1 in `fit1`), type

```
skill(fit1, fit2).
```

Doing this gives a skill score of 0.13, which is certainly greater than 0 and indicates M_1 is better than M_2. This is not surprising given the exploratory analysis we did in Chapter 12. Incidentally, this code takes some time to run as it has to loop over every observation in the old data and perform computations on it. So do not worry if it takes a while; all is well.

Replace M_2 with a model that contains just `Appendicitis`; i.e., no age. M_1 stays the same, i.e. with age. If M_1 has skill with respect to M_2, then this tells us that age is adding something useful in our knowledge of white blood count. Doing this (and you will do this in the homework) gives a skill of 0.002[2], so age does help us predict white blood count, but it is a trivial amount. Dropping age from our model will do no harm and make the model easier to compute and work with.

Skill works the same way for logistic regression. Let M_2 be

```
fit2 = glm(Appendicitis ~ 1, family=binomial)
```

and M_1 be the full model with `White.Blood.Count` and `Age`. Running `skill` gives a skill score of 0.19, so M_1 is better than M_2. Again, let M_2 be the same model without age; M_1 stays the same. The skill score is -0.008[3], which is of course less than 0 and which indicates M_2 is better than M_1. Adding age actually harms the model; you might say it is adding unnecessary information which adds to our uncertainty instead of decreases it. Incidentally, the classical p-value on (the coefficient of) Age is publishable; i.e., it equals 0.03, which means that, classically, people would have (wrongly) announced that age plays a "statistically significant" role in predicting appendicitis. The classical estimate for the age coefficients is $\widehat{\beta_a} = -0.02$ which implies increasing age *decreases* the probability of appendicitis. The modern posterior distribution of the coefficient of age even says that the probability that $\beta_a < 0$ is 0.98. The skill score, which is a function of the actual observables and the predictions of those observables, tells us that age does nothing at best, and even adds to our uncertainty at worst; it says we should not consider age. But if all you thought about were the unobservable parameters (coefficients), then you would have believed the exact opposite of what was true.

Before we wrap up, suppose all we know is the score for M_1. It is, say, 24.2. What does it mean? Nothing. Isolated scores without reference to something are meaningless, this is why we have to take a skill score and not just examine the score. We can only know how one model has done with respect to another (on the same data and evidence). We could take as a reference a

[2]Because of the way skill is computed, your value might be slightly different than this, or even be slightly negative. This behavior is normal and is another reason to round.

[3]Your value may again be slightly different.

hypothetical perfect model, which is the one that guesses the exact values of white blood count every time (in advance, of course). The score_2 for that model would be 0, which means the skill score is undefined (dividing by 0 is naughty in math). Perfect models are not possible anyway, since the values of observables like white blood count are contingent, their value is conditional on the universe being in a certain way, and therefore (in advance) predicting it must be done using probability which guarantees that the score for any model is (strictly) greater than 0 (and so the skill score will always be defined).

Incidentally, the skill or skill score you calculate for the data set at hand is, and I hope this doesn't sound silly, applies to the data at hand. This means we cannot be certain that the skill of any one model over another would continue in the same way for future data. We can, of course, quantify our uncertainty in the skill score and ask what the is the probability that future skill scores are greater than 0 (for example). In other words, we have to treat skill like any other observable piece of data. But that is a story for another day (or another book).

Classical statistics does something different. It picks a best estimate for the parameters (the $\widehat{\mu}_N$s etc.), and then plugs in the values of the variables to make a guess of each white blood count. Not a probability guess, but a $\widehat{\text{WBC}}$, a guess which says, "Yes, the future white count will *be* $\widehat{\text{WBC}}$." It then computes measures like R^2, which compares $\widehat{\text{WBC}}$ with the old data (something like the sum of $(\text{WBC}_i - \widehat{\text{WBC}}_i)^2$), and the `AIC` found in the output of `summary(fit)`. Since these methods to do not take into account the uncertainty in the predictions, I don't discuss them further.

4. Homework

(1) In the TSD example, I claimed that as the number of future observables increased, it would be increasingly unlikely that the exact same number of men and women would wear a TSD. Why is this true? HINT: Use exaggeration to solve this.
(2) In the TSD example, in the old data, what is the probability that the number of women who wore a TSD outnumbered the number of men?
(3) Suppose the TSD example were instead the result of a clinical trial, where instead of men and women, we had two different treatments, M and F (which might stand for two drug names). Which treatment—and why–would you recommend? What if drug F costs four times as much as M, which treatment would you recommend then?
(4) In the appendicitis example, why isn't the parameter β_0 of interest?
(5) How is the "null" model used to compute skill different than the "null" hypothesis of classical statistics.
(6) Compute the skill of the white blood count model with and without age. Does putting age and appendicitis status as an interaction (`Appendicitis*Age`) add any useful probative information on white blood count?
(7) This is where the data you have been saving is finally used. The goal is to build a regression model to explain an observable in which you

have an interest. Start with exploratory data analysis, then build models. Do the complete classical and Bayesian parameter analysis. Then do the real analysis: quantify uncertainty in future observables. Compute several different scenarios for the two data sets you collected and compute their probabilities. Investigate interesting questions about the data! Find the skill of your model against the standard null model on the two sets of data you collected. This is your final class project, so do a good job!

CHAPTER 14

Cheating

This Chapter is in the spirit of and is dedicated to Darrell Huff who, in 1954, published *How to Lie With Statistics*, a wonderful book that guided generations of statistical cheaters. That book is still in print. Most of an issue of *Statistical Science* (Steele, 2005) in 2005 contained homages from well-known authors on how to lie in areas which Huff had not touched on. I try not to cover the same ground as Huff or the *Stat. Sci.* authors and have angled my tips especially for those who use statistics in their academic papers, or want to discover how others might have cheated in theirs.

1. Statistics on the loose

Here is a case study to show you how easy it is to cheat with statistics. This kind of cheating is common in advertisements (some more are listed on the book website; see also the homework).

I saw a commercial for Glad ForceFlex trash bags[1], in which they said, in bold, animated letters, that "7 out of 10 *consumers*[2] preferred" ForceFlex (then in small small print) "over the other leading brand." So what is the probability that a "consumer" would prefer a Glad bag? You'll be forgiven if you said 0.7. That is exactly what the advertiser wants you to think. But it is wrong, wrong, wrong. Why? Let's parse the phrase they used and see how you can learn to cheat from it.

The first notable comment is "over the other leading brand." This heavily implies, but of course does not absolutely prove, that Glad commissioned a market research firm to survey "consumers" about what trash bag they preferred. The best way to do this is to ask people, "What trash bag do you prefer?" But evidently, this is not what happened. Here, the "consumer" was given a choice, "Would you rather have Glad? Or *this other particular* brand?" Here, we have no idea what that brand was, nor what was meant by "*leading* brand." Do you suppose it's possible that the advertiser gave in to temptation and chose, for his comparison bag, an inferior one? One that, in his opinion, is obviously substandard to Glad (but maybe cheaper)? It certainly *is* possible. So we already suspect that the 0.7 guess is off. But we're not finished yet.

[1]Viewed on Channel 11, WPIX, 19 July 2007, at 6:56 pm.

[2]This is one of the most idiotic terms invented by businessmen. "Hey, I just saw a *consumer* walking down the street!"

In tiny type at the bottom of the screen, we find these words: "Versus the other leading brand's Tall Kitchen Drawstring trash bag" and "Among those with a preference." So now we know that the "other leading brand" was not just some other bag, but a very specifically chosen one, just as we suspected. But how about that other bit? The phrase "Among those with a preference" should have your system announce *Red Alert!* Because it tells us that there were some people who just didn't give a damn about trash bags, or, at least, the two trash bags presented to them. How many people? We have no idea. But we might suspect it's a lot. Which means that the original guess of 0.7 for the implied, but false, question "What proportion of people prefer Glad", is way off, and certainly far too large.

The commercial had wanted you to believe that the background premise was E = "7 out of 10 people prefer ForceFlex" therefore there would be a 70% chance that *you* would prefer the bag. Closer inspection showed that the evidence E was very different, such that that we can't adequately identify the exact premises. Lesson 1 in how to cheat is obvious: conceal contrary evidence in small print, or somehow obfuscate it.

Incidentally, it is also reasonable to infer that the real evidence is such that the probability you would prefer a bag is less than 0.7 based on the premise that if the advertiser did have better evidence in his favor, he certainly would have used it. He did not, ergo, etc.

The moral of the story is: always be suspicious of other people's statistics, especially when somebody is trying to sell you something.

2. Who are the results valid for?

Remember, as always, the job of probability and statistics is to say something about data not yet seen. In this section, we'll primarily think about the human data. What I have to say applies to physical measurements, too, but it's far easier to cheat with humans. The type I have in mind here are those experiments coming from universities. These are many in number and variety, so we'll only use a couple of common types to illustrate the methods.

One typical type of academic study is one, say, that gathers two groups of college kids (they are always at hand), maybe about 40 in each set, and has them do some task or asks them to rate something. Another common type of study, a poll or survey, gathers data from a small area, say a neighborhood in a city, where the sample size may be as high as a few hundred, and asks sociological and economic questions of the people that live there. A medical experiment might try two treatments in two groups of a hundred or so people. When the data from any of these studies are in, the results are compiled and papers are published. Certain claims are made in these papers, usually about favored theories. The college kids paper will say that people act one way and not another; the city-survey paper will say that

poor people have less money; and the medical paper will claim treatment A is better than treatment B.

We already know that if all these researchers wanted to do was to say something about their datasets—just the people they measured and no others—then they do not probability models. They can look at their data and say for example, yes, more people got better under treatment A than under treatment B. They would be finished. Evidently, however, the creators of these studies do not want to make statements only about past data; they want to imply their findings are more widely applicable.

As said, all these kinds of studies concern humans. As of this writing, there are over 6.6 billion humans alive, about 100 billion are dead, and God only knows how many more are yet to live. The first way to cheat is to *not* mention these facts in your results (unless, of course, you happen to be writing about demography), it will weaken your argument. If you do mention them, your sample size will seem paltry and insignificant.

Are the results from the college kids study applicable to all humans? All those that lived in the past, those that will live in the future, even those that live now but not in the town in which the college lies? Those who are in their 50s?, 80s? who are less than 10? Poorer people and those with enough money to "get a degree[3]"? Kids at other universities? Let's be clear: researchers will gather data on their 100 kids, create a probability model, and since they have read this book, they will not just make a statement about the parameters, but calculate the probability distribution of future observables. The only problem is, to what people do we apply this probability distribution?

Before we answer that, let's think about the medical trial, which was conducted at a hospital in a city on the East Coast of the United States of America. The physicians used their data to create a probability distribution of future patients. But who exactly are these patients? People who live in other cities on the east coast?, anywhere in the USA? Canada, too? Or only cities of a certain size? Or do the future patients merely have to "look like" the patients in the old data; that is, be of the same ages, sex ratio, weights, economic condition, have eaten the same things in their lifetimes, traveled to the same places, engaged in the same activities, and so on? Would it have applied to the people who used to be alive, and to people not yet born, indefinitely into the future?

Nobody knows the answers to these questions, which is highly in your favor if you want to publish a study like these. You certainly want to imply that your results are as broadly applicable as possible because this makes you more of an expert than somebody who merely claims to know the habits of a small group of college kids in the year 2008 in a small college town and kids who are unmarried, between 19 and 22 years old, and whose parents are upper middle class, etc. Openly stressing these limitations might be noble

[3]Kids go to college to "get a degree" nowadays, and not usually for anything else. Well, maybe socialization. These are rational choices given the way things are.

and correct, but it will not get you far. State, or at least broadly hint, that your results are in terms of all people. For example, say "People fail to think correctly when presented with our experiment, which gives weight to our theory of psychology." Do *not* say, "College kids in our freshman psychology class, who might not be anything like the rest of the population, carried out an experiment for us—and surely they took this task seriously—and..."

In short, be loose describing the nature of your sample; or, rather, say as much about your sample as you like, but say little or nothing about whom you expect your results are applicable. Certainly imply that all humanity falls under your results, especially if you are working in any non-physical area. With any luck, a reporter will find your paper and help you along this road by summarizing your results, leaving out all hint of limitation with a headline like "Kumquats reduce risk of toenail cancer."

We'll talk more about this subject—about to whom statistical results pertain—next Chapter.

3. Randomization

In classical statistics, all data, before it can be analyzed, must possess the mysterious quality of *randomness*. This is in reinforced by the mistake of calling data, say, "normal", as in "WBC is normal", when what really should be said is that "our knowledge of WBC is quantified by a normal distribution". A lot more words, but a lot more correct and surely less apt to be misleading. Anyway, data does not have to come "randomized", because that word only means unknown. Of course we do not know the values of data before we know them! Modern statistics takes data as they come, whether in a planned controlled study or where the data is at hand, an observational study. This is not to say that we should ignore any data's provenance. How the data was created and where it came from obviously becomes part of our background evidence and therefore must influence the probability statements we make about future data. Data gathered under suspicious or irregular circumstances should rightly not be fully trusted.

Common knowledge says that non-randomized trials aren't as trustworthy as randomized ones. For example, in a medical trial, a (computerized) coin is flipped as a patient walks in the door; if it is heads, he gets treatment A, else B. But what does *randomized* mean?

Data need to be "random"[4] to justify use of the classical theory, specifically in "randomized" trials. In those experiments, what we want is some mechanism to invoke that takes the decision out of human hands about how to allocate the groups in which we collect the data (like a medical trials with different treatment groups). For example, a set of "sealed envelopes"

[4]Does the piece of data itself contain this "randomness"? Do some pieces have more "randomness" than others? Can we extract it, put in a jar so to speak? Jaynes calls the old belief in randomness a "mind projection fallacy"

containing "random" numbers generated by a computer says which patient goes to which group.

You must understand that computerized coin flips, even the results from real coin flips, are not "random". The output from a "random number generator" on a computer is nothing but a deterministic sequence of numbers: if you know the starting point, you know (I mean *know*) every number in order the computer will show you. A real coin flip is constrained by the same laws of motion that were responsible for dropping the apple on Newton's head. If we knew the initial conditions of the flip (weight, air viscosity, amount of spin), we could predict exactly what the result would be (Jaynes, 2003).[5] These events—computerized numbers, actual coin flips—appear "random" because we turn a blind eye to the initial conditions and to the equations that govern the outcome. We want the outcomes to be unpredictable—they are *not* unpredictable, we just act as if they are. These acts work as "randomizers" because, even if we turned our attention to the initial conditions and equations of motion, we would never have enough time to solve them before the outcome is realized.

Here is the real reason for "randomized" trials. It is solely because you cannot *trust* human beings that "random" trials are necessary. People will lie to others and to themselves, they will cheat when able, they will maneuver, shade, and finagle, they will engage in intrigue, they will contrive and conspire, they will duck, dodge, and double-deal; in short, they will use every method under the sun to "help" the results work out the way they want them to, even if they don't think they're doing it on purpose. In an medical trial, for example, we want to take the decision of who gets what treatment out of the hands of the human, and put on to a physical device that is not easily manipulated. The reason nobody trusts the results of a study, say, touted by a homeopath is not just because his method of treatment is ludicrous, it is because he has not conducted a "randomized" trial. That is, nobody will believe that he did not pick and choose this patients so that he could get the results he wanted.

Nobody will trust real doctors or researchers either when they report results from an "observational" or non "randomized" study, not because these doctors would always purposely lie to us, but they might lie to themselves. They might have picked data that confirmed their suspicions and not sought out data that was contrary to them. Out of sheer humanity, a physician might have let a sicker patient receive the new drug rather than the control drug, and so bias the results. Or a caring researcher might seek out data that proves some injustice befell a select group, but not look for data that shows this same injustice is common to most groups, or that it has nothing to do with the groups as he categorized them, but does have to do with some other feature that was ignored.

[5]Besides, it's an easy trick to learn how to flip a coin so that it comes up Heads every time.

"Randomized" trials show up in places besides formal studies. Because people are so wily is the reason nobody would trust a referee just picking one of the teams to receive the kickoff at a football game. He is forced to flip a coin to remove the suspicion that he favors one team over the other. The coin flip *can* be manipulated and predicted, but most people believe that nobody can. This is what makes the coin flip seem "random".

If you conducted a study where people ordinarily expect you to use "randomization" but you did not, then the best you can do is to not mention you failed to "randomize"; however, you surely will be caught.

4. Surveys Polls, & Questionnaires

"Ninety-eight percent of Americans like to read about opinion polls. This result is accurate to within plus or minus four points." There are (at least) two things wrong with that statement; by the time you finish this section you should be able to find both. If you are on the ball, you should be able to find the most glaring error immediately.

A survey or poll nearly always consists of a set of fixed questions together with pre-determined responses asked of small samples of people (kind of like multiple choice exams). The results from surveys and polls are obviously statistical. Why? Because the results are never intended to be just about the sample of people polled: they are meant to apply to larger groups of humans; usually all Americans, or even all of humanity.

You can work miracles with surveys and polls. Consider asking these two questions, "Would you support a law that requires the rich to pay their fair share?" and "Would you like your marginal income tax rate to go up by 15%?" Both are meant to show how much support there is for a new bill to be passed. Both could be true representations of what that bill would do. They are spin. A Congressman and newspaper who want the new tax will commission a survey wherein respondents are asked question 1; those representing folks who want to keep their money on the other side of the aisle will counter with a survey that asks question 2. Both will announce that "Americans support my position!" You will see the results, but you will *never—never*—see the actual wording of the questions. You will only hear the inferences made from them.

The true beauty of surveys and polls is that they are infinitely flexible. You can prove support for any point of view by creating a survey. All it requires is two things: clever question writing, and wild extrapolation. There is a professional class of people called pollsters or market researchers whose entire careers are devoted to the art of imaginative question writing. It isn't too hard to do yourself, but if you have doubts, it's easy to find these firms on the internet. Tell them what you hope to prove and they will provide the questions for a fee. It goes without saying that sometimes those who commission studies want to discover the truth of a matter; for example, a

business wants to know whether a new product will sell. But even then, the firms that do the surveys know well the adage of bearing bad news...

Obviously, the firms cannot reach all Americans, or all "Europeans", or whomever. So they call up a few people on the phone (land lines, not cell phones usually), or head to the mall with clipboard in hand. This small sample (who was at home at the right time or who passed by Auntie Annie's Pretzels) is invariably said to represent fairly the entire population. Don't worry about this; nobody ever questions your sampling method, or if they do, it's easy enough to overwhelm them with technicalities (this is done by citing published works that suffer the same problems you do).

It is also imperative to ignore the above-mentioned fact that people lie. They lie like dogs and often, particularly when presented with the question, "How much money do you make?" Credit card companies which ask this question on applications know that nearly everybody claims to make over $100 thousand. Even if the question isn't so bold as "Have you recently committed tax fraud?" and is simple as "What is your age?", people will lie. They sometimes lie because of the pure pleasure of it, or they lie to help you out by giving you answers they think you want to hear. Or maybe they answer wrongly because they didn't understand what you asked. These facts, known to everybody, should decrease the certainty in all survey and poll results. Strangely, however, they never do.

In some fields, such as medicine and psychology, a *survey* goes by the glorified name of *instrument*. You do not just ask people a bunch of questions, as you do on a survey, you *administer* an *instrument*, which certainly sounds a hell of a lot more impressive, but in the end you are doing nothing more than asking a bunch of questions and hoping for the best. This is a very technical subject, but I will try to summarize adequately the main problems with a typical example. I mention no names not wanting to hurt any body's feelings.

An example. Two groups, one fat the other thin (classified by body mass index[6]), are administered an "instrument" intended to measure their "depressive status", that is, whether they are depressed. The instrument consists of several questions such as "I feel sad" to which the respondents rate on a scale from 1 to 5, higher numbers indicating stronger agreement. A score is created by, more or less, adding up the ratings across all questions. If the person has a score higher than some cut off, they are said to be depressed. Average scores for both fat and thin people are computed and a classical test is performed which, of course, gives us a p-value, which we can imagine is 0.05, and is therefore publishable. A paper is written announcing "Thin people suffer more depression than fat ones." What is the probability that this proposition (call it S) is true? We already know that is has nothing to do with the p-value (and we know that no proposition has a probability

[6]An imperfect measure of fatness. Calculated by weight (in kg) divided by height (in m) squared.

independent of some evidence, i.e. some "given"). But ignore that problem and let's think about the data itself.

First is that the questionnaire—the instrument, I mean—is said to measure depression. Does it measure it exactly? Nobody makes this claim for any instrument, nobody claims that instruments exactly measures the thing it purports to; unless that thing is a real, physical entity, like weight. But even though everybody knows this, not everybody remembers this at all times. Too, psychiatrists and psychologists will not always agree whether a given patient is in fact depressed. So here are two sources of error: (1) the instrument does not and cannot measure depression exactly, and (2) depression itself is hard to define. These two sources of error have to be incorporated into our probability of S. Error (1) is usually large, (2) is smaller, but is not negligible.

It is the case that people, if given the instrument twice, will not answer in the same way. Their internal state might have changed between the times between administration or they might just answer differently because they do not think too much or cannot recall their previous answers (what's the difference between a "4" and a"5" on the question "I feel sad"?). They also might mark incorrectly. These possibilities provides two more sources of error, (3) the internal states of people might fluctuate too rapidly to be of use, and (4) inconsistency in answers. Error (3) is probably small or even negligible, but error (4) is not and it is well known not to be. These further make the probability of S less certain.

Any more? Well, people lie, they either don't want people to know the truth or they angle it towards what people want to hear; that's error source (5). The size of this error is usually unknown, but, ever hopeful, people assume it is near zero. Another source: not only might people not be able to distinguish between "4" and "5" on the scale, they might not know what you are asking. For example, one instrument asks something like "I feel blue", which surely depends on cultural information not possessed by all. Confusion about the questions is error source (6). This source is generally acknowledged.

So how to cheat? Well, same way a lot of people do. Just do not mention or downplay the sources of error. Ignore or dismiss them. This allows you to claim your results are far more certain than they truly are. People will see your small p-value and assume S is very probable, or even true.

Actually, you *can* mention the sources of error. People will nod their heads when they read your caveats and know you are being intellectually honest. But don't sweat it, either. This is because, even though you mention the (at least six) sources of error, you will not have to incorporate them into your p-value calculation. This is an important fact: everybody, even though they might discuss limitations, they all ignore the error when computing their final statistics. Thus, the p-values, or posterior distributions of the parameters, or even the predictive distributions of the observables will all be too sure of themselves. Nobody will ever tag you for leaving out the error

because nobody wants to give up on these paper-generating questionnaires. If a referee for your paper questions your validated (which usually means you gave your instrument in at least two samples and got similar answers) instrument because of the errors mentioned above, it means he would have to question and give up on his own. And nobody wants to do that.

Finally, you can easily create your own instrument, but it's far easier to use a well-established one on a new source of data. The vast number of previously-published studies that use that instrument give weight to the idea that this is a reasonable thing to do.

5. Publishable p-values

Most journals, say in medicine or those serving fields ending with "ology", are slaves to p-values. Papers have a difficult, if not impossible, time getting published unless authors can demonstrate for their study a p-value that is publishable, that is, that is less than 0.05. Sometimes, the data are not cooperative and the p-value that you get from using a common statistic is too large to see the light of print. This is bad news, because if you are an academic, you must publish papers else you can't get grants, and if you don't get grants, then you do not bring money into your university, and if you don't bring money into your university, the Dean is unhappy, and if they Dean is unhappy you do not get tenure, and if you do not get tenure, then you are out the door and you feel shame.

So small p-values are important. I of course advise against using classical statistics methods, but if you are forced to (and some journal editors insist on it[7]), then all is not lost if an initial large p-value is found. In fact, I would go so far to say that if you cannot find a publishable p-value in any situation, then you are not trying hard enough. There are several ways to lower your p-value.

The most well known is *to increase your sample size*. This one is a lock. Let's take a look at the t-test statistic from Chapter 10 to see why.

$$t(x) = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

There is a mathematical phrase that begins "without loss of generality" which I now invoke by letting, for ease of notation, $n_A = n_B = n$ and $s_A^2 = s_B^2 = s^2$, so that $t(x)$ becomes

$$t(x) = \sqrt{n}\frac{(\overline{x}_A - \overline{x}_B)}{s}$$

Remember that we want a large statistic, a large t, the larger the better, because larger ts mean smaller p-values. Do you see the trick? A larger n means a larger t! All you have to do is to increase your sample size and just

[7] I often get referee and editor comments either saying they do not understand the modern statistical methods so they are inappropriate, or could they please also have the p-value. I am not kidding.

wait for the small p-values to start rolling in. This trick *always* works in any classical situation, even when the difference $\overline{x}_A - \overline{x}_B$ is too small to be of interest to anybody. This is why having a small p-value is called attaining *statistical* significance and not practical, or useful, or clinical significance.

Incidentally, this trick also works in Bayesian statistics in the sense that, with large samples, the posterior distribution of $\mu_A - \mu_B$ will have most probability above or below zero. But it fails miserably in modern observable statistics because a trivial difference in $\mu_A - \mu_B$ won't make a tinker's dam worth of difference in the probability distribution of future observables. Your model will not have skill, of example.

The next trick, if you cannot increase your sample size, is to *change your statistic*! This comes from the useful loophole in classical theory that there is no rule which specifies which statistic you can use in any situation. Thus, though some creativity and willingness to spend time with your statistical software, you can create small p-values where others see only despair. This isn't so easy to do in `R` because you have to know the names of the alternate statistics, but it's cake in software like `SAS`, which usually prints out dozens of statistics in standard situations, which is one reason `SAS` is worth its exorbitant price. Look around at the advertising brochures of statistical software and you will see that the openly boast of the large number of tests on offer.

For example, for use in "testing differences between proportions", just off the top of my head I can think of the z statistic, the proportions test with and without correction for continuity (two or three to choose from here), χ^2 test, Fisher's exact test, McNemar's test, logistic regression. There are dozens more and teams of academic statisticians constantly add to the pile. Don't believe it? Here's a small table of these tests for the TSD/Sex data from Chapter 11.

Test	p-value
Prop test	0.78
Fisher's	0.70
Logistic Reg.	0.52
χ^2	0.50
z test	0.49
McNemar's	0.24

Because I was only able to get to 0.24 just means I didn't try hard enough. Which is *the* correct p-value? They *all* are! That's the beauty of this trick. It's thrilling! Not one of these p-values is more "right" than any other one. Each is valid: there is no way to prove which is best. If all you know is classical statistics, let this knowledge sink in. It should prove to you that p-values are not what you probably thought they were.

For "testing differences between means", there is the t-test (a couple of versions of this, actually), Wilcox test (also called Mann-Whitney), sign

tests, Spearman correlation tests, Kendall's τ, Kruskal-Wallis test, Kolmogorov-Smirnov test, permutation test, Friedman two-way analysis of variance—I'm running out of breath—and many more. Here's some of those tests for the advertising data:

Test	p-value
Spearman	0.87
Perm.	0.20
t-test	0.19
Wilcox	0.14
Kol.-Smi.	0.08

Nearly there!

Please remember that in this example, like the previous one, the *data is the same*; the only thing that changes is that classical statistical test.

The key to this deceit is to never admit what you did. When it comes time to write up your result, boldly and authoritatively state, "We used Johnston's (Johnston, 1983) frammilax test for differences in means." Tossing in a citation always cows potential critics; tossing in two or more guarantees editorial acquiescence. Do not tell the reader that you went through a dozen tests to find the lowest p-value. Act as if "Johnston's test" was what you had in mind all along.

Think this doesn't happen and old Briggs is exaggerating? Then turn to the 16 August 2008 *Wall Street Journal* (WSJ) article by Gabe Thornhill wherein he describes the "controversy" over a study Boston Scientific published about their new stent, the Taxus Liberte. Boston Scientific, testing whether its stent and a standard one were equivalent, used the Wald Test and got the magic p-value of 0.0487. Less than 0.05! Investors could now invest with confidence. Well, somebody—a competitor?—didn't like this and said this p-value was too close to the magic level. The WSJ actually got the data and found that using "a number of other methods of calculation—including 14 available in off-the-shelf software programs— the Liberte study would have been a failure by the common standards of statistical significance in research." In other words, they tried 13 other tests (many of which are different from the ones I listed above—there are so many!) and all of those tests gave p-values *larger* than 0.05. The WSJ even solicited help from some very prominent statisticians who agreed with their findings. Poor Boston Scientific!

But do you know, after all this hoopla, the *largest* p-value the WSJ was able to find was only 0.0547, surely not an insanely large departure from the publishable limit—which shows you how silly the whole story is, and how irrational people are when it comes to making decisions based on p-values

This cheating technique is unavailable in Bayesian or observable statistics. True, you can change your default prior distribution on the parameters or even change the model (see below), but editors in most fields are still suspicious of modern methods and tend to be conservative and will likely

insist on a well-known default. There will be more room for creativity in, say, ten years when modern methods become familiar.

Our last option, if you cannot lower your p-value any other way, is to change what is accepted as publishable. So, instead of a p-value of 0.05, use 0.10 and just state that this is the level you consider as statistically significant. I haven't seen any other number besides 0.10, however, so if your p-value is larger than this the best you can do is to claim that your results are "suggestive" or "in the expected direction." Don't scoff, because this sometimes works. You can really only get away with this in secondary and tertiary journals (which luckily are increasing in number) or in certain fields where the standard of evidence is low, or when your finding is one which people want to be true. This worked for second-hand smoking studies, for example, and currently works for anything said to be negatively associated with global warming.

6. Expand Your Data

Here is ancient wisdom:

Seek and ye shall find.

Nowhere does this better apply than in data analysis. Sometimes, despite all your efforts, you can not find a way to produce a publishable p-value with a given set of data. You tried all the tricks above, you can't increase the sample size and all the classical tests under the sun bring no joy. What to do? Increase your data! No, I don't mean increase the sample size, but increase the data on which you are making tests.

Everybody is constrained by time at least, but by budget usually, which puts a cramp on the sure-fire method of increasing sample size to get a small p-value. Well, friends, I am here to tell you, you can leave your "significant" p-value set at 0.05, leave your sample size as it is. You can still find a significant result with the method of *multiple testing*. This one requires a little more planning because you have to think of it before you start collecting data. For example, in the TSD example, don't just collect the fact that there were men and women; also observe the age, the weight, the race, day of the week, hour of the day, whether the person wore jeans, or a hat; stop and ask the people their income, their political party, their views on this and on that, whether the day was sunny, whether it was raining, the traffic density, and any other thing you can imagine. The only trick is to record as many different things as possible. Five is too few, fifteen is better, a hundred or more is practically a guarantee. I once was the statistician on a study that collected over 5000 items per person! I promise this is true. It was a medical study, wherein everything in a patient's chart was recorded, not once, but five to six times over a period of time, plus the individual questions from several "instruments", some homemade, some "validated" (which means more than one person in print used it).

Your main interest, in the TSD example, is still whether or not there is a difference between men and women, which we have already seen only gets our p-value (after trying several tests) to a non-publishable 0.24. The next thing to try is *sub-group analysis*. See if there is a difference between men and women on just the sunny days or the cloudy, or when the traffic density was high or when it was low, or whether it was a weekday or weekend, and on every other possible cut of the other variables. Race is always popular. One of these differences is bound to give you a publishable p-value.

Statisticians are on to this one, so be careful in how you describe your results. Whatever you do, do *not* say you tried every possible combination. You will be busted. Some statistician will immediately point out that you should have used so-and-so's method of correcting for multiple testing (the result of which is to inflate all your p-values). So be daring and just state, "Our results indicate that among poor Hispanics, more men than women wear TSDs" and nobody will ever question you, especially if you mention a disadvantaged group (e.g. the poor). Try to angle your writing towards the idea that this subgroup was your main interest all along.

It can still happen that, even after exhaustive efforts, you still cannot find a difference between men and women in any of the subgroups. This kind of thing is rare, and its more likely you will have got bored of looking than there isn't a statistically significant result lurking somewhere. Can you guess what to do next? Right! Abandon the quest to find differences between men and women and simply find a difference between some other group; sunny and cloudy days, or whatever. If you have collected enough data, you simply cannot go wrong.

7. Models

Suppose you ran a classical regression (the `glm` model) and found that some of the coefficients of interest did not have a small enough p-value. You can try the tricks above, but you could also scan through the data itself to make sure that nothing is causing problems.

It happens that sometimes in your data an exceptionally large or small value appears. Statisticians call these *outliers*. I don't mean bad data values that arise from, say, bad typing, or by transposition, transcription, or some other honest mistaken entry. You'll find those when you look through the data and remove them anyway. No—what I mean are large or small values that are real, that were really measured, but that stick out and which cause your model to go astray. This happens a lot with medical and economic data where the use of the normal distribution to quantify uncertainty is ubiquitous. Very large and small data values show up all the time. What to do? Smack the label *outlier* on those extreme values, and then shun them, by which I mean, toss them out. Recompute your model after this and you will usually find improvement (smaller p-values). I have seen this done on my presence on more than one occasion.

What's an outlier? A piece of data that does not fit your expectations. With surgical precision, then, you can cut out any offending variable so that, in the end, your data will act just like you wanted it to. Your chosen model will now fit. Of course, you will have learned nothing new, you will merely have reinforced your preconceptions, but that is always a comfort, isn't it?

Gottfried Leibniz, co-discover of calculus, said this

> Let us suppose for example that some one jots down a quantity of points upon a sheet of paper helter skelter, as of those who exercise the ridiculous art of Geomancy; now I say that it is possible to find a geometrical line whose concept shall be uniform and constant, that is, in accordance with a certain formula, and which line at the same time shall pass through all of those points...Leibniz (2005)

What the old boy is saying is that it is always possible, given *any* set of data, to find a model that fits those data to any level of exactness, even perfectly, even if that data is completely arbitrary. The implication, of course, is that if you can't find a model that fits your data well, then you aren't trying.

In linear models, such as regression, it's easy to find good fits. You have n data points (different people, say). One variable you want to predict, the remaining variables help you predict it. There are p of these. If you read the section above, you know you want p to be a big number. A well known trick is to let p get close to n in size. If $p = n$ then, with a regression model, you will meet Leibniz's criterion exactly, such that you will have found a (p-dimensional) line that goes through each data point perfectly. Now, chances are that if you do this the p-values on the coefficients will likely not be publishable, so you have to change strategy. Do not tout p-values, trumpet your model fit. I earlier skipped over the measure R^2 (which you can get from running `lm` instead of `glm`, which gives you AIC instead; I skipped these measures because they don't take the uncertainty of your model's guesses into account, which here works in your favor). The highest and best R^2 can be is 1 and the lowest and worst is 0. If $p = n$ your R^2 will equal 1, no matter what set of data you have! Obviously, you cannot report an $R^2 = 1$. This is like a psychic reading your mind exactly. People would be suspicious that a fast one is being pulled.

Take a few variables out of your model and report a modest, say, $R^2 = 0.6$, which sounds low, but believe me, it is not. Some fields would celebrate a value this high (these fields, which shall remain nameless, routinely see R^2s in the 0.1 to 0.2 range). You can try this with regression models, and it's an OK trick, but if you do people usually get curious about the p-values on the coefficients, which is an annoyance because we know these won't pass inspection. To get around this, skip regression and move on to what are called *latent variable* models. These go by names like *path* and *factor* analysis.

The way these work is that you have an observable y and a bunch of observable xs which are used to help explain y. So far, this is the introduction to regression, which isn't that exciting. Now here's the beautiful part. What you do is to pretend that there are a series of hidden, unobservable or *latent* variables $\alpha_1, \alpha_2, \ldots, \alpha_q$ that lie between x and y. Since x has p different variables, and there is no limitation on how many hidden variables that can lie between each x_j and y, and how many different paths can be between each of the α_i (yes! they can be connected too!), you have an inexhaustible supply of models. Ready for the best part? These are usually used just to find something like high R^2, the influence of the observable coefficients (hence, their p-values) are deemphasized. You can report on the p-values of the unobservable latent variables instead!

Besides the usually academic specialty suspects, these kinds of models find favor in marketing. Latent variable models go by the name "neural nets" among the highly computer literate (there are differences in internal structure between neural nets and other latent variable models, but they all share the idea that hidden forces are at work).

8. Sleight of hand

I obviously cannot teach you every possible way to turn leaden data into gold (peer-reviewed) published papers. Statistics is too big a field and the number of methods is huge and ever growing. The techniques I have given are the easiest and most reliable and you can nearly always get away with them as long as you are careful about your language.

As I mentioned above, boldness is imperative. Simply write your results as if the findings where what you were looking for and expected all along. If you have to use some obscure classical test or model, be sure to include at least two references that show that other papers (they don't have to be in your field) have used them. People, especially journal editors, dislike novelty. Reassure them that what you are doing *everybody* does.

Vagueness is ever useful. Do not confess all the steps you had to go through to get your desired finding. Let people think you are as honest as they are.

At the worst, if all else fails, then at least claim that your results are *suggestive* or *in the direction* one expects if your beloved theory were true.

9. Homework

(1) Find one use of statistics in an advertisement. Print is best, because you can just clip it out or photocopy and hand it in. If it's television or advertising, try to tape it and then copy down *exactly* what was said and done, and *exactly* where and under what circumstances you heard or saw it (what channel, time, web page, etc. etc.). Investigate how the copywriters might have cheated.

(2) Crack open a journal in a field which routinely uses statistical methods on human data. Find an example of a paper that might have—just *might*

have, I emphasize (don't accuse anybody unless you're sure you can get away with it)—used some of the creative techniques of statistical analysis mentioned in this chapter.

(3) As a group, redo the thinking suppression device data collection, this time making note of as many other variables as possible. Each member of the group can then take different slices (sunny vs. cloudy days, etc.) and compute the standard tests. Each person should write up his results to paint them in the best light. See if it fools your classmates.

(4) Try Leibniz's Geomancy example. Take out a piece of paper and draw a cross on it right down the center. This will be our standard mathematical axes. Get a pen, close your eyes, and stab at the paper at least a dozen times. Open your eyes (if I didn't specify this step, some clown would insist that he couldn't perform the remaining steps) and stare at the points. Try to find a smooth curve that passes through most of the points. Come up with a story for your curve.

(5) A more ambitious class project is to do a survey. Pick a topic which might have some controversy. Split the class in two. One side writes a question (or two) that seeks to gain support for the topic; the other sides seeks to gain dissent. The questions should be as fair as those found in any survey. Both sides should not reveal their questions until after all the results are in and analyzed. The main goal is to see how much you can get away with. Go to the mall with a clipboard and collect people's sex, age, college status (i.e. freshman, etc.), birthcountry (USA or not), and anything else you can think of. Particularly ask about income. The answers you get from this question should forever after instill skepticism in any survey which claims its results are influenced by income (rich or poor, and so on). In the end, write up the results and only then reveal to the other side what you have done. Contact me at matt@wmbriggs.com and let me know how it went.

CHAPTER 15

The final chapter

1. What is statistics?

The state of statistics is in somewhat of an odd place. The actual practice of statistics—as opposed to its theory—has evolved into procedures that make people more certain than they should be. This is because the methods of classical statistics have been designed, or co-opted, to do the thinking for you. Now, there are always careful statisticians out there, even those who exclusively practice classical statistics, who routinely warn against the excess of confidence, but their warnings never seem to stick. Pick up almost any academic journal and you will find examples of "statistically significant" results and conclusions presented with authority.

The practices that engender over-confidence need to change. This is why I have been so insistent that you think about observable, verifiable data. Any mistakes or over-confidence in your statistical procedures will become quickly and glaringly obvious if observable data is your focus. This is not so using the old ways. Let's look at the outline of the old and the new and contrast them.

Here are the steps in how classical statistical problems are handled:

(1) Start with quantifying your uncertainty in an *observable* using a probability distribution.
(2) The distribution will have unobservable parameters which *you do not know* but which must be specified.
(3) Collect observable data.
(4) Look up formula in cook book and use this to make a guess at the parameter's values.
(5) Compute your p-value. If it is less than the magic numbers, state the result is "significant".

Sometimes people substitute confidence intervals for p-values, but we know that these give identical answers. The problem is obvious. The focus is on the unobservable parameter. A guess is made of it and if the p-value is publishable, the words " statistically significant" will be read by most to mean "true", as in "My theory or hypothesis is true." In this system, no matter in what circumstance or type of problem, you will be too sure of yourself in the end.

Now here is a guide to how statistics problems should be solved:

(1) Start with quantifying your uncertainty in an *observable* using a probability distribution.
(2) The distribution will have unobservable parameters which *you do not know* but which must be specified.
(3) Quantify your uncertainty in these parameters using probability distributions.
(4) Collect logical evidence and observable data, which provide updated information about the parameters which you still do not know and which still have to be quantified with probability distributions.
(5) Since you do not care about the parameters, and you do care about *future observables*, quantify your uncertainty in these future observables by accounting for the uncertainty you still have in the parameters.
(6) Check to see if your model has skill, that is, that it can beat a "model" that just guesses.
(7) If your model does have skill, use it to make statements and conclusions about future observable data.

If you stop at step 4, you are a typical Bayesian (subjective, usually). If you stop at step 2 and branch off to make statements about functions of observable data assuming that certain details about the parameters are true, you are a frequentist. By far, the vast majority of people who practice statistics are frequentists. A growing number of people will use Bayesian methods to say something about the parameters. Shockingly few people go all the way to make statements about actual observables.

The beginning part of Step 4 should have more emphasis when it comes time to write up the results. Medical papers usually do a good job with this. The process even has the informal name "Table 1," which is found in most papers. It is a fairly complete description of the demographics of the population. This is necessary because, remember, the probabilities for the future observables are for data that "looks like" the old data you collected. This warrants a heavier emphasis than is normally given on the description of your sample. What are the characteristics of your sample? This is discussed in detail below.

If your data has anything to do with humans you have to keep in mind that your old data exists not only in space but in time. Don't just read a sentence like that and go on with your business. Think about it. How far into the future are your results expected to hold? Days, weeks, years? If history has proven anything, it has shown that people are rotten at predicting the future, while conversely still believing they are good at it. Your results are also probably much more limited than you first thought.

2. Randomized trials

As we previously discussed, there is a near religious belief in "randomized trials." Here is an example of one (I met a prominent physician at a cigar lounge who suggested this[1]). Come to New York City and on Tuesday walk across 5th Avenue in the middle of the block at 8:45 am. I will "randomize" participants to either close their eyes or leave them open. Everybody that makes it to the other side we will label "survivors". We want to test whether those with their eyes open survive more than do those with their eyes closed. Ok, everybody who volunteers to participate in this "randomized controlled trial" raise their hands? According to federal guidelines, we cannot know what the answer to this experiment would be until we actually ran it. Your intuition is not good enough! So if I don't get enough volunteers, I'll have to start picking people. You might think of offering some lame excuse about "danger" or "traffic". But your intuition from which these objections arise counts for nothing. According to classical statistics—and the "scientific method"—there is no way to know what the outcome of this experiment will be unless we actually ran the "trial."

Gordon Smith and Jill Pell, in the *British Medical Journal*[2], wrote a paper called "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomized controlled trials." They pointed out that people are using parachutes without any statistically significant evidence that they are effective. They said, "We were unable to identify any randomized controlled trials of parachute intervention." Their conclusion:

> As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evaluation by using randomized controlled trials. Advocates of evidence based medicine have criticised the adoption of interventions evaluated by using only observational data. We think that everyone might benefit if the most radical protagonists of evidence based medicine organised and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute.

That is, they want to run a trial similar to my doctor's 5th Avenue experiment. Some will be "randomized" to receive parachutes, others not. At the end we'll compute a Fisher's exact test on the proportion of survivors.

Incidentally, the decision you make not to participate in either of these trials is based on an inductive argument (how?), which as we all remember is forbidden thinking in classical statistics. In many fields, like physics, meteorology, chemistry, parachute studies, and pedestrian crosswalk-ology most experiments are not "randomized". They are controlled. All the variables

[1] Which is evidence that smoking is good for you.

[2] 2003, volume 327, pp 1459–1461

that might influence the outcome are controlled for as carefully as possible. Experimenters in these fields do make mistakes, of course, but they are rare. It is also case that in physical experiments people have the *luxury* of control, whereas in experiments with people the opportunities for control are limited.

One example of a human-centered partially-controlled experiment is a drug trial. We can control who gets what drug. Sometimes, we can further control for the number of men and women who get it, but that's less common. We should be controlling for many physically or biologically measurable quantities, like sex or weight, but we often do not. We instead rely on "randomization" to equal out the differences between groups in trials. Obviously, there is no way to control for everything (there is an infinite amount of everything; see the homework), but we should be able to control for the items most likely to be associated with the outcome. This means that we have to use extra-data evidence, such as historical data, other deductive and inductive arguments, and intuition.

There is a general distrust in intuition (intuition also contributed to your refusal to enlist in the above trials). Now, there is some evidence that in certain experimental situations (some of which are highly artificial), people's intuition can mislead them. And it is certainly true that as circumstances become increasingly complex (such as in individual or group human behavior), intuition can often lead you astray (Gilovich et al., 2002). But because intuition sometimes misleads us does not mean that it always does, especially in circumstances which are simple. Like looking both ways before crossing the road. See Kadane (1996) for examples of how intuition, and historical and other evidence can be used in clinical trials.

3. Parameters and Observables

I have repeatedly emphasized that we are interested in quantifying uncertainty in *observables*, which are, for the most part, real physical entities. I have also said that our lone goal should almost never be to make statements about *unobservable parameters*. Probability statements inferred from data about parameters will always be more certain than statements about what we can actually see and test.

Sometimes, though, all you can do is to make a decision about the value of something you cannot see because it is impossible to observe the observable. The most prominent example is a jury trial. Did Tom kill Wendy? Tom has pleaded innocent, the prosecutor presents damning evidence about the strange relationship Tom has with his dog, the defense lawyer points out Tom's sweet smile and facility with FORTRAN. In the end, the jury has to compute this probability

$$\Pr(\text{Tom is guilty}|\text{Evidence presented})$$

The statement "Tom is guilty" is a *parameter* in the sense that it cannot be observed, yet the jury still has to estimate a (non-numerical) probability

"beyond reasonable doubt" in order to make the decision that Tom is guilty. In most trials, the jury will never *know* that it made the correct decision because the event can never be observed.

Other examples of impossible to observe events are *counterfactuals*. Here is a common example of one:

$$\Pr(\text{Germany wins WWII}|\text{Hitler does not invade Russia}).$$

Many people have argued that the statement "Germany wins WWII" given the evidence above has a high probability of being true. Now, it is certain that the proposition "Hitler does not invade Russia" is false since in fact Hitler duplicated Napoleon's folly. It is no difficulty for logical probability to give estimates for counterfactuals, however, since probability statements are matters of logic. My favorite example is from the philosopher David Stove (1986)

$$\Pr(\text{Bob is a horse}|\text{Bob is a winged horse}) = 1.$$

Obviously, the probability (given our experience) of "Bob is a winged horse" is 0, but this is no bar to making the logical statement that *if*, in fact, Bob was a winged horse he is also certainly a horse.

This is a very important point in the favor of logical probability because counterfactuals are everywhere in human affairs and decision making. Some examples: "She would have got better had the doctor not administered the drug", "I would have had a better day if I didn't hit the snooze alarm", "I'd would have gone if only my mom would have let me" and on and on.

Incidentally, classical probability and statistics cannot deal with any of these situations. Neither is most of the mathematical apparatus of modern probability equipped to handle unobservable events and counterfactuals. This very naturally leads us to our next topic.

4. Not all uncertainty can be quantified

It is, among the more mathematical of statistical and economical circles, somewhat of a controversial statement to say that not all probability and risk can be quantified. However, it is true and easily proved.

Let some evidence we have collected—never mind how—be E = "Most people enjoy Butterfingers". We are interested in answering the truth of this statement: A = "Joe enjoys Butterfingers." We do not know whether A is true or false, and so we will quantify our uncertainty in A using probability, that is written like $\Pr(A|E)$, and which reads "The probability that A is true *given* the evidence E". If we can recall Chapter 1, this should all be review.

In English, the word *most* usually means *more than half*; it could even mean *a lot more than a half*, or even *nearly all*; in some cases it merely means a plurality[3], in which case we can change the evidence to E = "At least half of all people enjoy Butterfingers." There is certainly ambiguity in

[3]Thanks to my friend Raphael for pointing this out.

its definition. But since *most* usually means *more than half*, we can partially answer our question, which is written like this

$$0.5 < \Pr(\mathrm{A}|\mathrm{E}) < 1 \tag{40}$$

and which reads "The probability that A is true is greater than a half but not certain *given* the evidence E." This answer is the best we can do with the given evidence. This answer is a quantification of sorts, but it is not a direct quantification like, say, the answer "The probability that A is true is 0.673" which we always get using the computer.

It is because there is ambiguity in the evidence that we cannot completely quantify the uncertainty in A, no matter what the computer tells us. That is, the inability to articulate the precise definition of "most people" is the reason we cannot exactly quantify the probability of A.

The first person to recognize this, to my knowledge, was John Maynard Keynes is his gorgeous, but now little read, *A Treatise on Probability* (2004), a book which argued that all probability statements were statements of logic. To Keynes—and to us—all probability is conditional; you cannot have a probability of A, but you can have a probability of A with respect to certain evidence. Change the evidence and change the probability of A. Stating a probability of A unconditional on any evidence disconnects that statement from reality, so to speak. All this we have learned so far.

For many reasons, Keynes's eminently sensible idea never caught on and instead, around the same time his book was published, probability theory bifurcated into two antithetical paths. The first, as we know, was called *frequentism*: probability was defined to be that number which is the ratio of experiments in which A will be true divided by the total numbers of experiments as that number of experiments goes to infinity.[4] This definition makes it *difficult* (an academic word meaning *impossible*) to answer what is the probability that *Joe*, our Joe, likes Butterfingers. It also makes it *difficult* to define the probability for any event or events that are constrained to occur less than an infinite number of times (so far, this is all events that I know of).

The second branch was *subjective Bayesianism*. To this group, all probabilities are experiences, feelings that give rise to numbers which are the results of bets you make with yourself or against Mother Nature (nobody makes bets with God anymore). To get the probability of A you poll your inner self, first wondering how you'd feel if A were true, then how you'd feel if A were false. The sort of ratio, or cut point, where you would feel equally good or bad becomes the probability. Subjective Bayesianism, then, was a perfect philosophy of probability for the twentieth century. It spread like mad starting in the late 1970s and still holds sway today; it is even

[4]Another, common, way to say infinity is the euphemism "in the long run." Keynes has famously said that "In the long run we shall all be dead." It's always been surprising to me that the same people who giggle at this quip ignore its force.

gaining ground on frequentism. In its favor, it should be noted that, after we get past the bare axioms and talk of "feelings", the math of subjective Bayesianism and logical probability is the same.

What both of these views have in common is the belief that any statement can be given a precise, quantifiable probability. Frequentism does so by assuming that there always exists a class of events—which is to say, hard data—to which you can compare the A before you. Subjective Bayesianism, as we have seen, can always pull probabilities for any A out of thin air. In every conceivable field, journal articles using these techniques are multiplying. It doesn't help that the many times probability estimates are offered in learned publications, they are written in dense mathematical script. Anything that looks so complicated *must* be right!

The problem is not that the mathematical theories are wrong; they almost never are. But because the math is right does not imply that it is applicable to any real-world problems. We have already talked about how normal distributions are used too often. The math often is applicable, of course; usually for simple problems and in small cases the results of which would not be in much dispute even without the use of probability and statistics. Take, for example, a medical trial with two drugs, D and P, given to equal numbers of patients for an explicitly definable disease that is either absent or present. As long as no cheating took place and the two groups of patients balanced, then if more patients got better using drug D, that drug is probably better. In fact, just knowing that drug D performed better (and no cheating and balance) is evidence enough for a rational person to prefer D over P.

All that probability can do for you in cases like this is to clean up the estimates of how much better D might be than P in new groups of patients. As long as no cheating took place and the patients were balanced and typical, the textbook methods will give you reasonable answers. But suppose the disease the drugs treat is not simply defined and the patients aren't "typical" or balanced. Let's write what we just said in mathematical notation so that certain elements become obvious.

$$\Pr(\text{D} > \text{P} | \text{Trial Results} \quad \& \quad \text{No Cheating} \; \& \; \text{Patients Like Before}) > 0.5. \tag{41}$$

This reads, the probability that somebody gets better using drug D rather than P *given* the raw numbers we had from the old trial (including the old patient characteristics) *and* that no cheating took place in that trial *and* the new patients who will use the drugs "look like" the patients from the previous trial, is greater than 50% (and less than certain).

Now you can see why I repeatedly emphasized that part of the evidence that usually gets no emphasis: no cheating and patients "like" before. Suppose the outcome of applying a sophisticated probability algorithm gave us

the estimate of 0.728 for equation (41). Does writing this number more precisely help if you suppose you are the doctor who has to prescribe either D or P? Assume that no cheating took place in the old trial, then drug D is better if the patient in front of you is "like" the patients from the old trial. What is the probability she is so (given the information from the old trial)?

The word *like* is positively loaded with ambiguity. It is of utmost importance that we write out the last question mathematically:

(42) $\Pr(\text{My patient like others} | \text{Characteristics from previous trial})$

The reason to be verbose in writing out the probability conditions like this is that it puts the matter starkly. It forces you, unlike the old ways of frequentism and subjective Bayesianism, to specify as completely as possible the circumstances that form your estimate. Since all probability is conditional, it should always be written as such so that it is always seen as such. This is necessary because it is not just the probability from equation (41) that is important, equation (42) is, too. If you are the doctor, you do not—you *should* not—focus solely on probability (41) because what you really want is this:

$$\begin{aligned} &\Pr(\text{D} > \text{P \& My patient like before} | \\ &\qquad \text{Trial Results \& No Cheating \& Patients' Character}) \end{aligned} \tag{43}$$

which is just (41)×(42). I am in no way arguing that we should abandon formal statistics which produces quantifications like equation (41), i.e. the very ones we have been using in this book. But I am saying that since, as we already know, exactly quantifying (42) is nearly impossible, we will be too confident of any decisions we make if we, as is common, substitute probability (41) for (43) because, not matter what, the probability of (43) is always less than the probability of (41).

Appropriate caveats and exceptions are usually delineated in journal articles when using the old methods, but the results are buried in the text, which causes them to be weighed more or less importantly, and which give the reader a false sense of security. Because, in the end, they are left with the suitably highlighted number from equation (41), that comforting exact quantification reached by implementing impressive mathematical methods. That final number, which we can now see is not final at all, is tangible, and is held on to doggedly. All the evidence to the right of the bar is forgotten or downplayed because it is difficult to keep in mind.

The result to equation (41) is produced, too, only from the "hard data" of the trial, the actual physical measurements from the patients. These numbers have the happy property that they can be put into spreadsheets and databases. They are real. So real that their importance is magnified far beyond their capacity to provide all the answers. They fool people into thinking

that equation (41) is the final answer, which it never is. It is always equation (43) that is important to making new decisions. Sometimes, in simple physical cases—like blocks on inclined planes but not like crops in blocks of fields—probabilities (41) and (43) are so close as to be practically equal; but when the situation is complex, as it always is when involving humans, these two probabilities are not close (Gill, 2004).

The situation is actually even worse than what we have discussed so far. Probability models, the kind that spit out equation (41), are fit to the "hard data" at hand. The models that are chosen are usually picked because of habit and familiarity, but responsible practitioners also choose the models so that they fit the old data well. This is certainly a rational thing to do. The problem is that, since probability models are only designed to say something about *future* data, the *old* data does not always encompass everything that can happen and so we are limited in what we can say about the future. All we can say for certain is what has happened before might happen again. But it's any body's guess whether what *hasn't* happened before might happen in the future.

The probability models fit the *old* data well, but nobody can ever know how well they will fit *future* data. The result is that over reliance on "hard data" means that probabilities of extreme events are underestimated and mundane events overestimated. The simple way to state this is the system is built to engender over-confidence. Memorize this and apply it every result you compute or that you read from somebody else.

5. Decision Analysis

You're still the doctor and you still have to prescribe D or P (or nothing). No matter what you prescribe something will happen to the patient. What? And when? Perhaps the malady clears up, but how soon? Maybe the illness is merely mitigated, but by how much? You not only have to figure out what treatment is better, but what will happen if you apply that treatment. This is a very tricky business, and is why, incidentally, there is such a variance in the ability of doctors.[5] Part of the problem is explicitly defining what is meant be "the patient improves." There is ambiguity in that word *improve*, in what will happen with either of the drugs is administered.

Jinnah Mohammed[6] reminds us that even if the probability of D being better than P is huge (or there is a "statistically significant" difference) this is *in no way tells us the probability a patient gets better using D (or P)*. It might be that, in the trial of 1000 patients each of D and P, 10 got better under D, and none under P. D is better, but D is certainly no miracle cure.

[5]A whole new field of medicine has emerged to deal with this topic. It is called *evidence-based medicine*. Sounds good, no? What could be wrong with evidence? And its not entirely a bad idea, but there is an over reliance on the "hard data" and a belief that only this hard data can answer questions. We have already seen that this cannot be the case.

[6]LivingManicDepressive.com

There are two separate questions here: (1) defining events and estimating their probability of occurring and (2) estimating what will happen given those events occur. Going through both of the steps is called computing a *risk* or *decision analysis*. This is an enormously broad subject which we won't do more than touch on, only to show where more uncertainty comes in (for applications in economics, see Lancaster, 2004).

We have already seen that there is ambiguity in computing the probability of events. The more complex these events the more imprecise the estimate. It is also often the case that part (2) of the risk analysis is the most difficult. The events themselves cannot be articulated, either completely or unambiguously. In simple physical systems they often can be, of course, but in complex ones like the climate or ecosystems they are not. Anything involving humans is automatically complex.

What the older statistical methods and the strict reliance on hard data and fancy mathematics have done is to create a system where there is too much certainty when making conclusions about complex events. We should all, always, take any result and realize that it is conditional on everything being just so. We should realize those just so conditions that obtained in the past might not hold in the future. We should be far less certain than we are.

6. Homework

(1) Write down five counterfactuals, being explicit about the evidence, and estimate the probability of them being true. There is no need, and might even be impossible, to give a precise number to these estimates.

(2) Write down five statements and evidence such that the statements cannot be given precise probabilities. Recall Chapter 1 where the evidence was "M might happen" and we wanted to quantify the probability of "M will happen", the best we could do is to say that the probability was greater than 0 but less than 1.

(3) What is the probability that D = "The next child born will be a genius"? Be explicit in your list of premises/evidence. Is your final answer too specific, not specific enough?

(4) EXTRA Imagine a "randomized" medical experiment in which we will control only whether patients are placed in one of two divisions, control or new treatment. The treatment and control are designed to lower weight. A computer will flip a coin to assign patients to one of the two divisions. Next, list all the separate factors or variables that can also effect the outcome. Be as thorough as possible. Suppose, to simplify matters, that all of these factors can be adjusted so that they can be split into two groups such that there is a (roughly) 50% chance any patient would be in one of the groups. For example, sex (male or female), physical activity (a cutoff chosen so that there are two groups high or low), thyroid activity (normal or abnormal), and so on. Now, all of these factors will not be independent from one another, but for the purposes of this exercise, suppose that they are. We can say that a division is *imbalanced* if it has less than 10% or

more than 90% of patients from any group (e.g. the treatment group has only 5% men). What is the probability that the two divisions are balanced by randomizing?

(5) Find, in the popular press, an announcement of the result of some new medical trial or finding. Scrutinize this report, locate the original paper if you can, probe for weaknesses and explicate exceptions. Try to find everything that could have gone wrong and how the researchers and the reporters might have been too certain in their claims. This task sounds difficult, but experience has shown that it is not.

(6) EXTRA: If your class is one of specialists (for example, physicians), do the previous homework problem but on a paper chosen from a recent important journal.

APPENDIX A

List of R commands

The book website `http://wmbriggs.com/book` contains several files of `R` commands which can be downloaded. Look especially for the file `worked.example.R`, which contains examples of analyses from start to finish. On that site, there is also a `README` file which will detail any updates to the code. I anticipate making changes and improvements to the code over time, but the original function names will not be touched so that anybody who reads this book will be able to perform any of the calculations outlined in the text. There may of course be new functions added.

It is important to remember that these methods are not perfect and that many better exist (see Sivia, 1996). Mine are not bad, but they are very limited. Software for general modern analysis does exist, but it is complicated—far more complicated than any method here. A favorite is JAGS (2008).

This Appendix contains simple summaries of all the `R` commands used in the text, plus some commands you have not yet seen. `R` is a simple and a very complicated language, so do not be frustrated by an inability to master it in one sitting. On-line help is as close as any internet search service. Chances are if you have a question, somebody else has had the same one and posted the answer on some web site.

Basic help

Command	Function
`apropos('abc')`	This will list all the functions/commands that have `abc` in their names. Be sure you use quotes around the text.
`?command`	A shortened version of `help(command)`.
`help(command)`	Obvious. Be sure to look at the end of the help file because there are usually pre-worked `examples` that you can cut and paste. Do not use quotes around the `command`.

Basic data manipulation

Command	Function

`attach(x)`	Attaches, or makes "visible", the variables names in the data `x` available to call directly. The reason for this is that several different data sets can have the some of the same variable names.
`detach(x)`	Detaches, or makes "invisible", the variables names in the data `x`. See `attach()`.
`x = read.csv('file.csv')`	Reads in the CSV file `file`. Do not forget to put the proper path in front of the file names, e.g. `'C:/Documents/mydata/file.csv'`. `R` cannot guess where you have put your file.
`source(url("http://wmbriggs.com/book/Rcode.R"))`	This loads the book's `R` functions into memory so they are ready to use. You can also download `Rcode.R` and access it from your favorite directory. This is safer because you do not have to be online. Try `source("c:/mydirectory/Rcode.R")`.

Basic graphics

Command	Function
`boxplot(y∼x)`	Creates several boxplots of `y` by the different factors of `x`: `x` must be a categorical variable for this to work best. You can always leave off the `∼x` bit.
`density(x)`	Forms an estimate of the density of approximately continuous data. Use with `plot(density(x))`
`hist(x)`	Form a discrete histogram of the data.
`plot(x)`	Plots the object `x`; also try `plot(x,y)`. This function also takes an enormous number of arguments that you can use to make the picture prettier.
`stem(x)`	Plots a stem-and-leaf.

Summaries & Classical estimation

Command	Function
`summary(x)`	Compute min, max, median, and mean, plus counts the missing values of `x`. Also prints out the summary of regression models and so on. Always try this commands on any object.
`table(x)`	Computes a count of each level of `x`.

`prop.table(table(x))`	Computes a percent, or `prop`ortion, of each level of `x`.
`mean(x,na.rm=T)`	Computes the mean and removes missing values, the `na.rm=T` is optional.
`sd(x,na.rm=T)`	Computes the mean and removes missing values.
`binom.test(k,n)`	Estimate the binomial parameter θ (with standard confidence interval).

Modern estimation

Command	Function
`newdbinom(x, n_new, k_old, n_old)`	Probability distribution for new observables for binomial distribution given old data and guess of how many new data points there will be `n_new`.
`newpnorm(x, x_old)`	Probability distribution for new observables for normal distribution given old data.

Classical testing

Command	Function
`t.test(y~x)`	This is the t-test written in regression form. You can also write it the next way.
`t.test(y, x), paired = T`	This is the same thing but tells `R` the data are paired like in the Army training example. Leave off `paired=T` is the data are not pairs.
`prop.test(c(k1 , k2), c(n1, n2))`	This proportions test where you plug in the number of successes in the first group `k1` then the second `k2`, followed by the sample size in each group.

Classical linear models

Command	Function
`fit = glm(y~x1+x2)`	Standard linear regression for the formula $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$. To get the interaction term $\beta_3 x_{1,i} x_{2,i}$ change the syntax to `x1*x2`. The results are stored in `fit`.
`fit = glm(y~x1+x2, family=binomial)`	Same thing except gives logistic regression.
`summary(fit)`	Shows the results of the linear models.
`confint(fit)`	Gives the classical 95% confidence intervals of the linear models.

Modern linear models

Command	Function
`glm.posterior(fit)`	A textual and graphical summary of the posterior distribution of the linear models *parameters*. `fit` is from the output of `glm` above.
`s1 = obs.glm(fit, newdata)`	Creates the posterior distribution of the future *observables* given a scenario `newdata`. Whatever variables went into the linear model *must* be in `newdata`; e.g. `newdata=data.frame(White.Blood.Count=9.5, Age=22, n=1000)`. The scenario is stored in `s1`. Two scenarios may have the same `newdata` but have different `fit`s resulting in two different runs of `glm`; e.g. `fit1 = glm(y ~ x1)` and `fit2 = glm(y ~ 1)`, which is the usual "null" model.
`obs.glm.prob(s1, s2)`	A textual and graphical summary of the posterior distribution of the linear's models *future observables*. Gives the probability that a new observation from scenario 1 (`s1`) is less than a new observation from scenario 2 (`s2`). Also shows the most probable value under each scenario and the probability that `s1` and `s2` is greater than the most probable value for `s1`.
`skill(fit1 ,fit2)`	Gives the score and skill score comparing how well the models `fit1` and `fit2` fit the observed data, fully accounting for the uncertainty in the observables.

Bibliography

Adams, E. W. (1998). *A Primer of Probability Logic*. CSLI Publications, Leland Stanford Junior University.

Benenson, F. C. (1984). *Probability, Objectivity, and Evidence*. Routledge & Kegan Paul, London.

Berger, J. O. and Selke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence. *JASA.*, 33:112–122.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Wiley, New York.

Birkhahn, R., Briggs, W. M., Datillo, P., Deusen, S. V., and Gaeta, T. (2006). Classifying patients suspected of appendicitis with regard to likelihood. *American Journal of Surgery*, 191(4):497–502.

Briggs, W. M. (2006). Broccoli reduces the risk of splenetic fever! the use of induction and falsifiability in statistics and model selection. *arxiv.org/pdf/math.GM/0610859*.

Briggs, W. M. (2007). Positive evidence for non-arbitrary assignment of probability. In Knuth, K. H., Caticha, A., Center, J. L., Giffin, A., and Rodriguez, C. C., editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 101–108. American Institute of Physics, Melville, New York.

Briggs, W. M. and Ruppert, D. (2005). Assessing the skill of yes/no predictions. *Biometrics*, 61(3):799–807.

Briggs, W. M. and Zaretzki, R. A. (2008). The skill plot: a graphical technique for evaluating continuous diagnostic tests. *Biometrics*, 64:250–263. (with discussion).

Brown, F. M. (2003). *Boolean Reasoning*. Dover, Mineola, NY, second edition.

Campbell, S. and Franklin, J. (2004). Randomness and induction. *Synthese*, 138:79–99.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago University Press, Chicago.

Chaitin, G. (2005). *Meta Math! The Quest for Omega*. Vintage, New York.

Cook, D. B. (2002). *Probability and Schrädinger's Mechanics*. World Scientific, Singapore.

Cox, R. T. (1961). *Algebra of Probable Inference*. Johns Hopkins University Press, Baltimore.

de Laplace, M. (1996). *A Philosophical Essay on Probabilities*. Dover, Mineola, NY.

Dupre, M. J. and Tipler, F. J. (2007). The Cox theorm: unknowns and plausible value. *arxiv.org/pdf/math.PR/0611795v1*.

Fine, T. L. (1973). *Theories of Probability: An Examination of Foundations*. Academic Press, New York.

Fisher, R. (1970). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, fourteenth edition.

Fisher, R. (1973). *Statistical Methods and Scientific Inference*. Hafner Press, New York, third edition.

Franklin, J. (2001a). Resurrecting logical probability. *Erkenntnis*, 55:277–305.

Franklin, J. (2001b). *Theory of Probability:evidence and probability before Pascal*. Johns Hopkins, Baltimore.

Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall, New York.

Gill, J. (2004). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Chapman & Hall, Boca Raton, FL.

Gilovich, T., Griffin, D., and Kahneman, D., editors (2002). *Heuristics and Biases*. Cambridge Univeristy Press, Cambridge.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *JASA*, 102:359–378.

Hájek, A. (1997). Mises redux—redux: Fifteen arguments against finite frequentism. *Erkenntnis*, 45:209–227.

Halpern, J. Y. (1999a). A counterexample to theorems of Cox and Fine. *J. of Artificial Intelligence Research*, 10:67–85.

Halpern, J. Y. (1999b). Cox's theorem revisited. *J. of Artificial Intelligence Research*, 11:429–435.

Howson, C. and Urbach, P. (1993). *Scientific Reasoning: the Bayesian Approach*. Open Court, Chicago, second edition.

Hume, D. (2003). *A Treatise of Human Nature*. Oxford Univeristy Press, Oxford, corrected edition.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Jeffrey, R. (2004). *Subjective Probability*. Cambridge University Press, Cambridge.

Jeffreys, H. (1998). *Theory of Probability*. Oxford University Press, Oxford.

Kadane, J. B. (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*. Wiley, New York.

Keynes, J. M. (2004). *A Treatise on Probability*. Dover Phoenix Editions, Mineola, NY.

Kline, M. (1980). *Mathematics: the Loss of Certainty*. Oxford, Oxford.

Kyburg, H. E. and Smokler, H. E. (1964). *Studies in Subjective Probability*. Krieger Publishing Co., New York.

Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*. Blackwell Publishing.

Lee, J. C., Johnson, W. O., and Zellner, A., editors (1996). *Modelling and Prediction: Honoring Seymour Geisser*. Springer, New York.

Leibniz, G. (2005). *Discourse on Metaphysics and The Monadology*. Dover, Mineola, NY.

Little, R. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *American Statistician*, 60(3):1–11.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A*, 236:333–380.

Plato, J. V. (1998). *Creating Modern Probability*. Cambridge University Press, Cambridge.

Plummer, M. (2008). *JAGS: Just Another Gibbs Sampler*. International Agency for Research on Cancer, Lyon, France. http://www-fis.iarc.fr/∼martyn/software/jags/.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ross, S. (1988). *A First Course in Probability*. Macmillan Publishing Company, New York, third edition.

Sivia, D. (1996). *Data Analysis: A Bayesian Tutorial*. Clarendon Press, Oxford.

Steele, J. M. (2005). Darrell huff and fifty years of How to Lie with Statistics. *Statistical Science*, 20:205–209.

Stove, D. (1973). *Probability and Hume's Inductive Scepticism*. Clarendon, Oxford.

Stove, D. (1982). *Popper and After: Four Modern Irrationalists*. Pergamon Press, Oxford.

Stove, D. (1986). *The Rationality of Induction*. Clarendon, Oxford.

Sullivan, M. (2007). *Statistics: Informed Decisions Using Data*. Pearson, Upper Saddle River, New Jersey, second edition.

Tipler, F. J. (2008). What about quantum theory? Bayes and the Born interpretaion. *arxiv.org/pdf/quant-ph/0611245v1*.

Wasserman, L. (2000). Bayesian model selection and model averaging. *J. Mathematical Psychology*, 44:92–107.

West, M. (1986). Bayesian model monitoring. *J. R. Statist. Soc. B*, 48(1):70–78.

Williams, D. (1947). *The Ground of Induction*. Russell & Russell, New York.

Index

About Your Author

Your author has a Ph.D. in mathematical statistics from Cornell University, and is now a statistician at New York Methodist Hospital and consultant in New York City. He has long been interested in the math, philosophy, probability, and statistics of how well people make predictions and understand uncertainty. He works mainly in medicine, meteorology, climatology, and in any area in which people are willing to pay him. He has concluded that too many people are too certain about too many things. He dislikes writing of himself in the third person. The web site where more about this particularly fascinating individual may be found is WMBRIGGS.COM.