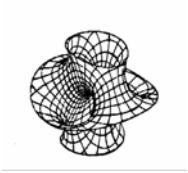
UNIVERSIDADE FEDERAL FLUMINENSE
CENTRO DE ESTUDOS GERAIS
INSTITUTO DE MATEMÁTICA



DEPARTAMENTO DE ESTATÍSTICA

ESTATÍSTICA DESCRITIVA

Ana Maria Lima de Farias
Luiz da Costa Laurencel

Setembro 2006

Conteúdo

Capítulo 1

Apresentação de Dados

1.1 Pesquisa estatística - conceitos básicos

1.1.1 População e amostra

Estatística é a ciência da aprendizagem a partir dos dados. Em geral, fazemos levantamentos de dados para estudar e compreender características de uma população. Por exemplo, um grande banco, querendo lançar um novo produto, precisa conhecer o perfil sócio-econômico dos seus clientes e, nesse caso, a população de interesse é formada pelos clientes de todas as agências do banco. A Federação das Indústrias do Estado do Rio de Janeiro - FIRJAN - mede o grau de confiança dos empresários industriais através de uma pesquisa junto às empresas industriais, sendo a população de interesse, aqui, o conjunto das empresas industriais do estado do Rio de Janeiro. Com esses dois exemplos apenas, já podemos ver que o conceito de *população de uma pesquisa estatística* é mais amplo, não se restringindo a seres humanos; ela é definida exatamente a partir dos objetivos da pesquisa. Mais precisamente, *população é o conjunto de elementos para os quais se deseja estudar determinada(s) característica(s).*

Embora tenham populações bastante distintas, essas duas pesquisas têm em comum o fato de os resultados desejados serem obtidos a partir de dados levantados junto a um subconjunto da população - uma *amostra*. Há várias razões para se trabalhar com *pesquisas por amostragem* - custo e tempo, em geral, são as mais comuns. Mas, além de serem mais baratas e rápidas, as pesquisas por amostragem, se bem planejadas, podem fornecer resultados quase tão precisos quanto aqueles fornecidos por *pesquisas censitárias,* onde todos os elementos da população são investigados. Exemplos clássicos de pesquisa censitária são os Censos Demográficos realizados a cada dez anos no Brasil e em outros países. O objetivo desses censos é levantar informações sobre toda a população do país, de modo a fornecer subsídios para os governantes definirem as políticas públicas.

1.1.2 Variáveis qualitativas e quantitativas

Nas pesquisas estatísticas, as características sobre as quais queremos obter informação são chamadas *variáveis*. Em uma pesquisa domiciliar sobre emprego e renda, algumas variáveis de interesse são sexo, raça, grau de instrução e valor dos rendimentos do morador. Em uma pesquisa sobre o estado nutricional dos brasileiros, o peso e a altura dos moradores de cada domicílio da amostra foram medidos. Para o acompanhamento da atividade industrial no Rio de Janeiro, a FIRJAN obtém informações junto às empresas industriais sobre tipo de atividade econômica, número de empregados, número de horas trabalhadas, valor da folha de pagamento. É importante diferenciar entre variáveis qualitativas e variáveis quantitativas. Sexo, raça, religião e atividade econômica de

uma empresa são exemplos de *variáveis qualitativas*. Já valor dos rendimentos, peso, altura, número de empregados, valor da folha de pagamento são exemplos de *variáveis quantitativas*. Podemos ver, então, que as variáveis qualitativas *descrevem* características dos elementos de uma população, enquanto as variáveis quantitativas *mensuram* características desses elementos.

As variáveis quantitativas, por sua vez, podem ser classificadas em *discretas* ou *contínuas*. As variáveis discretas, em geral, envolvem contagens, enquanto as contínuas envolvem mensurações. Assim, número de filhos, número de empregados, número de caras em 10 lançamentos de uma moeda, etc, são exemplos de variáveis discretas. Já peso, altura, renda, pressão sanguínea, etc, são exemplos de variáveis contínuas. A definição formal é a seguinte:

Definição 1 *Uma variável quantitativa é discreta se os seus valores possíveis formam um conjunto finito ou enumerável. As variáveis contínuas são aquelas cujos valores possíveis encontram-se em aglum intervalo da reta.*

Exercício 1.1 *O texto a seguir foi extraído da página do IBOPE na Internet: www.ibope.com.br. Aí temos parte da descrição da pesquisa sociodemográfica realizada por esse instituto. Identifique as variáveis pesquisadas, classificando-as como qualitativas ou quantitativas.*

"O Levantamento Socioeconômico (LSE) é a pesquisa do IBOPE Mídia que mapeia as características sociais, demográficas e econômicas das famílias das principais regiões metropolitanas do país. Oferece também outros dados essenciais para traçar a estratégia de marketing para um produto. Com uma base de dados estendida em relação às outras pesquisas do IBOPE Mídia, o LSE serve de base para outros estudos.

São levantados dados sobre a condição do domicílio entrevistado (condição da rua, tipo de imóvel) e sobre a condição socioeconômica do domicílio (informações sobre renda e classificação econômica). Também são pesquisados o número de pessoas no domicílio, a presença e a quantidade de crianças e adolescentes, a idade, grau de instrução e condição de atividade do chefe da casa e da dona-de-casa. A pesquisa levanta também dados sobre a posse de bens, como geladeira, máquina de lavar, automóvel, rádio, computador, telefone, entre outros, e acesso a serviços de mídia, como TV por Assinatura, Internet, etc."

1.2 Apresentação de dados qualitativos

Vamos considerar o seguinte exemplo fictício, mas verossímil. A direção de uma empresa está estudando a possibilidade de fazer um seguro saúde para seus funcionários e respectivos familiares. Para isso, ela faz um levantamento junto a seus 500 funcionários, obtendo informação sobre sexo, estado civil, idade, número de dependentes e salário. Como são 500 funcionários, temos que achar uma forma de resumir os dados. Nesta aula você irá aprender a resumir dados qualitativos em forma de uma distribuição (ou tabela) de freqüência e também em forma gráfica. Você verá que os gráficos complementam a apresentação tabular.

1.2.1 Distribuições de freqüência

Consideremos inicialmente a variável qualitativa sexo. O que interessa saber sobre essa variável não é que João é do sexo masculino e Maria é do sexo feminino, mas, sim, quantos funcionários e quantas funcionárias há na empresa. Esse resultado pode ser resumido em uma tabela ou distribuição de freqüências da seguinte forma:

Sexo	Número de Funcionários
Masculino	270
Feminino	230
Total	500

Os números 270 e 230 resultaram da contagem das freqüências de ocorrência de cada uma das categorias da variável sexo. Essa contagem é também chamada de *freqüência simples absoluta* ou simplesmente *freqüência*. O total de 500 é obtido somando-se o número de homens e de mulheres.

É interessante também expressar esses resultados em forma relativa, ou seja, considerar a *freqüência relativa* de cada categoria em relação ao total:

$$\frac{270}{500} = 0,54$$

ou seja, 54% dos funcionários da empresa são do sexo masculino e

$$\frac{230}{500} = 0,46$$

ou seja, 46% dos funcionários são mulheres. A Tabela 1.1 apresenta a versão completa.

Tabela 1.1: Distribuição do número de funcionários por sexo

Sexo	Freqüência Simples	
	Absoluta	Relativa
Masculino	270	0,54
Feminino	230	0,46
Total	500	1,00

Note que a soma das freqüências relativas é sempre 1, enquanto a soma das freqüências absolutas deve ser igual ao número total de elementos sendo investigados.

De maneira análoga, obteríamos a Tabela 1.2 para a variável estado civil. Note que, aí, a freqüência relativa está apresentada em forma percentual, ou seja, multiplicada por 100. Por exemplo, para os casados temos:

$$\frac{280}{500} \times 100 = 0,56 \times 100 = 56\%$$

Em geral, essa é a forma mais usual de se apresentarem as freqüêcias relativas e neste caso, a soma deve dar 100%.

Exemplo 1

Consideremos que, na situação descrita anteriormente, os dados tenham sido levantados por departamento, para depois serem totalizados. Para o Departamento de Recursos Humanos, foram obtidas as seguintes informações, apresentadas na Tabela 1.3:

Tabela 1.2: Distribuição do número de funcionários por estado civil

Estado Civil	Freqüência Simples	
	Absoluta	Relativa %
Solteiro	125	25,0
Casado	280	56,0
Divorciado	85	17,0
Viúvo	10	2,0
Total	500	100,0

Tabela 1.3: Funcionários do Departamento de RH

Nome	Sexo	Estado civil	Número de dependentes
João da Silva	M	Casado	3
Pedro Fernandes	M	Viúvo	1
Maria Freitas	F	Casada	0
Paula Gonçalves	F	Solteira	0
Ana Freitas	F	Solteira	1
Luiz Costa	M	Casado	3
André Souza	M	Casado	4
Patrícia Silva	F	Divorciada	2
Regina Lima	F	Casada	2
Alfredo Souza	M	Casado	3
Margarete Cunha	F	Solteira	0
Pedro Barbosa	M	Divorciado	2
Ricardo Alves	M	Solteiro	0
Márcio Rezende	M	Solteiro	1
Ana Carolina Chaves	F	Solteira	0

Para pequenos conjuntos de dados, podemos construir a tabela à mão e para isso precisamos contar o número de ocorrências de cada categoria de cada uma das variáveis. Varrendo o conjunto de dados a partir da primeira linha, podemos ir marcando as ocorrências da seguinte forma:

Masculino	///// ///	Solteiro	///// /
Feminino	///// //	Casado	///// /
		Divorciado	//
		Viúvo	/

Obtemos, então, as seguintes tabelas:

Sexo	Freqüência Simples	
	Absoluta	Relativa %
Masculino	8	53,33
Feminino	7	46,67
Total	15	100,0

Estado Civil	Freqüência Simples	
	Absoluta	Relativa %
Solteiro	6	40,00
Casado	6	40,00
Divorciado	2	13,33
Viúvo	1	6,67
Total	15	100,00

No exemplo anterior, a divisão de algumas freqüencias absolutas pelo total de 15 resultou em dízimas. Nesses casos, torna-se necessário arredondar os resultados, mas esse arredondamento deve ser feito com cautela para se evitarem problemas tais como a soma não ser igual a 1 ou 100%.

A primeira etapa no processo de arredondamento consiste em se decidir o número de casas decimais desejado. Em geral, freqüências relativas percentuais são apresentadas com, no máximo, 2 casas decimais. Isso significa que temos que descartar as demais casa decimais. Existe a seguinte regra de arredondamento:

Regra 1 ***Arredondamento de Números***

Quando o primeiro algarismo a ser suprimido é menor ou igual a 4 (ou seja, é igual a 0, 1, 2, 3 ou 4), o último algarismo a ser mantido permanece inalterado. Quando o primeiro algarismo a ser suprimido é igual a 5, 6, 7, 8 ou 9, o último algarismo a ser mantido é acrescido de 1.

Na distribuição de freqüências da variável sexo, temos os seguintes resultados:

$$\frac{8}{15} \times 100 = 53,33333\ldots$$
$$\frac{7}{15} \times 100 = 46,66666\ldots$$

No primeiro caso, o primeiro algarismo a ser suprimido é 3; logo, o último algarismo a ser mantido (3) não se altera e o resultado é 53,33. No segundo caso, o primeiro algarismo a ser suprimido é 6; logo, o último algarismo a ser mantido (6) deve ser acrescido de 1 e o resultado é 46,67. Tente sempre usar essa regra em seus arredondamentos; com ela, você evitará erros grosseiros.

Exercício 1.2 *Para o Departamento Financeiro, obteve-se a seguinte informação sobre o sexo dos 23 funcionários:*

$$\begin{array}{cccccccccccc} M & F & F & M & M & M & F & F & M & M & M & M \\ M & F & M & M & F & F & M & M & M & F & F & \end{array}$$

onde M = Masculino e F = Feminino. Construa uma tabela de freqüências para esses dados.

1.2.2 Gráficos

As distribuições de freqüência para dados qualitativos também podem ser ilustradas graficamente através de gráficos de colunas ou gráficos de setores, também conhecidos como gráficos de pizza. Na Figura 1.1 temos os gráficos de coluna e de setores para os dados da Tabela 1.2, referentes ao estado civil dos funcionários.

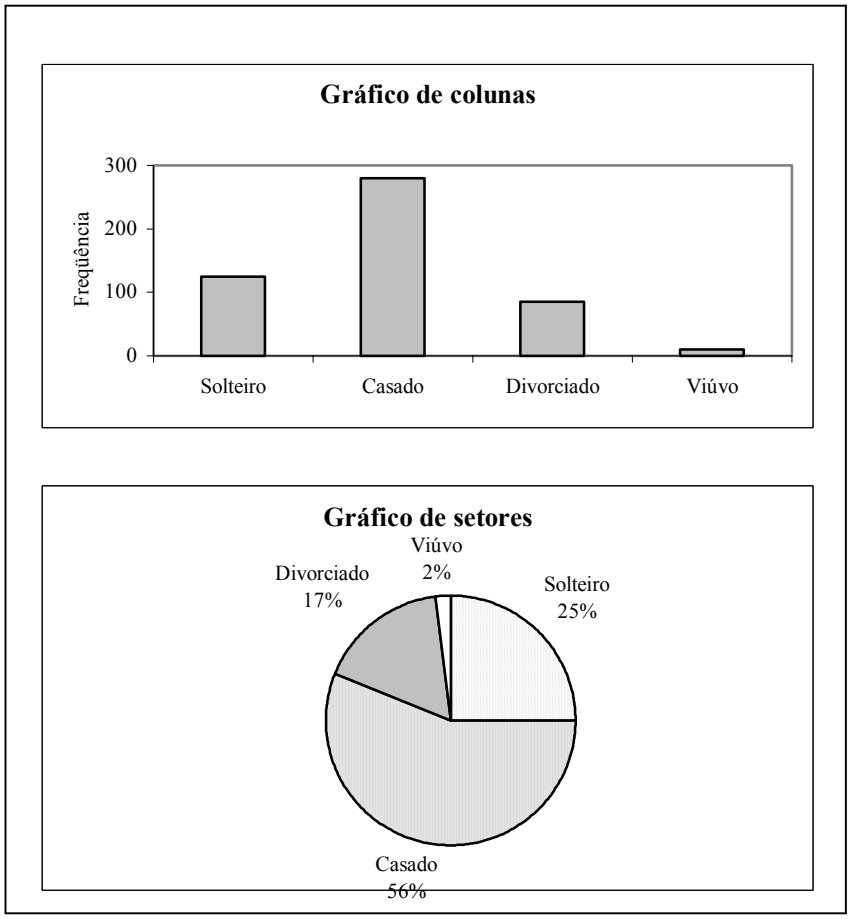


Figura 1.1: Distribuição do número de funcionários por estado civil

No *gráfico de colunas*, a altura de cada coluna representa a freqüência da respectiva classe e o gráfico pode ser construído com base nas freqüências absolutas ou relativas. Para diferenciar um do outro, coloca-se no título do eixo o tipo de freqüência utilizada. Note que, no eixo horizontal, não há escala, uma vez que aí se representam as categorias da variável, que devem ser equi-espaçadas.

No *gráfico de setores*, a freqüência de cada categoria é representada pelo tamanho (ângulo) do setor (ou fatia da pizza). Para construir um gráfico de setores à mão, você precisa de um compasso para fazer um círculo de raio qualquer. Em seguida, trace um raio qualquer no círculo e a partir daí, comece a marcar os raios de acordo com os ângulos de cada setor, utilizando um transferidor. Para determinar o ângulo de cada setor, você deve usar a seguinte regra de proporcionalidade: o ângulo total - 360^o- corresponde ao número total de observações; o ângulo de cada setor corresponde à freqüência da respectiva classe. Dessa forma, você obtém a seguinte regra de três para os Solteiros:

$$\frac{360^o}{500} = \frac{x}{125} \Rightarrow x = 90^o$$

Esses gráficos podem ser construídos facilmente com auxílio de programas de computador, como, por exemplo, o programa de planilhas Excel da Microsoft ®.

Exercício 1.3 *Construa os gráficos de setores e de colunas para os dados do Exercício 1.2.*

1.3 Apresentação de dados quantitativos discretos

Quando uma variável quantitativa discreta assume poucos valores distintos, é possível construir uma distribuição de freqüências da mesma forma que fizemos para as variáveis qualitativas. A diferença é que, em vez de termos categorias nas linhas da tabela, teremos os distintos valores da variável. Continuando com o nosso exemplo, vamos trabalhar agora com a variável número de dependentes. Suponha que alguns funcionários não tenham dependentes e que o número máximo de dependentes seja 7. Obteríamos, então, a seguinte distribuição de freqüências:

Número de	Freqüência Simples	
dependentes	Absoluta	Relativa %
0	120	24,0
1	95	19,0
2	90	18,0
3	95	19,0
4	35	7,0
5	30	6,0
6	20	4,0
7	15	3,0
Total	500	100,0

O processo de construção é absolutamente o mesmo mas, dada a natureza quantitativa da variável, é possível acrescentar mais uma informação à tabela. Suponha, por exemplo, que a empresa esteja pensando em limitar o seu projeto a 4 dependentes, de modo que funcionários com mais de 4 dependentes terão que arcar com as despesas extras. Quantos funcionários estão nessa situação? Para responder a perguntas desse tipo, é costume acrescentar à tabela de freqüências uma coluna com as *freqüências acumuladas*. Essas freqüências são calculadas da seguinte forma: para cada valor da variável (número de dependentes), contamos quantas ocorrências correspondem a valores menores ou iguais a esse valor. Por exemplo, valores da variável menores ou iguais a 0 correspondem aos funcionários sem dependentes. Logo, a freqüência acumulada para o valor 0 é igual à freqüência

simples: 120. Analogamente, valores da variável menores ou iguais a 1 correspondem aos funcionários sem dependentes mais os funcionários com 1 dependente. Logo, a freqüência acumulada para o valor 1 é igual a 120+95 = 215. Para o valor 2, a freqüência acumulada é igual a 120+95+90 = 215+90 = 305. Repetindo esse procedimento, obtemos a Tabela 1.4. Note que aí acrescentamos também as

Tabela 1.4: Distribuição de Freqüências para o Número de Dependentes

Número de	Freqüência Simples		Freqüência Acumulada	
dependentes	Absoluta	Relativa %	Absoluta	Relativa %
0	120	24,0	120	24,0
1	95	19,0	215	43,0
2	90	18,0	305	61,0
3	95	19,0	400	80,0
4	35	7,0	435	87,0
5	30	6,0	465	93,0
6	20	4,0	485	97,0
7	15	3,0	500	100,0
Total	500	100,0		

freqüências acumuladas em forma percentual. Essas freqüências são calculadas como a proporção da freqüência acumulada em relação ao total; por exemplo,

$$87,0 = \frac{435}{500} \times 100$$

Exercício 1.4 *Construa a distribuição de freqüência para o número de dependentes dos funcionários do Departamento de Recursos Humanos, conforme dados a seguir (ver no Exemplo 1):*

3 1 0 0 1 3 4 2 2 3 0 2 0 1 0

A representação gráfica da distribuição de freqüências de uma variável quantitativa discreta pode ser feita através de um gráfico de colunas. A única diferença nesse caso é que no eixo horizontal do gráfico é representada a escala da variável quantitativa e tal escala deve ser definida cuidadosamente de modo a representar corretamente os valores. Na Figura 1.2 temos o gráfico de colunas para o número de dependentes dos 500 funcionários.

Embora não seja incorreto, não é apropriado representar dados quantitativos discretos em um gráfico de setores, uma vez que, nesse gráfico, não é possível representar a escala dos dados.

Exercício 1.5 *Construa o gráfico de colunas para representar a distribuição de freqüências obtida no Exercício 1.4.*

No Exemplo 1, há duas variáveis quantitativas discretas: número de dependentes e idade. A diferença entre elas é que a idade pode assumir um número maior de valores, o que resultaria em uma tabela grande, caso decidíssemos relacionar todos os valores. Além disso, em geral não é

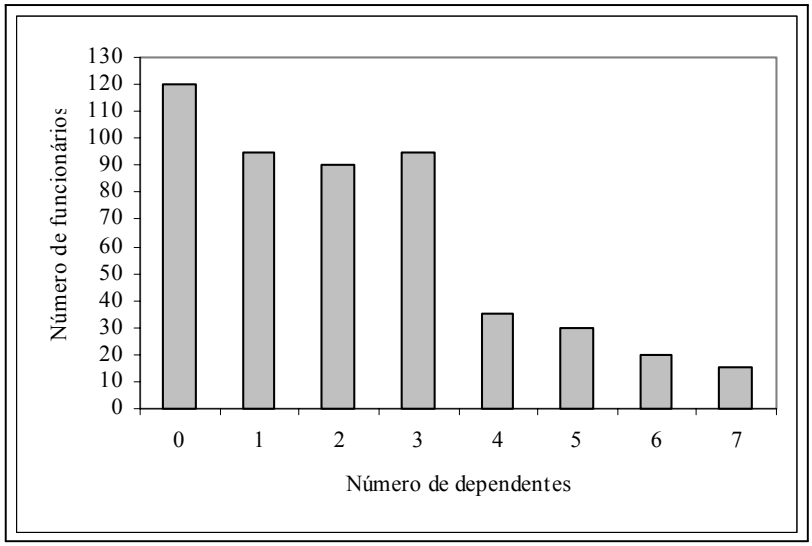


Figura 1.2: Distribuição do número de dependentes de 500 funcionários

necessário apresentar a informação em tal nível de detalhamento. Por exemplo, para as seguradoras de planos de saúde, as faixas etárias importantes - aquelas em que há reajuste por idade - são 0 a 18; 19 a 23; 24 a 28; 29 a 33; 34 a 38; 39 a 43; 44 a 48; 49 a 53; 54 a 58 e 59 ou mais. Sendo assim, podemos agrupar os funcionários segundo essas faixas etárias e construir uma *tabela de freqüências agrupadas* da mesma forma que fizemos para o número de dependentes, só que agora cada freqüência corresponde ao número de funcionários na respectiva faixa etária. Na Tabela 1.5 temos a tabela resultante.

Tabela 1.5: Distribuição de freqüência da idade de 500 funcionários

Faixa	Freqüência Simples		Freqüência Acumulada	
Etária	Absoluta	Relativa %	Absoluta	Relativa %
$19 - 23$	1	0,2	1	0,2
$24 - 28$	23	4,6	24	4,8
$29 - 33$	103	20,6	127	25,4
$34 - 38$	246	49,2	373	74,6
$39 - 43$	52	10,4	425	85,0
$44 - 48$	50	10,0	475	95,0
$49 - 53$	25	5,0	500	100,0
Total	500	100,0		

Exercício 1.6 *Na Tabela 1.6 temos as informações sobre idade e salário para os 15 funcionários do Departamento de Recursos Humanos. Construa uma tabela de freqüências para a idade, levando em conta as mesmas faixas etárias utilizadas acima.*

Tabela 1.6: Idade e salário dos funcionários do Departamento de RH

Nome	Idade	Salário
João da Silva	36	6300
Pedro Fernandes	51	5700
Maria Freitas	26	4500
Paula Gonçalves	25	3800
Ana Freitas	29	3200
Luiz Costa	53	7300
André Souza	42	7100
Patrícia Silva	38	5600
Regina Lima	35	6400
Alfredo Souza	45	7000
Margarete Cunha	26	3700
Pedro Barbosa	37	6500
Ricardo Alves	24	4000
Márcio Rezende	31	5100
Ana Carolina Chaves	29	4500

1.4 Apresentação de dados quantitativos contínuos

Para as variáveis quantitativas contínuas, devemos também trabalhar com distribuições de freqüências agrupadas. O processo de construção é idêntico ao visto para as variáveis discretas, mas aqui devemos tomar um cuidado especial na construção das classes. A escolha dos limites das classes deve ser feita com base na natureza, valores e unidade de medida dos dados. As únicas regras que têm que ser seguidas são as seguintes.

Regra 2 ***Definição das classes em uma distribuição de freqüências agrupadas***

1. *As classes têm que ser exaustivas, isto é, todos os elementos devem pertencer a alguma classe.*
2. *As classes têm que ser mutuamente exclusivas, isto é, cada elemento tem que pertencer a uma única classe.*

O primeiro passo é definir o número de classes desejado; esse número, de preferência, deve estar entre 5 e 25. Em seguida, devemos determinar a *amplitude* dos dados, ou seja, o intervalo de variação dos valores observados da variável em estudo.

Definição 2 *A* ***amplitude*** *de um conjunto de dados, representada por Δ_{total}, é definida como a diferença entre os valores máximo e mínimo:*

$$\Delta_{total} = V_{Máx} - V_{Mín} \tag{1.1}$$

Como regra geral, considere o primeiro múltiplo do número de classes maior que o valor da amplitude. Dividindo esse valor pelo número de classes, você obtém o comprimento de cada classe e

os limites de classe podem ser obtidos somando-se o comprimento de classe a partir do valor mínimo dos dados.

Exemplo 2

Suponha que entre os 500 funcionários da nossa empresa, o menor salário seja 2800 e o maior salário seja de 12400. Para agrupar os dados em 5 classes devemos fazer o seguinte:

$$\Delta_{total} = V_{\text{Máx}} - V_{\text{Mín}} = \quad 12400 - 2800 = 9600$$

$$\text{Próximo múltiplo de } 5 = \quad 9605$$

$$\text{Comprimento de classe} = \quad \frac{9605}{5} = 1921$$

Os limites de classe, então, são:

$$\begin{aligned} & & 2800 \\ 2800 + 1921 &= 4721 \\ 4721 + 1921 &= 6642 \\ 6642 + 1921 &= 8563 \\ 8563 + 1921 &= 10484 \\ 10484 - 1921 &= 12405 \end{aligned}$$

e as classes podem ser definidas como

$$\begin{aligned} &[2800, 4721) \\ &[4721, 6642) \\ &[6642, 8563) \\ &[8563, 10484) \\ &[10484, 12405) \end{aligned}$$

Essa é uma regra que resulta em classes corretamente definidas, mas nem sempre as classes resultantes são apropriadas ou convenientes. No exemplo acima, poderia ser preferível trabalhar com classes de comprimento 2000, definindo o limite inferior dos dados como 2500. Isso resultaria nas classes [2500,4500); [4500,6500); [6500,8500); [8500, 10500) e [10500, 12500), que são classes corretas e mais fáceis de ler.

Exercício 1.7 *Construa uma distribuição de freqüências agrupadas em 5 classes de mesmo comprimento para os dados de salários da Tabela 1.6.*

Tabela 1.7: Distribuição dos salários dos funcionários do Departamento de RH

Classe	Freqüência Simples		Freqüência Acumulada	
de salário	Absoluta	Relativa %	Absoluta	Relativa %
[3200,4021)	4	26,67	4	26,67
[4021,4842)	2	1,33	6	40,00
[4842,5663)	2	1,33	8	53,33
[5663,6484)	3	20,00	11	73,33
[6484,7305)	4	26,67	15	100,00
Total	15	100,00		

1.4.1 Histogramas e polígonos de freqüência

O histograma e o polígono de freqüências são gráficos usados para representar uma distribuição de freqüências simples de uma variável quantitativa contínua.

Um *histograma* é um conjunto de retângulos com bases sobre um eixo horizontal dividido de acordo com os comprimentos de classes, centros nos pontos médios das classes e *áreas proporcionais ou iguais às freqüências*. Vamos ilustrar a construção de um histograma usando como exemplo a distribuição de freqüência dos dados sobre salários do Exercício 1.7, reproduzida na Tabela 1.7.

Como as classes têm o mesmo comprimento, o histograma, nesse caso, pode ser construído de tal modo que as alturas dos retângulos sejam iguais às freqüências das classes. Dessa forma, as áreas serão *proporcionais* (e não iguais) às freqüências, conforme ilustrado no gráfico superior da Figura 1.3. No gráfico inferior nessa mesma figura, a área de cada retêngulo é *igual* à freqüência relativa da classe e a altura de cada classe é calculada usando-se a expressão que dá a área de um retângulo. Por exemplo, para a classe [3200,4021), a freqüência (área) é $\frac{4}{15} = 0,266667$ e a base do retângulo (comprimento de classe) é 821. Logo, a altura h do retângulo correspondente é

$$h = \frac{0,266667}{821} = 0,000325$$

O resultado dessa divisão é denominado *densidade*, uma vez que dá a concentração em cada classe por unidade da variável. Em ambos os gráficos, a forma dos retângulos é a mesma; o que muda é a escala no eixo vertical.

De modo geral, quando as classes têm o mesmo comprimento - e essa é a situação mais comum - podemos representar as alturas dos retângulos pelas freqüências das classes, o que torna o gráfico mais fácil de interpretar.

Um *polígono de freqüências* é um gráfico de linha que se obtém unindo por uma poligonal os pontos correspondentes às freqüências das diversas classes, centradas nos respectivos pontos médios. Mais precisamente, são plotados os pontos com coordenadas (ponto médio, freqüência simples). Para obter as interseções da poligonal com o eixo, cria-se em cada extremo uma classe com freqüência nula. Na Figura 1.4 temos o polígono de freqüências para a renda dos funcionários do Departamento de Recursos Humanos.

Exercício 1.8 *Na Tabela 1.8 abaixo temos as notas de 50 alunos em uma prova. Construa uma tabela de freqüências agrupadas, usando as classes* $2 \vdash 3, 3 \vdash 4, 4 \vdash 5, \cdots, 9 \vdash 10$. *Construa o histograma e o polígono de freqüências.*

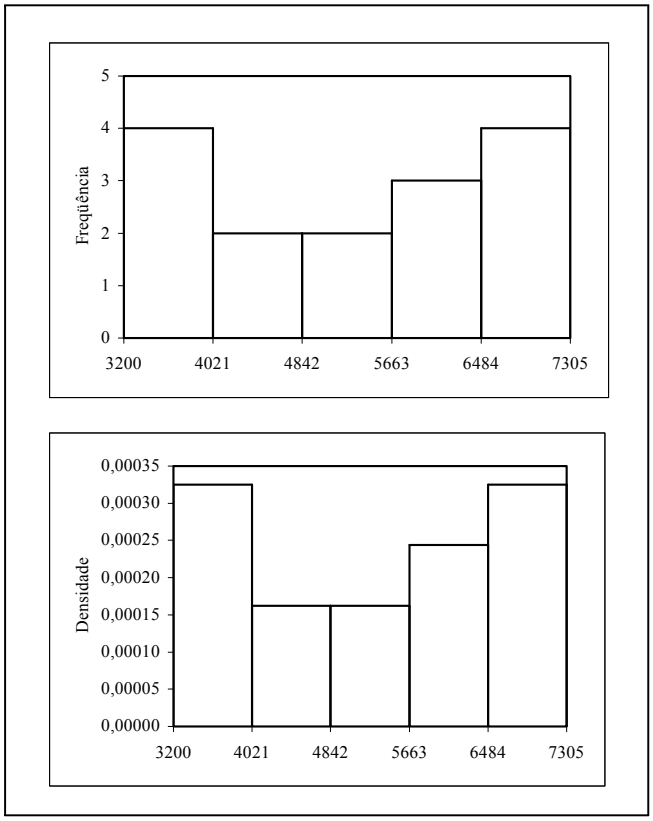


Figura 1.3: Histogramas da distribuição dos salários dos funcionários do Departamento de RH

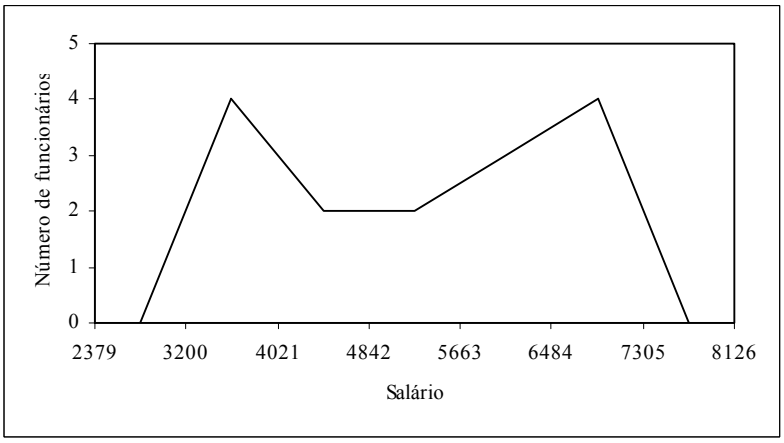


Figura 1.4: Polígono de freqüência para os salários dos funcionários do Departamento de RH

Tabela 1.8: Notas de 50 alunos para o Exercício 1.8

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

1.5 Diagrama de ramos e folhas

Um outro gráfico usado para mostrar a forma da distribuição de um conjunto de dados quantitativos é o diagrama de ramos e folhas, desenvolvido pelo estatístico americano John Tukey. Este gráfico é constituído de uma linha vertical, com a escala indicada à esquerda desta linha. A escala, naturalmente, depende dos valores observados, mas deve ser escolhida de tal forma que cada valor observado possa ser "quebrado" em duas partes: uma primeira parte quantificada pelo valor da escala e a segunda quantificada pelo último algarismo do número correspondente à observação. Os *ramos* do gráfico correspondem aos números da escala, à esquerda da linha vertical. Já as *folhas* são os números que aparecem na parte direita. Na Figura 1.5 temos o diagrama de ramos e folhas para as notas de 50 alunos dadas na Tabela 1.8. Nesse caso, a "quebra" dos valores é bastante natural: os ramos são formados pelo algarismo inteiro e as folhas pelos algarismos decimais, o que é indicado pela escala do gráfico.

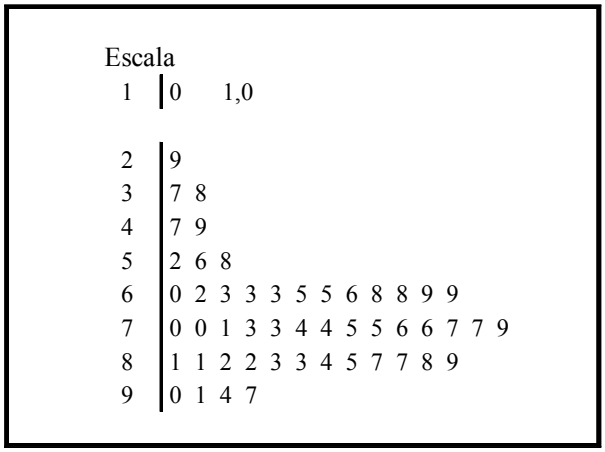


Figura 1.5: Notas de 50 alunos

Note a escala no gráfico: em geral, a escala representa a folha e nesse exemplo poderíamos indicar a escala como "folha = 0,1". Caso os dados fossem do tipo 29, 37, 38, etc, faríamos "folha = 1".

O diagrama de ramos e folhas também é útil na comparação de conjuntos de dados. Suponha que, no exemplo acima, a mesma prova tenha sido aplicada a duas turmas diferentes. Para comparar os resultados, podemos construir o diagrama que se encontra na Figura 1.6. Para facilitar a comparação, é usual indicar o número de dados em cada banda do diagrama. Note que, na parte esquerda do gráfico, as folhas são anotadas crescentemente da direita para a esquerda, enquanto que, na parte direita do gráfico, as folhas são anotadas crescentemente da esquerda para a direita.

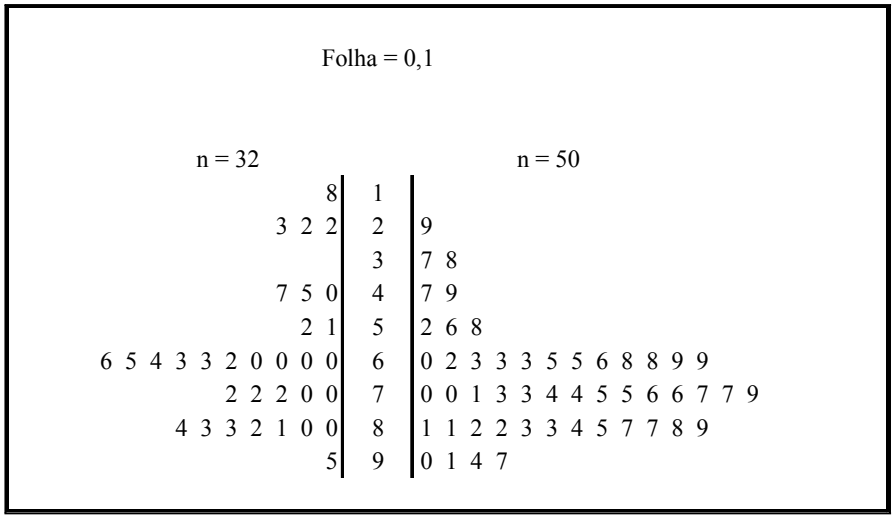


Figura 1.6: Comparação das notas de 2 turmas

Exercício 1.9 *Suponha que as idades dos 23 funcionários do Departamento Financeiro sejam 27; 31; 45; 52; 33; 34; 29; 27; 35; 38; 50; 48; 29; 30; 32; 29; 42; 41; 40; 42; 28; 36; 48. Usando esses dados e aqueles apresentados na Tabela 1.6 sobre os funcionários do Departamento de Recursos Humanos, construa um diagrama de ramos e folhas para comparar os 2 departamentos.*

1.6 Gráficos de linhas

O *gráfico de linhas* é usado principalmente para representar observações feitas ao longo do tempo, isto é, observações de uma *série de tempo*. No eixo horizontal colocam-se as datas em que foram realizadas as observações e no eixo vertical, os valores observados. Os pontos assim obtidos são unidos por segmentos de reta para facilitar a visualização do comportamento dos dados ao longo do tempo. Na Figura 1.7 temos o gráfico que ilustra o consumo de refrigerante (em milhões de litros) no período de 1986 a 2005, conforme dados da ABIR.

Na Tabela 1.9 temos dados sobre o número de homicídios nos estados do Rio de Janeiro e de São Paulo no período de 1980 a 2002. Para efeitos de comparação, é possível construir um gráfico de linhas em que as 2 séries são representadas conjuntamente. Veja a Figura 1.8.

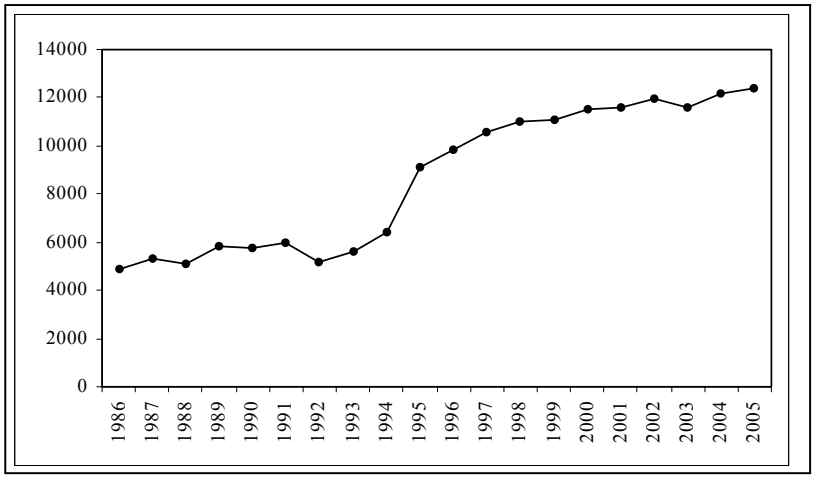


Figura 1.7: Consumo de refrigerante 1986-2005

Tabela 1.9: Número de homicídios - RJ e SP - 1980 a 2002

Ano	RJ	SP	Ano	RJ	SP
1980	2946	3452	1992	4516	9022
1981	2508	4187	1993	5362	9219
1982	2170	4183	1994	6414	9990
1983	1861	5836	1995	8183	11566
1984	2463	7063	1996	8049	12350
1985	2550	7015	1997	7966	12522
1986	2441	7195	1998	7569	14001
1987	3785	7918	1999	7249	15810
1988	3054	7502	2000	7337	15631
1989	4287	9180	2001	7304	15745
1990	7095	9496	2002	8257	14494
1991	5039	9671			

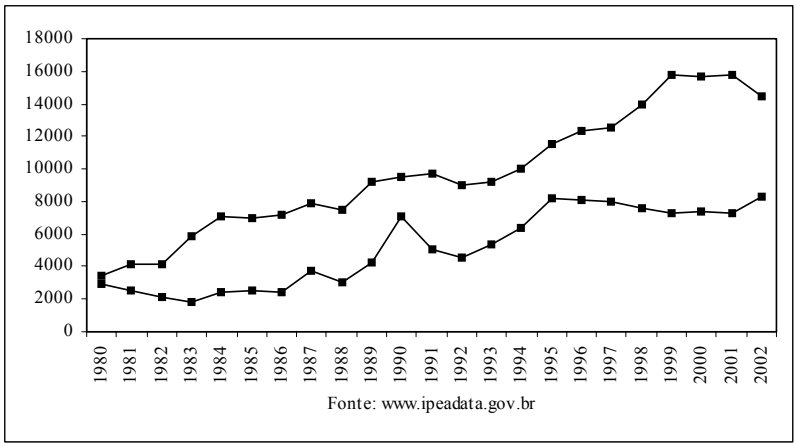


Figura 1.8: Número de homicídios nos estados do Rio de Janeiro e de São Paulo - 1980-2002

1.7 Exercícios Complementares

Exercício 1.10 *Na Tabela 1.10 temos informações sobre o sexo, a matéria predileta (**P**ortuguês, **M**atemática, **H**istória, **G**eografia ou **C**iências) no 2º grau e a nota (número de questões certas) em um teste de múltipla escolha com 10 questões de matemática, ministrado no primeiro dia de aula dos calouros de Economia de uma universidade (dados fictícios).*

1. *Classifique as variáveis envolvidas.*
2. *Construa a tabela de freqüências apropriada para cada uma das variáveis.*
3. *Construa gráficos apropriados para ilustrar as distribuições de freqüência.*

Tabela 1.10: Dados sobre sexo, matéria predileta e nota de alunos

Sexo	Predileta	Nota	Sexo	Predileta	Nota	Sexo	Predileta	Nota	Sexo	Predileta	Nota
F	H	5	M	M	2	M	H	3	F	M	8
M	M	8	M	G	4	M	M	5	M	P	5
F	P	8	M	G	9	F	P	5	M	G	6
F	H	6	M	M	7	F	G	5	F	M	7
M	C	5	M	M	1	M	C	7	M	P	5
M	H	6	F	P	8	M	H	4	F	M	5
F	M	8	F	G	5	F	M	7	F	M	5
F	P	4	M	G	9	F	P	7	F	P	9
F	H	2	M	P	5	F	M	6	M	M	8
M	C	6	F	M	8	M	G	6			
F	P	8	F	G	6	M	H	9			

Exercício 1.11 *Na Tabela 1.11 temos dados sobre o consumo de refrigerantes no Brasil em 2005, segundo dados da Associação Brasileira das Indústrias de Refrigerantes e de Bebidas Não Alcóolicas. Construa um gráfico apropriado para ilustrar esses dados.*

Tabela 1.11: Refrigerantes - Participação dos sabores - 2005

Refrigerantes	%
Colas	51,1
Guaraná	24,4
Laranja	10,9
Limão	5,9
Uva	3,2
Tuti Fruti	1,1
Tônica	0,7
Cítrico	0,1
Maçã	0,5
Outros sabores	2,1
Total	100,0

Fonte: ABIR - www.abir.org.br

Exercício 1.12 *Na Tabela 1.12 temos as freqüências acumuladas do número de sinistros por apólice de seguro do ramo Automóveis. Complete a tabela, calculando as freqüencias simples absolutas e relativas e também as freqüências acumuladas relativas.*

Tabela 1.12: Número de sinistros por apólice, para o Exercício 1.12

Número de sinistros	Número de apólices
0	2913
≤ 1	4500
≤ 2	4826
≤ 3	4928
≤ 4	5000

Exercício 1.13 *Para a seguinte notícia, extraída do jornal Folha de São Paulo, construa um gráfico para ilustrar o texto no segundo parágrafo da notícia.*

"Dentro de dez anos, 90% do mercado automobilístico mundial estará nas mãos de meia dúzia de conglomerados. A previsão consta de estudo produzido pela consultoria especializada britânica Autopolis, que dá assessoria técnica a montadoras que estão instaladas no Reino Unido.

... Dados levantados pela Autopolis mostram que, hoje, a concentração de mercado já é grande. Cerca de 75% do setor é dominado por somente seis conglomerados, liderados por General Motors (22,8%), Ford (16,8%), Volkswagen (9,4%), Toyota (9,2%, incluindo Daihatsu), Reanult-Nissan (8,7%) e Daimler-Chrysler (8,3%). Os outros 24,8% do mercado são dominados por uma infinidade de empresas pequenas e médias, como Fiat, BMW, Peugeot e Honda, entre outras."

Exercício 1.14 *Num estudo sobre a jornada de trabalho das empresas de Produtos Alimentares foram levantados os dados da Tabela 1.13 relativos ao total de horas trabalhadas pelos funcionários no mês de agosto (dados hipotéticos). Construa uma tabela de freqüências usando 5 classes de mesmo tamanho; construa também o histograma e o polígono de freqüências. Para facilitar a solução, os valores mínimo e máximo são: 1.815 e 118.800.*

Tabela 1.13: Jornada de trabalho de empresas alimentares para o Exercício 1.14

3.960	5.016	13.015	8.008	6.930	5.544	4.224	6.138
118.800	57.904	72.600	100.100	55.935	7.223	3.775	4.224
3.216	7.392	2.530	6.930	1.815	4.338	8.065	10.910
8.408	8.624	6.864	5.742	5.749	8.514	2.631	5.236
8.527	3.010	5.914	11.748	8.501	6.512	11.458	10.094
6.721	2.631	7.082	10.318	8.008	3.590	7.128	7.929
10.450	6.780	5.060	5.544	6.178	13.763	9.623	14.883
17.864	34.848	25.300	52.800	17.732	63.923	30.360	18.876
30.800	19.562	49.240	49.434	26.950	22.308	21.146	14.212
25.520	49.251	30.976	23.338	43.648	26.796	44.880	30.008
30.769	16.907	33.911	27.034	16.500	14.445	28.160	42.442
16.507	36.960	67.760	84.084	89.888	65.340	82.280	86.152
91.080	99.792	77.836	76.032				

Exercício 1.15 *Na Tabela 1.14 temos a população dos municípios de MG com mais de 50.000 habitantes, com base nos dados do Censo Demográfico 2000. Construa uma tabela de freqüências, trabalhando com as seguintes classes (em 1.000 hab.): [50,60), [60,70), [70,80), [80,100), [100,200), [200, 500) e 500 ou mais. Note que aqui estamos trabalhando com classes desiguais, o que é comum em situações desse tipo, onde há muitos observações pequenas e poucas grandes.*

Tabela 1.14: População dos municípios de MG com mais de 50.000 habitantes, para o Exercício 1.15

Município	População	Município	População	Município	População
Leopoldina	50.097	Timóteo	71.478	Varginha	108.998
Pirapora	50.300	Pará de Minas	73.007	Barbacena	114.126
três Pontas	51.024	Patrocínio	73.130	Sabará	115.352
São Francisco	51.497	Paracatu	75.216	Patos de Minas	123.881
Pedro Leopoldo	53.957	Vespasiano	76.422	Teófilo Otoni	129.424
Ponte Nova	55.303	Itaúna	76.862	Ibirité	133.044
S.Seb.do Paraíso	58.335	Caratinga	77.789	Poços de Caldas	135.627
Janaúba	61.651	S.João del Rei	78.616	Divinópolis	183.962
Formiga	62.907	Lavras	78.772	Sete Lagoas	184.871
Januária	63.605	Araxá	78.997	Santa Luzia	184.903
Cataguases	63.980	Itajubá	84.135	Ipatinga	212.496
Nova Lima	64.387	Ubá	85.065	Ribeirão das Neves	246.846
Viçosa	64.854	Ituiutaba	89.091	Gov.Valadares	247.131
Três Corações	65.291	Muriaé	92.101	Uberaba	252.051
Ouro Preto	66.277	Passos	97.211	Betim	306.675
João Monlevade	66.690	Cor. Fabriciano	97.451	Montes Claros	306.947
Alfenas	66.957	Itabira	98.322	Juiz de Fora	456.796
Manhuaçu	67.123	Araguari	101.974	Uberlândia	501.214
Curvelo	67.512	Cons.Lafaiete	102.836	Contagem	538.017
Unaí	70.033	Pouso Alegre	106.776	Belo Horizonte	2.238.526

Fonte: IBGE - Censo Demográfico 2000

Exercício 1.16 *Na Tabela 1.15 temos a densidade populacional (hab/km^2) das unidades da federação brasileira. Construa um gráfico ramo-e-folhas para esses dados. Para RJ e DF, você pode dar um "salto" na escala, de modo a não acrescentar muitos ramos vazios.*

Exercício 1.17 *Construa um gráfico de linhas para os dados da inflação brasileira anual medida pelo Índice Nacional de Preços ao Consumidor (INPC) apresentados na Tabela 1.16.*

Tabela 1.15: Densidade populacional dos estados brasileiros, para o Exercício 1.16

UF	Densidade Populacional (hab/km^2)	UF	Densidade Populacional (hab/km^2)
RO	6	SE	81
AC	4	BA	24
AM	2	MG	31
RR	2	ES	68
PA	5	RJ	328
AP	4	SP	149
TO	5	PR	48
MA	17	SC	57
PI	12	RS	37
CE	51	MS	6
RN	53	MT	3
PB	61	GO	15
PE	81	DF	353
AL	102		

Fonte: IBGE - Censo Demográfico 2000

Tabela 1.16: Índice Nacional de Preços ao Consumidor - 1995-2005

Ano	INPC (%)
1995	22,0
1996	9,1
1997	4,3
1998	2,5
1999	8,4
2000	5,3
2001	9,4
2002	14,7
2003	10,4
2004	6,1
2005	5,1

Capítulo 2

Principais Medidas Estatísticas

2.1 Introdução

A redução dos dados através de tabelas de freqüências ou gráficos é um dos meios disponíveis para se ilustrar o comportamento de um conjunto de dados. No entanto, muitas vezes queremos resumir ainda mais esses dados, apresentando um único valor que seja "representativo" do conjunto original. As medidas de posição ou tendência central, como o próprio nome está indicando, são medidas que informam sobre a posição típica dos dados. Na Figura 2.1 podemos notar os seguintes fatos: em (a) e (b), as distribuições são idênticas, exceto pelo fato de que a segunda está deslocada à direita. Em (c), podemos ver que há duas classes com a freqüência máxima e em (d), há uma grande concentração na cauda inferior e alguns poucos valores na cauda superior. As *medidas de posição* que apresentaremos a seguir irão captar essas diferenças.

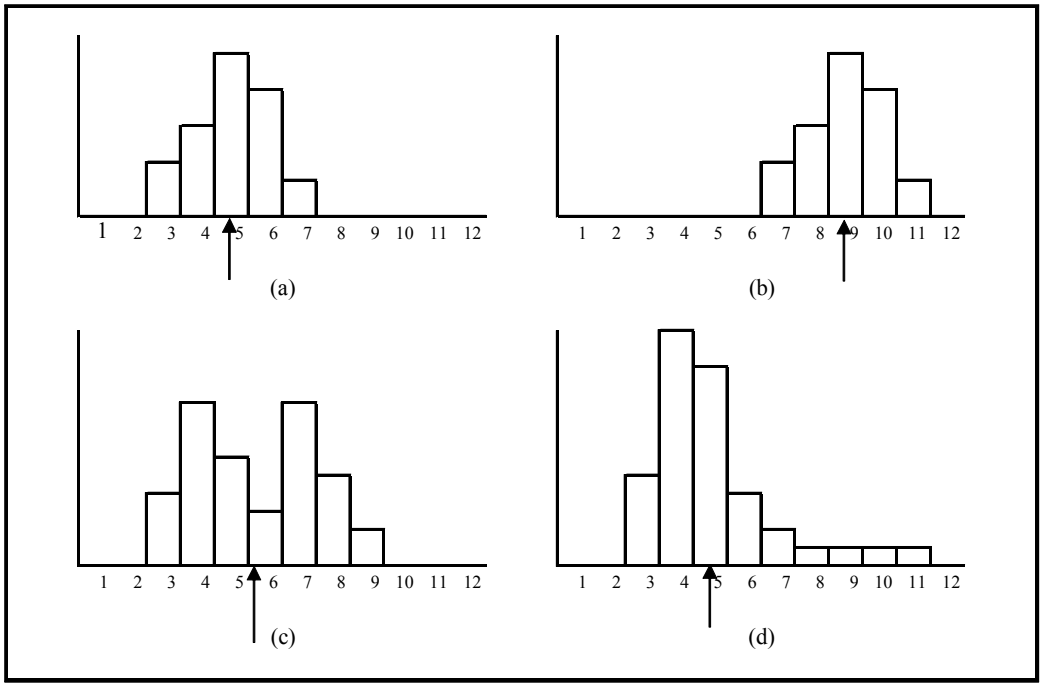


Figura 2.1: Exemplos ilustrativos do conceito de medidas de posição

Considere os conjuntos de dados apresentados por um *diagrama de pontos* na Figura 2.2. Nesse gráfico, as "pilhas" de pontos representam as freqüências de cada valor. Podemos ver facilmente que esses conjuntos têm o mesmo centro, mas diferem quanto à forma como os dados estão espalhados. As *medidas de dispersão* que iremos apresentar permitirão diferenciar esses três conjuntos de dados

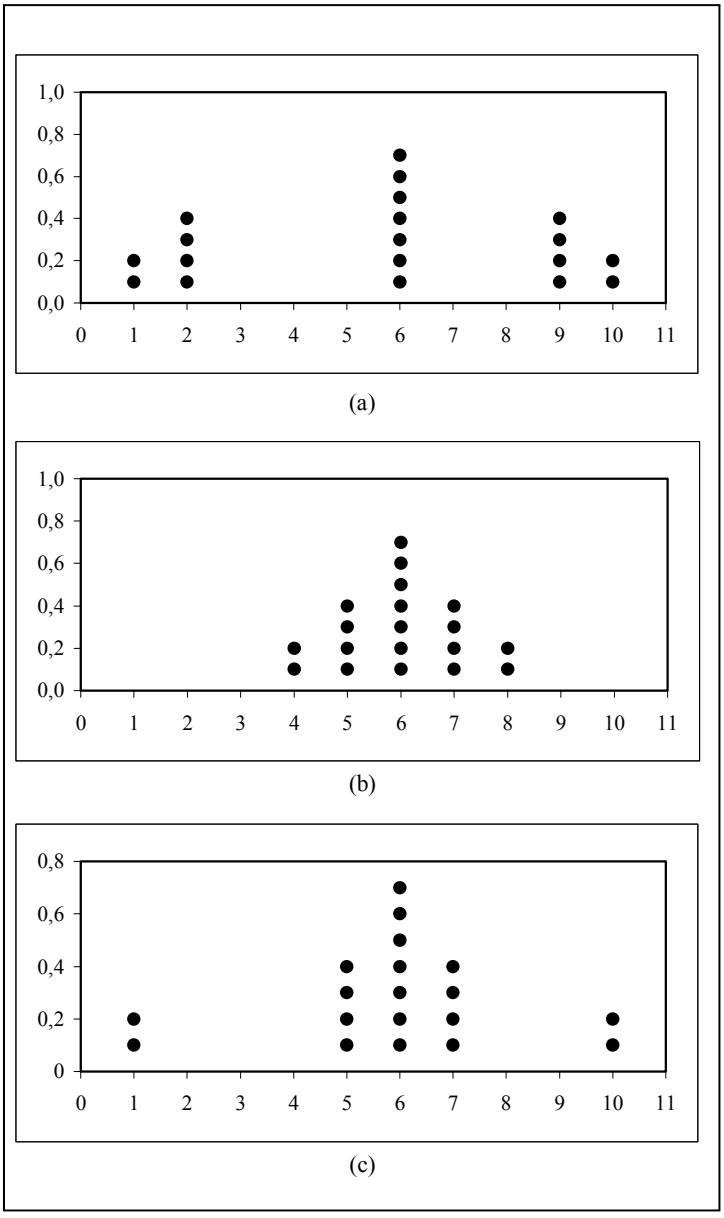


Figura 2.2: Exemplo ilustrativo do conceito de medidas de dispersão

2.2 Medidas de posição

2.2.1 Média aritmética simples

No nosso dia-a-dia, o conceito de média é bastante comum, quando nos referimos, por exemplo, à altura média dos brasileiros, à temperatura média dos últimos anos, etc.

Definição 3 *Dado um conjunto de n observações* $x_1, x_2, \ldots, x_n$*, a* ***média aritmética simples*** *é definida como*

$$\boxed{\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i}. \tag{2.1}$$

A notação $\overline{x}$ (lê-se x barra), usada para indicar a média, é bastante comum; em geral, usa-se a mesma letra utilizada para indicar os dados com a barra em cima. Na definição acima fazemos uso do símbolo de somatório, representado pela letra grega sigma maiúscula, Σ. Mais adiante você aprenderá mais sobre esse símbolo. Por enquanto, entenda como a média aritmética de um conjunto de dados é calculada. A primeira observação é que ela só pode ser calculada para dados quantitativos (não faz sentido somar masculino + feminino!). O seu cálculo é feito somando-se todos os valores e dividindo-se pelo número total de observações.

Consideremos as idades dos funcionários do Departamento de Recursos Humanos, analisadas no capítulo anterior e apresentadas no ramo e folhas da Figura 2.3.

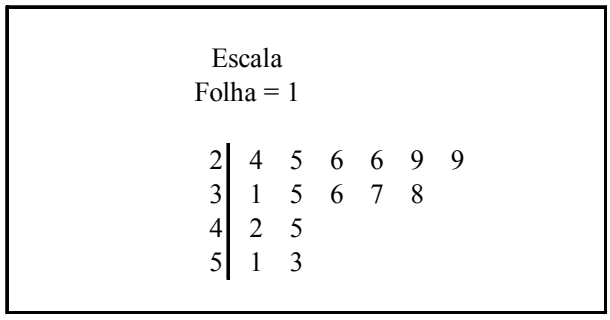


Figura 2.3: Idade dos funcionários do Departamento de RH

A idade média é:

$$\begin{aligned}\overline{x} &= \frac{24+25+26+26+29+29+31+35+36+37+38+42+45+51+53}{15} \\ &= \frac{527}{15} = 35,13\end{aligned}$$

Como as idades estão em anos, a idade média também é dada nessa unidade, ou seja, a idade média é 35,13 anos. Em geral, *a média de um conjunto de dados tem a mesma unidade dos dados originais.*

A interpretação física da média aritmética é que ela representa o centro de gravidade da distribuição; nos quatro histogramas da Figura 2.1, ela é o ponto de equilíbrio, indicado pela seta.

Note que o valor da média aritmética é um valor tal que, se substituíssemos todos os dados por ela, isto é, se todas as observações fossem iguais à média aritmética, a soma total seria igual à soma dos dados originais. Então, a média aritmética é uma forma de se distribuir o total observado pelos n elementos, de modo que todos tenham o mesmo valor. Considere os seguintes dados fictícios referentes aos salários de 5 funcionários de uma firma: $136, 210, 350, 360, 2500$. O total da folha de pagamentos é 3236, havendo um salário bastante alto, discrepante dos demais. A média para esses dados é 647,20. Se todos os 5 funcionários ganhassem esse salário, a folha de pagamentos seria a mesma e todos teriam o mesmo salário.

2.2.2 Moda

No histograma (c) da Figura 2.1, duas classes apresentam a mesma freqüência máxima. Esse é o conceito de *moda*.

Definição 4 *A **moda** de uma distribuição ou conjunto de dados, que representaremos por x^*, é o valor que mais se repete, ou seja, o valor mais freqüente.*

Podemos ter distribuições amodais (nenhum valor se repete), unimodais (uma moda), bimodais (duas modas), etc. Para os dados da Figura 2.3 temos as seguintes modas: $x^* = 26$ e $x^* = 29$ anos e, portanto, essa é uma distribuição bimodal. Assim como a média, *a moda sempre tem a mesma unidade dos dados originais.* É interessante observar que a moda é a única medida de posição que pode ser calculada tanto para dados qualitativos, quanto para dados quantitativos.

2.2.3 Mediana

Vamos analisar novamente os seguintes dados referentes aos salários (em R$) de 5 funcionários de uma firma: $136, 210, 350, 360, 2500$. Como visto, o salário médio é R$ 647,20. No entanto, esse valor não representa bem nem os salários mais baixos, nem o salário mais alto. Isso acontece porque o salário mais alto é muito diferente dos demais. Esse exemplo ilustra um fato geral sobre a média aritmética: ela é muito influenciada por *valores discrepantes* (em inglês, *outliers*), isto é, valores muito grandes (ou muito pequenos) que sejam distintos da maior parte dos dados. Essa situação está ilustrada pelo histograma na parte (d) da Figura 2.1: há uma grande concentração na cauda inferior e alguns poucos valores na cauda superior. Nesses casos é necessário utilizar uma outra medida de posição para representar o conjunto; uma medida possível é a *mediana*.

Definição 5 *Seja $x_1, x_2, \ldots, x_n$ um conjunto de n observações e seja $x_{(i)}, i = 1, \ldots, n$ o conjunto das observações ordenadas, de modo que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Então, a **mediana** Q_2 é definida como o valor tal que 50% das observações são menores que ela e 50% são maiores que ela. Para efeito de cálculo, valem as seguintes regras:*

$$\begin{aligned} n \text{ ímpar}: &\qquad Q_2 = x_{\left(\frac{n+1}{2}\right)} \\ n \text{ par}: &\qquad Q_2 = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \end{aligned} \tag{2.2}$$

Dessa definição, podemos ver que a mediana é o valor central dos dados e para calculá-la é necessário ordenar os dados. Para as idades na Figura 2.3, temos que o número total de observações é $n = 15$. Logo, a mediana é o valor central, que deixa 7 observações abaixo e 7 observações acima. Logo, a mediana é a oitava observação, uma vez que

$$\frac{n+1}{2} = \frac{15+1}{2} = 8$$

Sendo assim, a idade mediana é $Q_2 = 35$ anos. *A unidade da mediana é a mesma dos dados* e ela só pode ser calculada para dados quantitativos..

2.2.4 Separatrizes

A mediana é um caso particular de um conjunto mais amplo de medidas estatísticas, chamadas *separatrizes*. A separatriz de ordem p é o valor que deixa pelo menos $p\%$ dos dados abaixo dele e pelo menos $(1-p)\%$ acima dele. As separatrizes mais comuns são os *quartis*, *decis* e *percentis*, cujos fatores de divisão são 4, 10 e 100. Mais precisamente, existem 3 quartis, 9 decis e 99 percentis. Os quartis serão representados pela letra Q e são eles:

- primeiro quartil Q_1 : deixa pelo menos 25% das observações abaixo dele e pelo menos 75% acima;
- segundo quartil Q_2 : deixa pelo menos 50% das observações abaixo dele e pelo menos 50% acima; é a mediana;
- terceiro quartil Q_3 : deixa pelo menos 75% das observações abaixo dele e pelo menos 25% acima.

Os decis serão representados pela letra D e os percentis pela letra P; assim, por exemplo:

- o terceiro decil D_3 deixa pelo menos 30% das observações abaixo e pelo menos 70% acima;
- o quinto decil e o $50^{\underline{o}}$ percentil são a mediana;
- o octagésimo percentil deixa pelo menos 80% das observações abaixo e pelo menos 20% acima.

No cálculo das separatrizes quase sempre será necessário algum procedimento de arredondamento e aproximação. Consideremos novamente as idades dos 15 funcionários do Departamento de Recursos Humanos apresentadas no ramo e folhas da Figura 2.3.

Cálculo dos quartis

Para os quartis, podemos adotar o seguinte procedimento: depois de calculada a mediana, considere as duas partes dos dados, a parte abaixo da mediana e a parte acima da mediana, em ambos os casos excluindo a mediana. O primeiro quartil é calculado como a mediana da parte abaixo da mediana original e o terceiro quartil é calculado como a mediana da parte acima da mediana original. Para os dados acima, temos 15 observações. A mediana é a oitava observação. Então, as duas partes consistem nas 7 observações inferiores e nas 7 observações superiores, respectivamente (ver Figura 2.4). Como 7 é um número ímpar, a mediana é o valor central de cada uma dessas partes. Assim,

o primeiro quartil é a quarta observação da parte inferior e o terceiro quartil é a quarta observação da parte superior. Em termos dos dados completos, o primeiro quartil é a quarta observação e o terceiro quartil é a quarta observação a partir da mediana, ou seja, o terceiro quartil é a $(8+4) = 12^{\underline{a}}$ observação. Resulta, então, que $Q_1 = 26$ e $Q_3 = 42$.

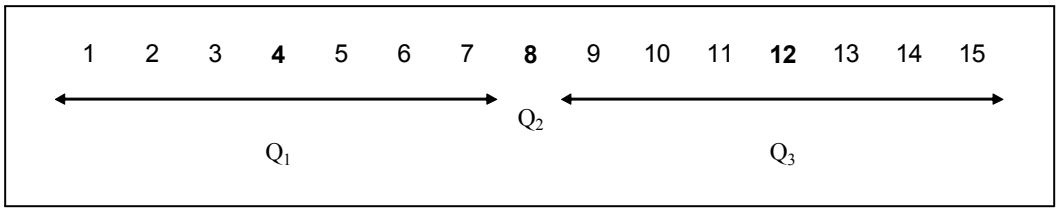


Figura 2.4: Cálculo dos quartis

Cálculo de separatrizes em geral

Vamos ilustrar o procedimento que adotaremos para calcular separatrizes de qualquer ordem com os dados acima. Se queremos calcular o primeiro decil, temos que ver que 10% de 15 observações correspondem a 1,5 observações, ou seja, o primeiro decil deixa 1,5 observações abaixo dele. Nesse caso, definimos o primeiro decil como sendo a segunda observação, ou seja, arredondamos 1,5 para 2. Para os nossos dados, $D_1 = 25$, a segunda menor observação. Por simetria, devemos tomar $D_9 = 51$, a segunda observação começando do máximo.

Para o segundo decil, temos que 20% de 15 observações correspondem a 3 observações, ou seja, o terceiro decil deixa 3 observações abaixo dele. Nesse caso, definimos o segundo decil como a média da terceira e quarta observações, ou seja, $D_2 = \frac{26+26}{2} = 26$. Por simetria, $D_8 = \frac{42+45}{2} = 43,5$.

Como regra geral, vamos adotar o seguinte procedimento:

Regra 3 *Cálculo de separatrizes*

Seja p a separatriz desejada em forma decimal; por exemplo, $p = 0,1$ para o cálculo de D_1, $p = 0,9$ para o cálculo de P_{90}, etc.

1. *Calcule $i = p \times n$, em que n é o número de observações.*
2. *Se i não for um inteiro, arredonde para o próximo maior inteiro para obter a posição da separatriz desejada.*
3. *Se i for inteiro, então a separatriz de ordem p é a média das observações de ordem i e $i+1$ na lista ordenada dos dados.*

A título de ilustração, vamos calcular P_{60} para as idades dos funcionários do Departamento de Recursos Humanos. Temos que

$$0,6 \times 15 = 9$$

Como o resultado é um número inteiro, P_{60} será a média da nona e da décima observações:

$$P_{60} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{36 + 37}{2} = 36,5$$

Para D_7, temos que

$$0,7 \times 15 = 10,5$$

Como o resultado não é um número inteiro, arredondamos para cima e D_7 é a décima primeira observação, ou seja:

$$D_7 = x_{(11)} = 38$$

2.2.5 Exemplo

Considere as idades dos 23 funcionários do Departamento Financeiro, analisadas no Exercício 1.9 do capítulo anterior, cujos valores são:. 27; 31; 45; 52; 33; 34; 29; 27; 35; 38; 50; 48; 29; 30; 32; 29; 42; 41; 40; 42; 28; 36; 48. Vamos calcular as medidas de posição para esses dados, os quartis, P_{20} e P_{90}. Ordenando os dados, temos o seguinte:

27	27	28	29	29	29	30	31	32	33	34	35
36	38	40	41	42	42	45	48	48	50	52	

A média é

$$\begin{aligned} \overline{x} &= \frac{2 \times 27 + 28 + 3 \times 29 + 30 + 31 + \cdots + 2 \times 42 + 45 + 2 \times 48 + 50 + 52}{23} \\ &= 36,78 \text{ anos} \end{aligned}$$

A moda é $x^* = 29$ anos e a mediana (n ímpar) é

$$Q_2 = x_{\left(\frac{23+1}{2}\right)} = x_{(12)} = 35 \text{ anos}$$

Com relação aos quartis, temos o seguinte: com 23 observações, temos 11 observações acima e abaixo da mediana. Com 11 observações, a mediana é a sexta observação; logo,

$$Q_1 = x_{(6)} = 29 \text{ anos}$$

e

$$Q_3 = x_{(12+6)} = x_{(18)} = 42 \text{ anos}$$

Note a simetria de Q_1 e Q_3 : abaixo de Q_1 temos 5 observações e acima de Q_3 temos 5 observações.

Para o cálculo de P_{20} temos que $0,20 \times 23 = 4,6$; logo, P_{20} é a quinta observação:

$$P_{20} = x_{(5)} = 29 \text{ anos}$$

Por simetria, $P_{80} = 45$ anos. Para P_{90}, temos que $0,9 \times 23 = 20,7$. Logo, $P_{90} = x_{(21)} = 48$ e por simetria, $P_{10} = x_{(3)} = 28$ anos.

Exercício 2.1 *Considere novamente os dados sobre os salários dos funcionários do Departamento de Recursos Humanos, cujos valores (em R$) são os seguintes:*

6300	*5700*	*4500*	*3800*	*3200*	*7300*	*7100*	*5600*
6400	*7000*	*3700*	*6500*	*4000*	*5100*	*4500*	

Calcule a média, a moda e a mediana para esses dados, especificando as respectivas unidades.

Exercício 2.2 *Calcule a nota média, a nota modal, os quartis e os decis para os dados da Tabela 2.1.*

Tabela 2.1: Notas de 50 alunos para o Exercício 2.2

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

2.2.6 Somatório

A notação de somatório é bastante útil na apresentação de fórmulas, pois ele resume de forma bastante compacta a operação de soma de várias parcelas. Para compreender as propriedades do somatório, basta lembrar as propriedades da adição.

Para desenvolver um somatório, temos que substituir o valor do índice em cada uma das parcelas e em seguida realizar a soma dessas parcelas. Por exemplo:

$$\sum_{i=1}^{5} i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2$$

Em termos mais gerais, temos as seguintes propriedades:

$$\begin{aligned}\sum_{i=1}^{n}(x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n) = \\ &= (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n) = \\ &= \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} kx_i &= kx_1 + kx_2 + \cdots + kx_n = \\ &= k(x_1 + x_2 + \cdots + x_n) = \\ &= k\sum_{i=1}^{n} x_i\end{aligned}$$

$$\sum_{i=1}^{n} k = k + k + \cdots + k = nk$$

É importante salientar algumas diferenças:

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$$

uma vez que

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

e

$$\left(\sum_{i=1}^{n} x_i\right)^2 = (x_1 + x_2 + \cdots + x_n)^2$$

Temos também que

$$\sum_{i=1}^{n} x_i y_i \neq \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

uma vez que

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

e

$$\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = (x_1 + x_2 + \cdots + x_n)(y_1 + y_2 + \cdots + y_n)$$

À medida do necessário iremos apresentando mais propriedades do somatório.

Exercício 2.3 *Calcule as seguintes quantidades para os dados abaixo:*

$$\sum_{i=1}^{6} x_i \qquad \sum_{i=1}^{6} f_i \qquad \sum_{i=1}^{6} f_i x_i \qquad \sum_{i=1}^{6} f_i x_i^2$$

i	1	2	3	4	5	6
f_i	3	5	9	10	2	1
x_i	10	11	15	19	21	26

2.2.7 Média aritmética ponderada

Vimos que a média aritmética equivale a dividir o "todo" (soma dos valores) em partes iguais, ou seja, estamos supondo que os números que queremos sintetizar têm o mesmo grau de importância. Entretanto, há algumas situações onde não é razoável atribuir a mesma importância para todos os dados. Por exemplo, o Índice Nacional de Preços ao Consumidor (INPC) é calculado com uma média dos Índices de Preço ao Consumidor (IPC) de diversas regiões metropolitanas do Brasil, mas a importância dessas regiões é diferente. Uma das variáveis que as diferencia é a população residente.

Nesse tipo de situação, em vez de se usar a média aritmética simples, usa-se a *média aritmética ponderada*, que será representada por $\overline{x}_p$.

Definição 6 *A **média aritmética ponderada** de números $x_1, x_2, \ldots, x_n$ com pesos $\rho_1, \rho_2, \ldots, \rho_n$ é definida como*

$$\overline{x}_p = \frac{\rho_1 x_1 + \rho_2 x_2 + \cdots + \rho_n x_n}{\rho_1 + \rho_2 + \ldots + \rho_n} = \frac{\displaystyle\sum_{i=1}^{n} \rho_i x_i}{\displaystyle\sum_{i=1}^{n} \rho_i}. \tag{2.3}$$

Se definimos

$$\omega_i = \frac{\rho_i}{\displaystyle\sum_{j=1}^{n} \rho_j} \tag{2.4}$$

então a média aritmética ponderada pode ser reescrita como

$$\overline{x}_p = \sum_{i=1}^{n} \omega_i x_i \tag{2.5}$$

onde $\sum_{i=1}^{n} \omega_i = 1$.

Note que a média aritmética simples é um caso particular da média aritmética ponderada, onde todas as observações têm o mesmo peso $\omega_i = \frac{1}{n}$.

Para a construção do Índice Nacional de Preços ao Consumidor - INPC, o peso de cada índice regional é definido pela população residente urbana, conforme dados da Tabela 2.2. Os pesos em porcentagem aí apresentados representam a participação da população residente urbana da região metropolitana no total da população residente urbana das 11 regiões metropolitanas pesquisadas. O índice geral é dado pela média ponderada:

$$\begin{aligned} \text{INPC}_{03/06} &= 0,0306 \times 0,75 + 0,0915 \times 0,64 + 0,0623 \times 0,55 + 0,0919 \times 0,52 + \\ &\quad 0,0749 \times 0,50 + 0,0425 \times 0,48 + 0,0378 \times 0,48 + 0,0385 \times 0,44 + \\ &\quad 0,3626 \times 0,37 + 0,0334 \times 0,37 + 0,1340 \times 0,18 \\ &= 0,427137 \end{aligned}$$

Tabela 2.2: Estrutura básica de ponderação regional para cálculo do INPC - Março 2006

Área Geográfica	Peso (%)	IPC - Mar/06
Brasília	3,06	0,75
Belo Horizonte	9,15	0,64
Salvador	6,23	0,55
Porto Alegre	9,19	0,52
Curitiba	7,49	0,50
Recife	4,25	0,48
Goiânia	3,78	0,48
Belém	3,85	0,44
São Paulo	36,26	0,37
Fortaleza	3,34	0,37
Rio de Janeiro	13,40	0,18
INPC - Geral		0,42

Fonte: IBGE

Exercício 2.4 *Segundo o critério de avaliação adotado pelo Departamento de Estatística, cada aluno será submetido a 2 provas, a primeira tendo peso 2 e a segunda tendo peso 3. Para ser aprovado sem ter que fazer prova final, a média nas 2 provas tem que ser, no mínimo, 6. Se um aluno tirar 5,5 na primeira prova, quanto deverá tirar na segunda prova para não ter que fazer prova final? E se as provas tivessem o mesmo peso?*

2.2.8 Propriedades das medidas de posição

Da interpretação física de média como centro de gravidade da distribuição, fica claro que a média é sempre um valor situado entre os valores mínimo e máximo dos dados. O mesmo resultado vale para a mediana e a moda, o que é imediato a partir das respectivas definições.

Propriedade 1

$$\begin{aligned} x_{\min} &\leq \overline{x} \leq x_{\max} \\ x_{\min} &\leq Q_2 \leq x_{\max} \\ x_{\min} &\leq x^* \leq x_{\max} \end{aligned} \tag{2.6}$$

Vamos apresentar as outras duas propriedades através do seguinte exemplo. Em uma turma de Estatística, os resultados de uma prova ficaram abaixo do que a professora esperava. Como todos os alunos vinham participando ativamente de todas as atividades, mostrando um interesse especial pela matéria, a professora resolveu dar 1 ponto na prova para todos os alunos. Além disso, ela deu os resultados com as notas variando de 0 a 10, mas a Secretaria da Faculdade exige que as notas sejam dadas em uma escala de 0 a 100. Sendo assim, a professora precisa multiplicar todas as notas por 10. O que acontece com a média, a moda e a mediana depois dessas alterações? Vamos ver isso com um conjunto de 5 notas: $5, 4, 2, 3, 4$. As notas ordenadas são $2, 3, 4, 4, 5$ e temos as seguintes medidas de posição:

$$\begin{aligned} \overline{x} &= \frac{5+4+2+3+4}{5} = \frac{18}{5} = 3,6 \\ Q_2 &= x^* = 4 \end{aligned}$$

Somando 1 ponto, as notas passam a ser $3, 4, 5, 5, 6$ com as seguintes medidas de posição:

$$\begin{aligned} \overline{y} &= \frac{3+4+5+5+6}{5} = \frac{23}{5} = 4,6 = 3,6+1 \\ Q_{2,y} &= y^* = 5 = 4+1 \end{aligned}$$

Ao somar 1 ponto em todas as notas, o conjunto de notas sofre uma translação, o que faz com que o seu centro também fique deslocado de 1 ponto. Sendo assim, todas as três medidas de posição ficam somadas de 1 ponto.

Multiplicando as novas notas por 10, obtemos $30, 40, 50, 50, 60$ e

$$\begin{aligned} \overline{z} &= \frac{30+40+50+50+60}{5} = \frac{230}{5} = 46,0 = 4,6 \times 10 \\ Q_{2,y} &= y^* = 50 = 5 \times 10 \end{aligned}$$

ou seja, todas as medidas de posição ficam multiplicadas por 10.

Esse exemplo ilustra as seguintes propriedades.

Propriedade 2

Somando-se um mesmo valor a cada observação x_i, obtemos um novo conjunto de dados $y_i = x_i + k$ para o qual temos as seguintes medidas de posição:

$$y_i = x_i + k \Rightarrow \begin{cases} \overline{y} = \overline{x} + k \\ Q_{2,y} = Q_{2,x} + k \\ y^* = x^* + k \end{cases} \tag{2.7}$$

Propriedade 3

Multiplicando cada observação x_i por uma mesma constante não nula k, obtemos um novo conjunto de dados $y_i = kx_i$ para o qual temos as seguintes medidas de posição:

$$y_i = kx_i \Rightarrow \begin{cases} \overline{y} = k\overline{x} \\ Q_{2,y} = kQ_{2,x} \\ y^* = kx^* \end{cases} \tag{2.8}$$

Exercício 2.5 *A relação entre as escalas Celsius e Fahrenheit é a seguinte:*

$$C = \frac{5}{9}(F - 32)$$

Se a temperatura média em determinada localidade é de $45°F$*, qual é a temperatura média em graus Celsius?*

Exercício 2.6 *Em uma certa pesquisa, foram levantados dados sobre o lucro líquido de uma amostra de grandes empresas, em reais, obtendo-se a média de R$ 1 035 420,00. Na divulgação dos resultados, os valores devem ser apresentados em milhares de reais. Qual é o valor a ser divulgado para o lucro médio?*

2.3 Medidas de dispersão

2.3.1 Amplitude

Considere novamente os conjuntos de dados na Figura 2.2. Uma forma de diferenciar os conjuntos (a) e (b) é através da *amplitude* dos dados, que é, como já visto, a diferença entre o maior e o menor valor.

Definição 7 *A* ***amplitude*** *de um conjunto de dados é a distância entre o maior valor e o menor valor.*

$$\Delta_{total} = V_{\max} - V_{\min}. \tag{2.9}$$

A amplitude tem a mesma unidade dos dados, mas ela tem algumas limitações, conforme ilustrado nas partes (a) e (c) da Figura 2.2. Lá, os dois conjuntos têm a mesma média, a mesma mediana e a mesma amplitude, mas essas medidas não conseguem caracterizar o fato de a distribuição dos valores entre o mínimo e o máximo ser diferente nos dois conjuntos. A limitação da amplitude também fica patente pelo fato de ela se basear em apenas duas observações, independentemente do número total de observações.

2.3.2 Desvio médio absoluto

Uma maneira de se medir a dispersão dos dados é considerar os tamanhos dos *desvios* $x_i - \overline{x}$ de cada observação em relação à média. Note, na Figura 2.2, que, quanto mais disperso o conjunto de dados, maiores esses desvios tendem a ser. Para obter uma medida-resumo, isto é, um único número, poderíamos somar esses desvios, ou seja, considerar a seguinte medida:

$$D = \sum_{i=1}^{n}(x_i - \overline{x}). \tag{2.10}$$

Vamos desenvolver tal fórmula, usando as propriedades de somatório e a definição da média amostral.

$$\begin{aligned} D &= \sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = \sum_{i=1}^{n} x_i - n\overline{x} = \\ &= \sum_{i=1}^{n} x_i - n \times \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0. \end{aligned}$$

Ou seja: essa medida, que representa a soma dos desvios em relação à média, é sempre nula, não importa o conjunto de dados! Logo, ela não serve para diferenciar quaisquer conjuntos!

Vamos dar uma explicação intuitiva para esse fato, que nos permitirá obter correções para tal fórmula. Ao considerarmos as diferenças entre cada valor e o valor médio, obtemos desvios negativos e positivos, pois, pela definição de média, sempre existem valores menores e maiores que a média; esses desvios positivos e negativos, ao serem somados, se anulam.

Bom, se o problema está no fato de termos desvios positivos e negativos, por que não trabalhar com o valor absoluto das diferenças? De fato, esse procedimento nos leva à definição de *desvio médio absoluto.*

Definição 8 *O* ***desvio médio absoluto*** *de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por*

$$DMA = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \tag{2.11}$$

onde as barras verticais representam o valor absoluto ou módulo.

Note que nesta definição estamos trabalhando com o desvio médio, isto é, tomamos a média dos desvios absolutos. Isso evita interpretações equivocadas, pois, se trabalhássemos apenas com a soma dos desvios absolutos, um conjunto com um número maior de observações tenderia a apresentar um resultado maior para a soma devido apenas ao fato de ter mais observações. Esta situação é ilustrada com os seguintes conjuntos de dados:

- Conjunto 1: $\{1, 3, 5\}$
- Conjunto 2: $\left\{1, \frac{5}{3}, 3, \frac{13}{3}, 5\right\}$

Para os dois conjuntos, $\overline{x} = 3$ e para o conjunto 1

$$\sum_{i=1}^{3}|x_i - \overline{x}| = |1-3| + |3-3| + |5-3| = 4$$

e para o conjunto 2

$$\sum_{i=1}^{5} |x_i - \overline{x}| = |1-3| + \left|\frac{5}{3} - 3\right| + |3-3| + \left|\frac{13}{3} - 3\right| + |5-3| = \frac{20}{3} = 6,667.$$

Então, o somatório para o segundo conjunto é maior, mas o desvio absoluto médio é o mesmo para ambos; de fato, para o primeiro conjunto temos

$$DMA = \frac{4}{3}$$

e para o segundo conjunto

$$DMA = \frac{\frac{20}{3}}{5} = \frac{4}{3}$$

Ao dividirmos o somatório pelo número de observações, compensamos o fato de o segundo conjunto ter mais observações que o primeiro.

O desvio médio absoluto tem a mesma unidade dos dados.

Exercício 2.7 *Para o conjunto de dados* $2, 4, 7, 8, 9, 6, 5, 8$, *calcule os desvios em torno da média e verifique que eles somam zero. Em seguida, calcule o desvio médio absoluto.*

2.3.3 Variância e desvio padrão

Considerar o valor absoluto das diferenças $(x_i - \overline{x})$ é uma das maneiras de se contornar o fato de que $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. No entanto, a função módulo tem a desvantagem de ser não diferenciável no ponto zero. Outra possibilidade de correção, com propriedades matemáticas e estatísticas mais adequadas, é considerar o quadrado das diferenças. Isso nos leva à definição de *variância*.

Definição 9 *A* ***variância***[1] *de um conjunto de dados* $x_1, x_2, \ldots, x_n$ *é definida por*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 . \tag{2.12}$$

Note que esta definição de variância nos diz que a variância é *a média dos desvios quadráticos*.

Suponhamos que os valores x_i representem os pesos, em quilogramas, de um conjunto de pessoas. Então, o valor médio $\overline{x}$ representa o peso médio dessas pessoas e sua unidade também é quilogramas, o mesmo acontecendo com as diferenças $(x_i - \overline{x})$. Ao elevarmos essas diferenças ao quadrado, passamos a ter a variância medida em quilogramas ao quadrado, uma unidade que não tem interpretação física. Uma forma de se obter uma medida de dispersão com a mesma unidade dos dados consiste em tomar a raiz quadrada da variância.

[1] É possível definir a variância usando o divisor $n-1$ no lugar de n; essa é a diferença entre os conceitos de variância populacional e variância amostral, que será mais relevante num segundo curso de estatística.

Definição 10 *O **desvio padrão** de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por*

$$\sigma = \sqrt{Variância} = \sqrt{\sigma^2} \tag{2.13}$$

A título de ilustração, vamos considerar novamente os dados analisados anteriormente, referentes à idade dos funcionários do Departamento de Recursos Humanos. Essas idades são:

$$24 \quad 25 \quad 26 \quad 26 \quad 29 \quad 29 \quad 31 \quad 35 \quad 36 \quad 37 \quad 38 \quad 42 \quad 45 \quad 51 \quad 53$$

e sua média é $\frac{527}{15} = 35,1\overline{3}$. Assim, a variância, em anos^2 é

$$\begin{aligned}\sigma^2 &= \frac{1}{15}\left[\begin{array}{c}(24-35,13)^2+(25-35,13)^2+2\times(26-35,13)^2+2\times(29-35,13)^2+\\(31-35,13)^2+(35-35,13)^2+(36-35,13)^2+(37-35,13)^2+(38-35,13)^2+\\(42-35,13)^2+(42-35,13)^2+(45-35,13)^2+(51-35,13)^2+(53-35,13)^2\end{array}\right] = \\ &= \frac{1213,73}{15} = 80,92\end{aligned}$$

e o desvio padrão, em anos, é

$$\sigma = \sqrt{80,92} = 8,995$$

Exercício 2.8 *Para o conjunto de dados do Exercício 2.7 − $\{2,4,7,8,9,6,5,8\}$ − calcule a variância e o desvio padrão.*

2.3.4 Fórmula alternativa para o cálculo da variância

Consideremos a Equação (2.12) que define a variância. Desenvolvendo o quadrado e usando as propriedades de somatório, obtemos:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}2\overline{x}x_i + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) + \frac{1}{n}n\overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}^2 + \overline{x}^2\end{aligned}$$

ou seja

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 \tag{2.14}$$

Essa forma de escrever a variância facilita quando os cálculos têm que ser feitos à mão ou em calculadoras menos sofisticadas, pois o número de cálculos envolvidos é menor. Note que ela nos diz que a variância é a *média dos quadrados menos o quadrado da média.*

Vamos calcular a variância das idades dos funcionários de RH usando essa fórmula:

$$\begin{aligned}\sigma^2 &= \frac{1}{15}\left[\begin{array}{c}24^2+25^2+25^2+2\times26^2+2\times29^2+31^2+35^2+36^2+\\37^2+38^2+39^2+42^2+45^2+51^2+53^2\end{array}\right] - \left(\frac{527}{15}\right)^2 = \\ &= \frac{19729\times15-527^2}{15^2} = \frac{295935-277729}{225} = \frac{18206}{225} = 80,916\end{aligned}$$

Na comparação dos resultados obtidos pelas duas fórmulas, pode haver alguma diferença por causa dos arredondamentos, uma vez que a média é uma dízima.

Exercício 2.9 *No Exercício 2.8 você calculou a variância do conjunto de dados* $\{2, 4, 7, 8, 9, 6, 5, 8\}$ *como a média dos desvios quadráticos. Calcule a variância novamente utilizando a fórmula alternativa dada na Equação (2.14).*

2.3.5 Exemplo

Vamos considerar novamente os dados referentes ao número de dependentes dos funcionários do Departamento de Recursos Humanos, apresentados novamente na tabela a seguir.

Nome	No.de dependentes	Nome	No.de dependentes
João da Silva	3	Patrícia Silva	2
Pedro Fernandes	1	Regina Lima	2
Maria Freitas	0	Alfredo Souza	3
Paula Gonçalves	0	Margarete Cunha	0
Ana Freitas	1	Pedro Barbosa	2
Luiz Costa	3	Ricardo Alves	0
André Souza	4	Márcio Rezende	1
Ana Carolina Chaves	0		

Como o menor valor é 0 e o maior valor é 4, temos que a amplitude dos dados é de 4 dependentes. A média calculada para esses dados foi $\overline{x} = \frac{22}{15} = 1,467$. Vamos calcular a soma dos desvios em torno da média, usando o fato de que temos observações repetidas.

$$\begin{aligned}\sum(x_i - \overline{x}) &= 5 \times \left(0 - \frac{22}{15}\right) + 3 \times \left(1 - \frac{22}{15}\right) + 3 \times \left(2 - \frac{22}{15}\right) + 3 \times \left(3 - \frac{22}{15}\right) + \left(4 - \frac{22}{15}\right) = \\ &= -\frac{110}{15} - \frac{21}{15} + \frac{24}{15} + \frac{69}{15} + \frac{38}{15} = -\frac{131}{15} + \frac{131}{15} = 0\end{aligned}$$

Caso trabalhássemos com o valor aproximado 1,467, o resultado aproximado seria -0,005.

O desvio médio absoluto é

$$\begin{aligned}DMA &= \frac{1}{n}\sum|x_i - \overline{x}| = \\ &= \frac{1}{15} \times \left[5 \times \left|0 - \frac{22}{15}\right| + 3 \times \left|1 - \frac{22}{15}\right| + 3 \times \left|2 - \frac{22}{15}\right| + 3 \times \left|3 - \frac{22}{15}\right| + \left|4 - \frac{22}{15}\right|\right] = \\ &= \frac{1}{15} \times \left[\frac{110}{15} + \frac{21}{15} + \frac{24}{15} + \frac{69}{15} + \frac{38}{15}\right] = \frac{1}{15} \times \left[\frac{131}{15} + \frac{131}{15}\right] = \frac{262}{225} = 1,1644\end{aligned}$$

A variância é

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum(x_i - \overline{x})^2 \\ &= \frac{1}{15} \times \left[5 \times \left(0 - \frac{22}{15}\right)^2 + 3 \times \left(1 - \frac{22}{15}\right)^2 + 3 \times \left(2 - \frac{22}{15}\right)^2 + 3 \times \left(3 - \frac{22}{15}\right)^2 + \left(4 - \frac{22}{15}\right)^2\right] = \\ &= \frac{1}{15} \times \left[\frac{2420}{225} + \frac{147}{225} + \frac{192}{225} + \frac{1587}{225} + \frac{1444}{225}\right] = \frac{5790}{15 \times 225} = 1,715556\end{aligned}$$

e

$$\sigma = \sqrt{\frac{5790}{15 \times 225}} = 1,3098$$

Vamos agora calcular a variância usando a fórmula alternativa:

$$\begin{aligned} \sigma^2 &= \frac{1}{15} \times \left(5 \times 0^2 + 3 \times 1^2 + 3 \times 2^2 + 3 \times 3^2 + 4^2\right) - \left(\frac{22}{15}\right)^2 = \\ &= \frac{3 + 12 + 27 + 16}{15} - \frac{484}{225} = \frac{58}{15} - \frac{484}{225} = \frac{58 \times 15 - 484}{225} = \frac{386}{225} = 1,715556 \end{aligned}$$

Note que com essa fórmula os cálculos ficam bem mais simples, uma vez que temos que fazer menos conta!

2.3.6 Propriedades das medidas de dispersão

Como visto para as medidas de posição, vamos ver as principais propriedades das medidas de dispersão.

Propriedade 1
Todas as medidas de dispersão são não negativas!

$$\begin{array}{l} \Delta \geq 0 \\ DMA \geq 0 \\ \sigma^2 \geq 0 \\ \sigma \geq 0 \end{array} \tag{2.15}$$

Propriedade 2
Somando-se uma mesma constante a todas as observações, as medidas de dispersão não se alteram. Essa propriedade é bastante intuitiva se notarmos que, ao somar uma constante aos dados, estamos simplesmente fazendo uma translação dos mesmos, sem alterar a dispersão.

$$y_i = x_i + k \Rightarrow \left\{ \begin{array}{c} \Delta_y = \Delta_x \\ DMA_y = DMA_x \\ \sigma_y^2 = \sigma_x^2 \\ \sigma_y = \sigma_x \end{array} \right. \tag{2.16}$$

Propriedade 3
Ao multiplicarmos todos os dados por uma constante não nula temos que:

$$y_i = kx_i \Rightarrow \left\{ \begin{array}{c} \Delta_y = |k|\ \Delta_x \\ DMA_y = |k|\ DMA_x \\ \sigma_y^2 = k^2 \sigma_x^2 \\ \sigma_y = |k|\ \sigma_x \end{array} \right. \tag{2.17}$$

Note que é razoável que apareça o módulo da constante, já que as medidas de dispersão são não negativas.

Exercício 2.10 *Se o desvio padrão das temperaturas diárias de uma determinada localidade é de* $5,2°F$*, qual é o desvio padrão em graus Celsius? Lembre-se que a relação entre as duas escalas é*

$$C = \frac{5}{9}(F - 32)$$

2.3.7 Intervalo interquartil

Na Figura 2.5 temos uma ilustração da definição de quartis. Analisando essa figura, podemos ver que entre Q_1 e Q_3, há sempre 50% dos dados, qualquer que seja a distribuição. Assim, quanto maior for a distância entre Q_1 e Q_3, mais dispersos serão os dados. Temos, assim, uma nova medida de dispersão, o *intervalo interquartil*.

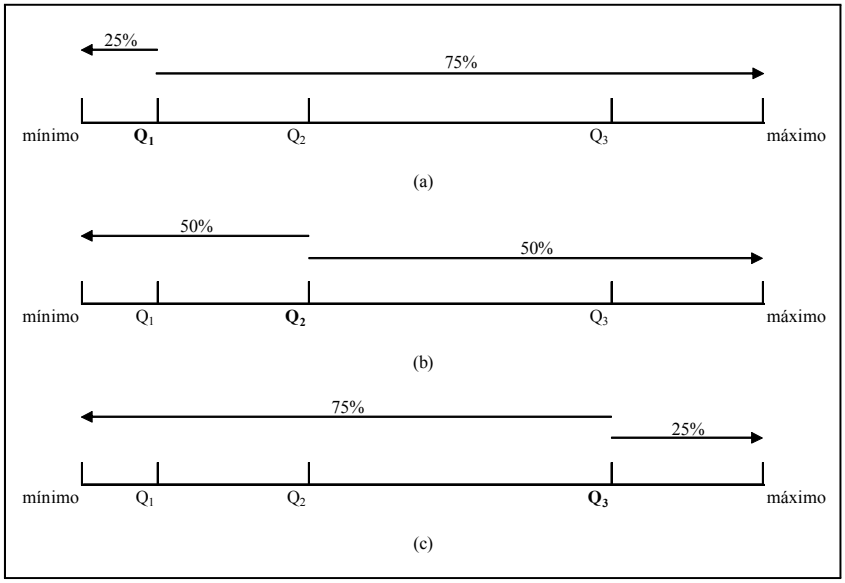


Figura 2.5: Ilustração da definição de quartis

Definição 11 *O **intervalo interquartil**, que denotaremos por IQ, é definido como a distância entre o primeiro e o terceiro quartis, isto é:*

$$IQ = Q_3 - Q_1 \tag{2.18}$$

O intervalo interquartil tem a mesma unidade dos dados. A vantagem do intervalo interquartil sobre o desvio padrão é que, assim como a mediana, o IQ não é muito influenciado por valores discrepantes.

Exercício 2.11 *Calcule todas as medidas de dispersão para as idades dos funcionários do Departamento Financeiro, cujos dados ordenados são*

27	*27*	*28*	*29*	*29*	*29*	*30*	*31*	*32*	*33*	*34*	*35*
36	*38*	*40*	*41*	*42*	*42*	*45*	*48*	*48*	*50*	*52*	

No Exemplo da Seção 2.2.5 foram calculadas algumas medidas de posição para esse conjunto de dados.

2.4 Medidas de posição e dispersão para dados agrupados

Considere a distribuição de freqüências do salário dos funcionários do Departamento de Recursos Humanos reproduzida na Tabela 2.3.

Tabela 2.3: Distribuição da renda dos funcionários do Departamento de RH

Classe	Ponto	Freqüência Simples		Freqüência Acumulada	
de renda	médio	Absoluta	Relativa %	Absoluta	Relativa %
[3200,4021)	3610,5	4	26,67	4	26,67
[4021,4842)	4431,5	2	1,33	6	40,00
[4842,5663)	5252,5	2	1,33	8	53,33
[5663,6484)	6073,5	3	20,00	11	73,33
[6484,7305)	6894,5	4	26,67	15	100,00
Total		15	100,00		

Essa tabela foi construída a partir dos dados da Tabela 1.6 analisada no capítulo anterior. Imagine, agora, que não dispuséssemos daqueles dados e só nos fosse fornecida a Tabela 2.3. Como poderíamos calcular as medidas de posição e dispersão para essa distribuição? Vamos começar com a média aritmética e as medidas de dispersão, pois estes cálculos partem do mesmo princípio.

2.4.1 Média aritmética

Quando agrupamos os dados em uma distribuição de freqüências, estamos perdendo informação, uma vez que não apresentamos os valores individuais. Informar apenas que existem 4 valores na classe $3200 \vdash 4021$ nos obriga a escolher um valor típico, representante de tal classe. Esse valor será sempre o *ponto médio* da classe. Então a informação anterior é interpretada como a existência de 5 valores iguais a 3610,5, que é o ponto médio dessa classe. Essa é a interpretação básica da tabela de freqüências: *todos os valores de uma classe são considerados iguais ao ponto médio da classe.* O ponto médio da classe, por sua vez, é calculado como a média dos limites de classe. Veja a coluna criada com esses valores na Tabela 2.3.

A interpretação da tabela de freqüências nos diz que há 4 observações iguais a 3610,5; 2 observações iguais a 4431,5; 2 iguais a 5252,5; 3 iguais a 6073,5 e 4 iguais a 6894,5. Então esses dados podem ser vistos como o seguinte conjunto de observações:

$$\left.\begin{array}{l} 3610,5 \\ 3610,5 \\ 3610,5 \\ 3610,5 \end{array}\right\} \text{4 ocorrências} \qquad (2.19)$$

$$\left.\begin{array}{l} 4431,5 \\ 4431,5 \end{array}\right\} \text{2 ocorrências}$$

$$\left.\begin{array}{l} 5252,5 \\ 5252,5 \end{array}\right\} \text{2 ocorrências}$$

$$\left.\begin{array}{l} 6073,5 \\ 6073,5 \\ 6073,5 \end{array}\right\} \text{3 ocorrências}$$

$$\left.\begin{array}{c} 6894,5 \\ 6894,5 \\ 6894,5 \\ 6894,5 \end{array}\right\} 4 \text{ ocorrências}$$

Para calcular a média desse novo conjunto de dados temos que fazer:

$$\begin{aligned}
\overline{x} &= \frac{4 \times 3610,5 + 2 \times 4431,5 + 2 \times 5252,5 + 3 \times 6073,5 + 4 \times 6894,5}{15} = \\
&= \frac{4}{15} \times 3610,5 + \frac{2}{15} \times 4431,5 + \frac{2}{15} \times 5252,5 + \frac{3}{15} \times 6073,5 + \frac{4}{15} \times 6894,5 = \\
&= 0,2667 \times 3610,5 + 0,1333 \times 4431,5 + 0,1333 \times 5252,5 + 0,20 \times 6073,5 + 0,2667 \times 6894,5 = \\
&= 5307,2333
\end{aligned}$$

Note, na penúltima linha da equação anterior, que os pontos médios de cada classe estão multiplicados pela freqüência relativa da classe. Então, a média dos dados agrupados em classes é uma *média ponderada dos pontos médios*, onde os pesos são definidos pelas freqüências das classes. Representando o ponto médio da classe por x_i e por f_i a freqüência relativa (não multiplicada por 100), temos que

$$\overline{x} = \sum_{i=1}^{k} f_i x_i \tag{2.20}$$

Os pesos (freqüências) aparecem exatamente para compensar o fato de que as classes têm números diferentes de observações.

2.4.2 Variância

Usando a interpretação da tabela de freqüências que nos diz que há 4 observações iguais a 3610,5, 2 observações iguais a 4431,5, 2 observações iguais a 5252,5, 3 observações iguais a 6073,5 e 4 observações iguais a 6894,5, calculamos a variância desses "novos" dados usando uma das fórmulas 2.12 ou 2.14.

Usando (2.12), a variância é calculada como:

$$\begin{aligned}
\sigma^2 &= \frac{1}{15} \times \left[\begin{array}{c} 4 \times (3610,5 - 5307,2333)^2 + 2 \times (4431,5 - 5307,2333)^2 + 2 \times (5252,5 - 5307,2333)^2 \\ +3 \times (6073,5 - 5307,2333)^2 + 4 \times (6894,5 - 5307,2333)^2 \end{array} \right] \\
&= \frac{4}{15} \times (3610,5 - 5307,2333)^2 + \frac{2}{15} \times (4431,5 - 5307,2333)^2 + \frac{2}{15} \times (5252,5 - 5307,2333)^2 \\
&\quad + \frac{3}{15} \times (6073,5 - 5307,2333)^2 + \frac{4}{15} \times (6894,5 - 5307,2333)^2 \\
&= 1659638,729
\end{aligned}$$

Note, na penúltima linha da equação anterior, que os desvios quadráticos de cada classe estão multiplicados pela freqüência relativa da classe. Dessa forma, chegamos à seguinte expressão para a variância de dados agrupados:

$$\sigma^2 = \sum f_i (x_i - \overline{x})^2 \tag{2.21}$$

onde x_i é o ponto médio da classe e f_i é a freqüência relativa.

Usando a Equação (2.12), a variância é calculada como:

$$\begin{aligned}\sigma^2 &= \frac{1}{15} \times \left[4 \times 3610,5^2 + 2 \times 4431,5^2 + 2 \times 5252,5^2 + 3 \times 6073,5^2 + 4 \times 6894,5^2\right] - 5307,2333^2 \\ &= \left[\frac{4}{15} \times 3610,5^2 + \frac{2}{15} \times 4431,5^2 + \frac{2}{15} \times 5252,5^2 + \frac{3}{15} \times 6073,5^2 + \frac{4}{15} \times 6894,5^2\right] - 5307,2333^2 \\ &= 1659638,729\end{aligned}$$

Note, na penúltima linha da equação anterior, que os quadrados dos pontos médios de cada classe estão multiplicados pela freqüência relativa da classe. Dessa forma, chegamos à seguinte expressão alternativa para a variância de dados agrupados:

$$\sigma^2 = \sum f_i x_i^2 - \overline{x}^2 \tag{2.22}$$

e mais uma vez, obtemos que a variância é a *média dos quadrados menos o quadrado da média*; a diferença é que aqui a média é uma média ponderada pelas freqüências das classes.

2.4.3 Desvio médio absoluto

Seguindo raciocínio análogo, obtemos que o desvio médio absoluto para dados agrupados é

$$DMA = \sum f_i \,|\, x_i - \overline{x} \,|$$

que é uma média ponderada dos desvios absolutos em torno da média.

Exercício 2.12 *Calcule a média, a variância e o desvio médio absoluto para a distribuição dada na seguinte tabela:*

Classes	*Freqüência*
$4 \vdash 6$	*10*
$6 \vdash 8$	*12*
$8 \vdash 10$	*18*
$10 \vdash 12$	*6*
$12 \vdash 14$	*4*
Total	*50*

2.4.4 Mediana

Como já visto, a mediana é o valor que deixa 50% das observações acima e 50% abaixo dela. Estando os dados agrupados em classes, existe um método geométrico que produz uma estimativa da mediana. As idéias subjacentes a esse método são que a mediana divide ao meio o conjunto de dados (ou seja, a definição de mediana) e que, no histograma da distribuição, as áreas dos retângulos são proporcionais às freqüências relativas.

Considere o histograma da Figura 2.6, referente aos salários dos funcionários do Departamento de Recursos Humanos. Nas duas primeiras classes temos 40% das observações e nas três primeiras

classes temos 53,33%; logo, a mediana é algum ponto da *classe mediana* $4842 \vdash 5663$ e abaixo desse ponto temos que ter 50% da distribuição, ou seja, as áreas dos 2 primeiros retângulos mais a área do retângulo hachurado representam 50% da freqüência. Então, para identificar a mediana, devemos notar que na classe mediana ficam faltando $50\% - 40\% = 10\%$ da distribuição para completar 50%. Então a área A_1 do retângulo hachurado deve ser igual a 10%, enquanto que o retângulo da classe mediana tem área $A_m = 13,33\%$. Usando a fórmula que dá a área de um retângulo obtém-se:

$$\begin{aligned} A_1 &= 0,10 = (Q_2 - 4842) \times h \\ A_m &= 0,1333 = (5663 - 4842) \times h \end{aligned}$$

onde h é a altura comum dos dois retângulos. Dividindo as duas igualdades termo a a termo obtém-se a seguinte regra de proporcionalidade:

$$\frac{0,10}{0,1333} = \frac{Q_2 - 4842}{821} \Rightarrow Q_2 = 5457,904$$

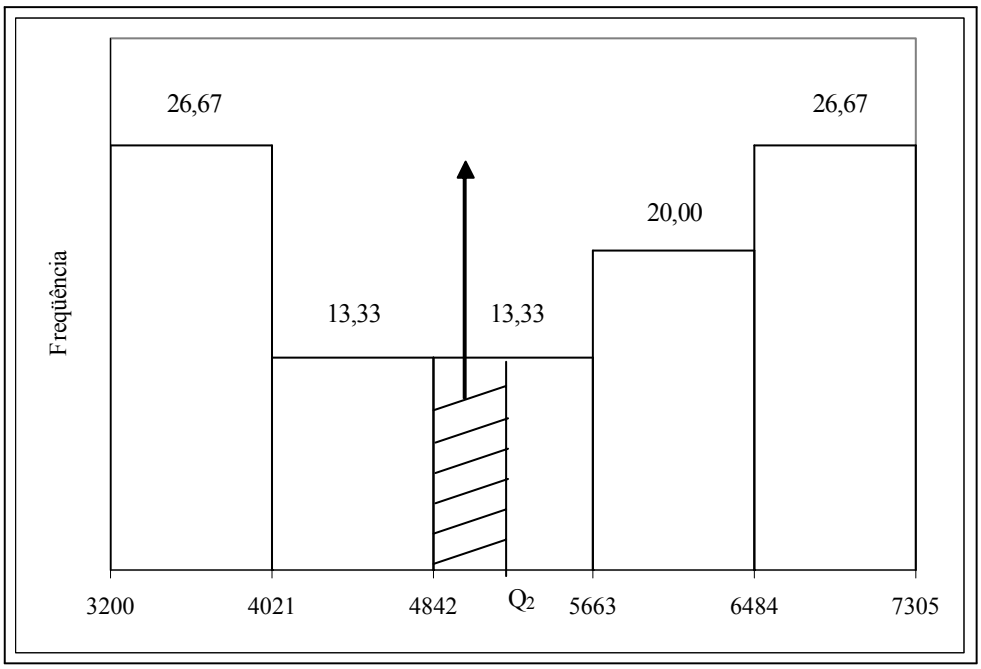


Figura 2.6: Cálculo da mediana dos salários dos funcionários de RH

Exercício 2.13 *Calcule a mediana dos dados da tabela a seguir.*

Classes	*Freqüência*
$1 \vdash 3$	*25*
$3 \vdash 5$	*30*
$5 \vdash 7$	*45*
$7 \vdash 9$	*15*
$\vdash 11$	*10*
Total	*50*

2.4.5 Cálculo de separatrizes de dados agrupados

Como as separatrizes são um caso mais geral da mediana, é possível calcular, de forma análoga, as separatrizes de dados agrupados em classe. Vamos ilustrar o procedimento através de alguns exemplos com base na seguinte distribuição:

Classes	Freqüência simples		Freqüência Acumulada	
	Absoluta	Relativa	Absoluta	Relativa
$4 \vdash 6$	10	0,20	10	0,20
$6 \vdash 8$	12	0,24	22	0,44
$8 \vdash 10$	18	0,36	40	0,80
$10 \vdash 12$	6	0,12	46	0,92
$12 \vdash 14$	4	0,08	50	1,00
Total	50	1,00		

- Cálculo de Q_1

O primeiro quartil está na segunda classe, $6 \vdash 8$. Até a classe anterior (a primeira) temos 20% dos dados; para completar 25%, faltam 5%. Veja a Figura 2.7.

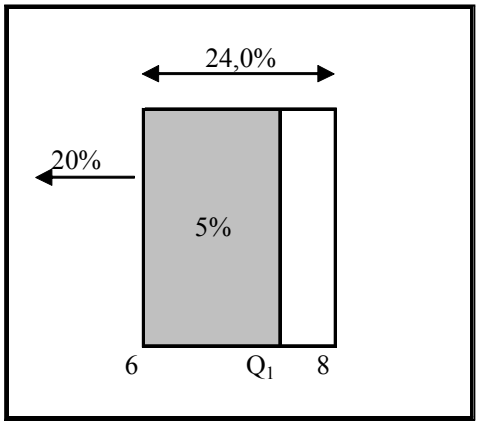


Figura 2.7: Cálculo de Q_1

Temos, então, a seguinte proporcionalidade de áreas:

$$\frac{(Q_1 - 6) \times h}{(8 - 6) \times h} = \frac{5}{24} \Longrightarrow \frac{Q_1 - 6}{8 - 6} = \frac{5}{24} \Rightarrow Q_1 = 6,4167$$

- Cálculo de Q_2

O segundo quartil está na terceira classe, $8 \vdash 10$. Até a classe anterior (a segunda) temos 44% dos dados; para completar 50%, faltam $50\% - 44\% = 6\%$. Veja a Figura 2.8.

Temos, então, a seguinte proporcionalidade de áreas:

$$\frac{(Q_2 - 8) \times h}{(10 - 8) \times h} = \frac{6}{36} \Longrightarrow \frac{Q_2 - 8}{2} = \frac{1}{6} \Rightarrow Q_1 = 8,33$$

- Cálculo de Q_3

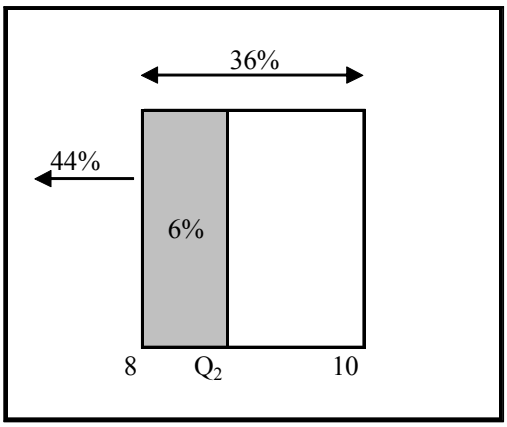


Figura 2.8: Cálculo de Q_2

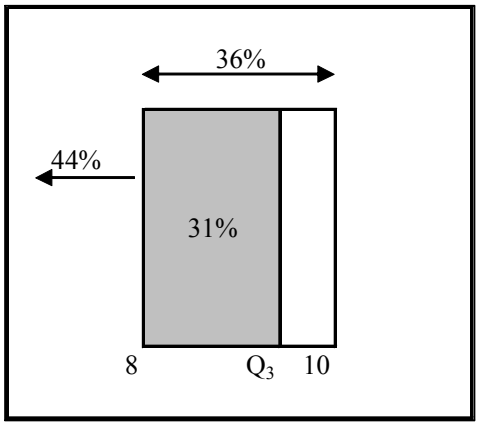


Figura 2.9: Cálculo de Q_3

O terceiro quartil está na terceira classe, $8 \vdash 10$. Até a classe anterior (a segunda) temos 44% dos dados; para completar 75%, faltam $75\% - 44\% = 31\%$. Veja a Figura 2.9.

Temos, então, a seguinte proporcionalidade de áreas:

$$\frac{(Q_3 - 8) \times h}{(10 - 8) \times h} = \frac{31}{36} \Longrightarrow \frac{Q_3 - 8}{2} = \frac{31}{36} \Rightarrow Q_3 = 9,72$$

- Cálculo de P_{80}

Ao final da terceira classe temos exatamente 80% dos dados; logo, $P_{80} = 10$.

Exercício 2.14 *Calcule os três quartis da seguinte distribuição:*

Classes	*Freqüência*
$10 \vdash 20$	*101*
$20 \vdash 30$	*125*
$30 \vdash 40$	*95*
$40 \vdash 50$	*64*
$50 \vdash 60$	*22*
Total	*407*

2.4.6 Moda

Como visto, a moda de um conjunto de dados é o valor mais freqüente. Para dados agrupados em classes, uma definição análoga seria a de *classe modal*, que é a classe de maior freqüência. Por ser o ponto médio o representante da classe, podemos definir a moda dos dados como sendo o ponto médio da classe modal; essa é a definição de *moda bruta*.

Existem, no entanto, alguns métodos que permitem obter uma estimativa mais refinada da moda. Todos esses métodos buscam, na classe modal, um ponto (valor) que seja representativo da moda dos dados. Embora não muito usados na prática, a apresentação de tais métodos tem a vantagem de desenvolver no aluno um raciocínio geométrico e intuitivo sobre essa medida de posição.

Consideremos a seguinte distribuição de freqüencias:

Clases	Ponto	Freq. Simples		Freq. Acumulada	
	médio	Absoluta	Relativa	Absoluta	Relativa
$4 \vdash 6$	5	10	0,20	10	0,20
$6 \vdash 8$	7	12	0,24	22	0,44
$8 \vdash 10$	9	18	0,36	40	0,80
$10 \vdash 12$	11	6	0,12	46	0,92
$12 \vdash 14$	13	4	0,08	50	1,00
Total		50	1,00		

A classe modal é $8 \vdash 10$ e a moda bruta é $x^* = 9$. Veremos dois métodos geométricos para calcular a moda e ambos utilizam apenas a classe modal e suas duas classes vizinhas. Na Figura 2.10 apresentamos parte do histograma da distribuição; veja aí que a moda está em algum ponto da classe modal. Para determinar esse ponto exatamente, temos que determinar a distância $x^* - 8$, o que automaticamente determina $10 - x^*$.

Em ambos os métodos, podemos pensar nas classes vizinhas "puxando" a moda dos dados; a classe que tiver mais "força", ganha, ou seja, a moda estará mais próxima dela. Então, *quanto maior*

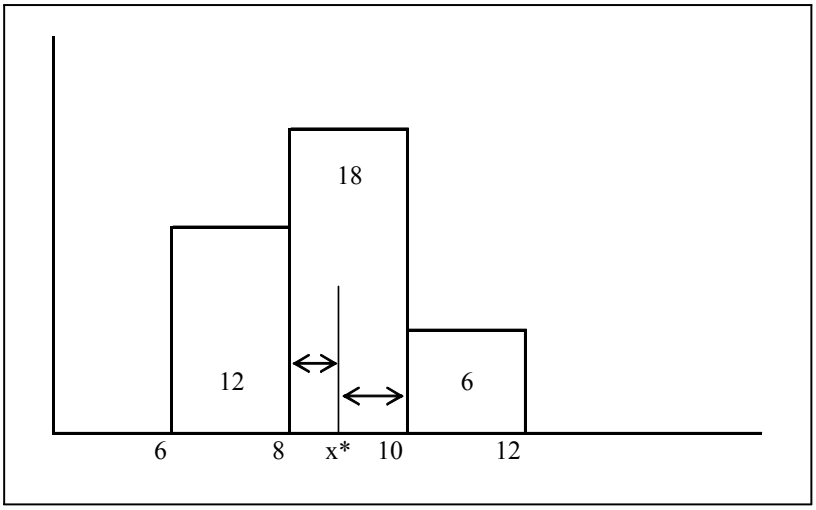


Figura 2.10: Ilustração do método de cálculo da moda de dados agrupados

a força, menor a distância. (Veja a Figura 2.11) Sem fazer qualquer cálculo, podemos deduzir que a moda estará mais próxima da classe inferior, ou seja, a moda tem que ser menor que 9.

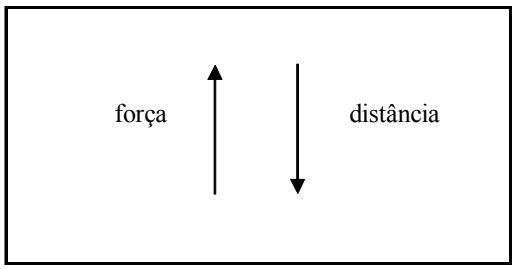


Figura 2.11: Ilustração dos métodos geométricos de cálculo da moda de dados agrupados

Método de King

No método de King, a "força" de cada classe vizinha é a sua freqüência; quanto maior essa freqüência, maior a força. Então, o esquema de proporcionalidade para o método de King é como ilustrado na Figura 2.12:

A distância da moda à classe anterior é $x^* - 8$ e à classe posterior é $10 - x^*$; como essas distâncias têm que ser *inversamente* proporcionais às freqüências das classes (veja a primeira e a última setas) temos que ter:

$$\frac{x^* - 8}{10 - x^*} = \frac{6}{12} \Rightarrow 12x^* - 96 = 60 - 6x^* \Rightarrow 18x^* = 156 \Rightarrow x^* = 8,67$$

Para generalizar esse método, vamos usar a seguinte notação:

- ℓ_I limite inferior da classe modal (ou limite superior da classe anterior);

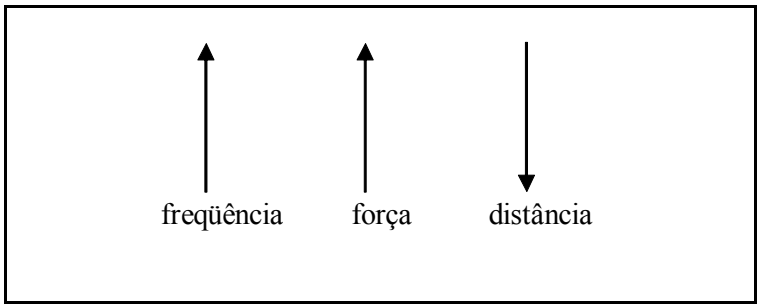


Figura 2.12: Ilustração do método de King

- ℓ_S limite superior da classe modal (ou limite inferior da classe posterior);
- n_I freqüência absoluta da classe inferior ou anterior à classe modal;
- n_S freqüência absoluta da classe superior ou posterior à classe modal;
- δ_m comprimento da classe modal ($\delta_m = \ell_S - \ell_I$).

Então, a moda é determinada pela seguinte regra de proporcionalidade *inversa*:

$$\frac{x^* - \ell_I}{\ell_S - x^*} = \frac{n_S}{n_I} \Rightarrow n_I x^* - n_I \ell_I = n_S \ell_S - n_S x^* \Rightarrow (n_I + n_S)x^* = n_I \ell_I + n_S \ell_S \Rightarrow$$

$$x^* = \frac{n_I}{n_I + n_S} \times \ell_I + \frac{n_S}{n_I + n_S} \times \ell_S \tag{2.23}$$

Note que a moda é uma média ponderada dos extremos da classe modal, ℓ_I e ℓ_S, onde os pesos são definidos pelas freqüências das classes vizinhas.

O método de King resulta de argumentos de semelhança de triângulos. Considere a Figura 2.13, onde temos representadas apenas as classes modal e suas vizinhas, como antes. O ponto A é marcado no lado da moda correspondente ao limite inferior, de modo que sua altura seja igual à freqüência da classe posterior à classe modal. O ponto B é marcado no lado da moda correspondente ao limite superior, mas na parte inferior, de modo que sua altura seja igual à freqüência da classe anterior à classe modal. Os triângulos $A\ell_I x^*$ e $B\ell_S x^*$ são semelhantes. Então, resulta a seguinte proporcionalidade entre os lados:

$$\frac{\overline{\ell_I x^*}}{\overline{\ell_S x^*}} = \frac{\overline{A\ell_I}}{\overline{B\ell_S}}$$

Pela construção desses triângulos, isso significa que:

$$\frac{x^* - \ell_I}{\ell_S - x^*} = \frac{n_S}{n_I}$$

a mesma igualdade obtida anteriormente.

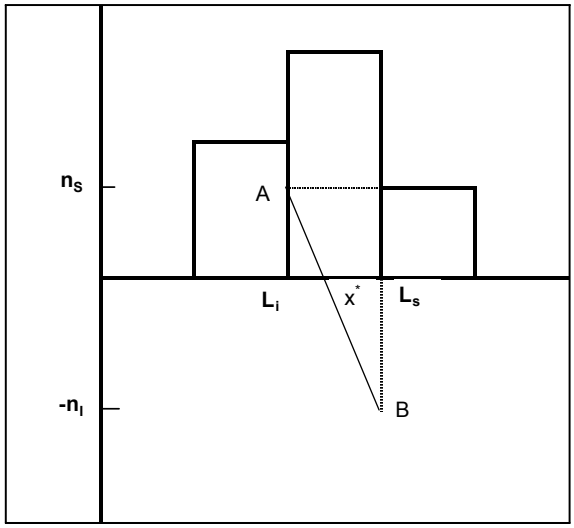


Figura 2.13: Representação geométrica do método de King

Método de Czuber

O método de Czuber é análogo ao de King, mas agora quem define a "força" da classe vizinha é a diferença entre a freqüência da classe modal e a freqüência da classe vizinha. Então, a moda estará mais próxima da classe vizinha que tiver freqüência mais próxima à freqüência modal. Então, quanto menor (maior) a diferença entre as freqüências, maior (menor) será a "força" da classe. Veja a Figura 2.14.

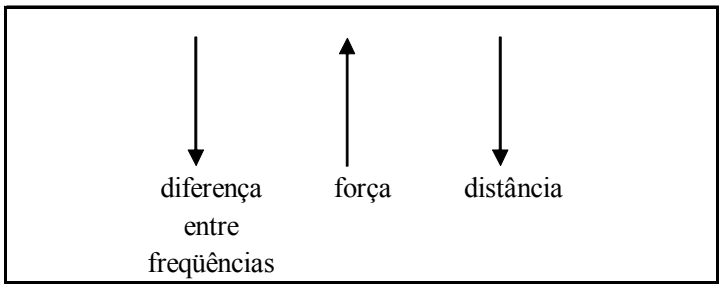


Figura 2.14: Ilustração do método de Czuber

Analisando a Figura 2.14, podemos ver que existe uma proporcionalidade *direta* entre "diferença entre freqüências" e "distância", ou seja, temos s seguinte regra de três:

$$\frac{x^* - 8}{10 - x^*} = \frac{18 - 12}{18 - 6} \Rightarrow 12x^* - 96 = 60 - 6x^* \Rightarrow 18x^* = 156 \Rightarrow x^* = 8,67$$

Nesse exemplo, as modas calculadas pelos dois métodos foram iguais, mas isso nem sempre ocorre, como você verá em outros exercícios.

Para o caso genérico, vamos acrescentar a seguinte notação:

- n_m freqüência da classe modal;

- Δ_I diferença entre a freqüência da classe modal e a freqüência da classe anterior ($\Delta_I = n_m - n_I$);
- Δ_S diferença entre a freqüência da classe modal e a freqüência da classe posterior ($\Delta_S = n_m - n_S$).

Então:

$$\begin{aligned} \frac{x^* - \ell_I}{\Delta_I} &= \frac{\ell_S - x^*}{\Delta_S} \Rightarrow \Delta_S x^* - \Delta_S \ell_I = \Delta_I \ell_S - \Delta_I x^* \Rightarrow \\ (\Delta_I + \Delta_S) x^* &= \Delta_S \ell_I + \Delta_I \ell_S \Rightarrow \\ x^* &= \frac{\Delta_S}{\Delta_I + \Delta_S} \ell_I + \frac{\Delta_I}{\Delta_I + \Delta_S} \ell_S \Rightarrow \end{aligned} \tag{2.24}$$

A interpretação geométrica também se baseia em argumentos de semelhança de triângulos, conforme ilustrado na Figura 2.15. Os triângulos RQS e TQU são semelhantes; resulta a seguinte proporcionalidade entre lados e alturas:

$$\frac{\overline{AQ}}{\overline{RS}} = \frac{\overline{BQ}}{\overline{TU}}$$

ou equivalentemente

$$\frac{x^* - \ell_I}{n_m - n_I} = \frac{\ell_S - x^*}{n_m - n_S} \qquad \Rightarrow \qquad \frac{x^* - \ell_I}{\Delta_I} = \frac{\ell_S - x^*}{\Delta_S}$$

a mesma proporção obtida anteriormente.

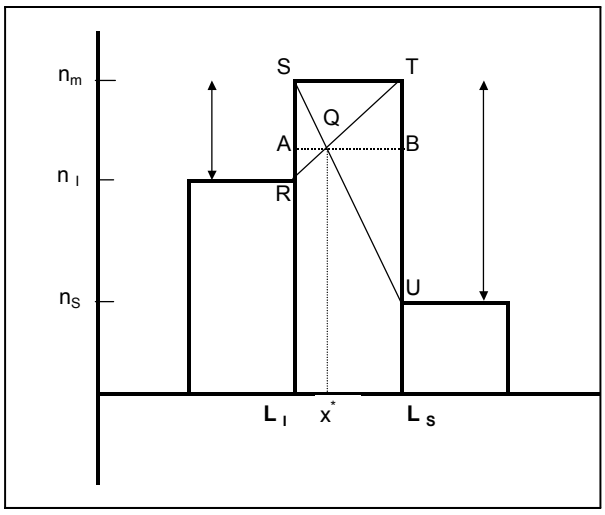


Figura 2.15: Representação geométrica do método de Czuber

Exercício 2.15 *Calcule a média, a mediana e a moda (King e Czuber) dos dados da tabela a seguir.*

Classes	*Freqüência*
$10 \vdash 20$	*101*
$20 \vdash 30$	*125*
$30 \vdash 40$	*95*
$40 \vdash 50$	*64*
$50 \vdash 60$	*22*
Total	*407*

2.4.7 Exemplo

Para fixar as idéias, vamos calcular a média, a variância, a moda (King e Czuber), a mediana, D_1 e D_9 da seguinte distribuição:

Classes	Freqüência Simples		Freqüência Acumulada	
	Absoluta	Relativa %	Absoluta	Relativa %
$0 \vdash 5$	5	6,25	5	6,25
$5 \vdash 10$	15	18,75	20	25,00
$10 \vdash 15$	22	27,50	42	52,50
$15 \vdash 20$	18	22,50	60	75,00
$20 \vdash 25$	12	15,00	72	90,00
$25 \vdash 30$	8	10,00	80	100,00
Total	80	100,00		

Os pontos médios das classes são

$$\frac{0+5}{2} = 2,5 \qquad \frac{5+10}{2} = 7,5 \qquad \cdots \qquad \frac{25+30}{2} = 27,5$$

e a média é calculada como

$$\begin{aligned} \overline{x} &= 0,0625 \times 2,5 + 0,1875 \times 7,5 + 0,2750 \times 12,5 + 0,2250 \times 17,5 + 0,15 \times 22,5 + 0,10 \times 27,5 = \\ &= 15,0625 \end{aligned}$$

Note que é preferível trabalhar com as freqüências relativas em forma decimal pois, se trabalhássemos com as freqüências relativas em forma percentual, teríamos que dividir o resultado por 100!

Para a variância, temos:

$$\begin{aligned} \sigma^2 &= 0,0625 \times 2,5^2 + 0,1875 \times 7,5^2 + 0,2750 \times 12,5^2 + 0,2250 \times 17,5^2 \\ &\quad +0,15 \times 22,5^2 + 0,10 \times 27,5^2 - (15,0625)^2 \\ &= 47,4961 \end{aligned}$$

A classe modal é a classe $10 \vdash 15$, cuja freqüência é 22. As freqüências das classes inferior e superior são, respectivamente, 15 e 18. O cálculo da moda pelos dois métodos é o seguinte:

$$\text{King} \quad : \qquad \frac{x^* - 10}{15 - x^*} = \frac{18}{15} \Rightarrow x^* = 12,73$$

$$\text{Czuber} \quad : \qquad \frac{x^* - 10}{15 - x^*} = \frac{22 - 15}{22 - 18} \Rightarrow x^* = 13,18$$

Da coluna de freqüências relativas acumuladas, vemos que a mediana está na terceira classe $10 \vdash 15$. Nas duas primeiras classes temos 25% dos dados; assim, está faltando 25% para completar 50%. Veja a Figura 2.16. A regra de três resultante é:

$$\frac{Q_2 - 10}{25} = \frac{15 - 10}{27,5} \Rightarrow Q_2 = 14,545$$

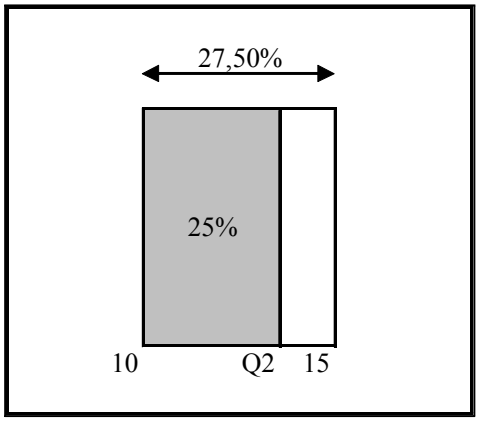


Figura 2.16: Cálculo da mediana para o Exemplo da Seção 2.4.7

De modo análogo, D_1 está na segunda classe e a regra de três é

$$\frac{D_1 - 5}{3,75} = \frac{10 - 5}{18,75} \Rightarrow D_1 = 6,0$$

O cálculo de D_9 é imediato, já que na quarta classe completam-se os 90% da distribuiçõa, ou seja, $D_9 = 25$.

Exercício 2.16 *Calcule a média, a moda (King e Czuber), os quartis e a variância para a seguinte distribuição:*

Classes	*Freqüência Simples*
	Absoluta
$12 \vdash 16$	*10*
$16 \vdash 20$	*18*
$20 \vdash 24$	*29*
$24 \vdash 28$	*10*
$28 \vdash 32$	*3*
Total	*70*

2.5 Medidas de assimetria

Considere os diagramas de pontos dados nas partes (a) a (c) da Figura 2.17, onde a seta indica a média dos dados. Analisando-os, podemos ver que a principal e mais marcante diferença entre eles diz respeito à simetria da distribuição. A segunda distribuição é simétrica, enquanto as outras duas são assimétricas.

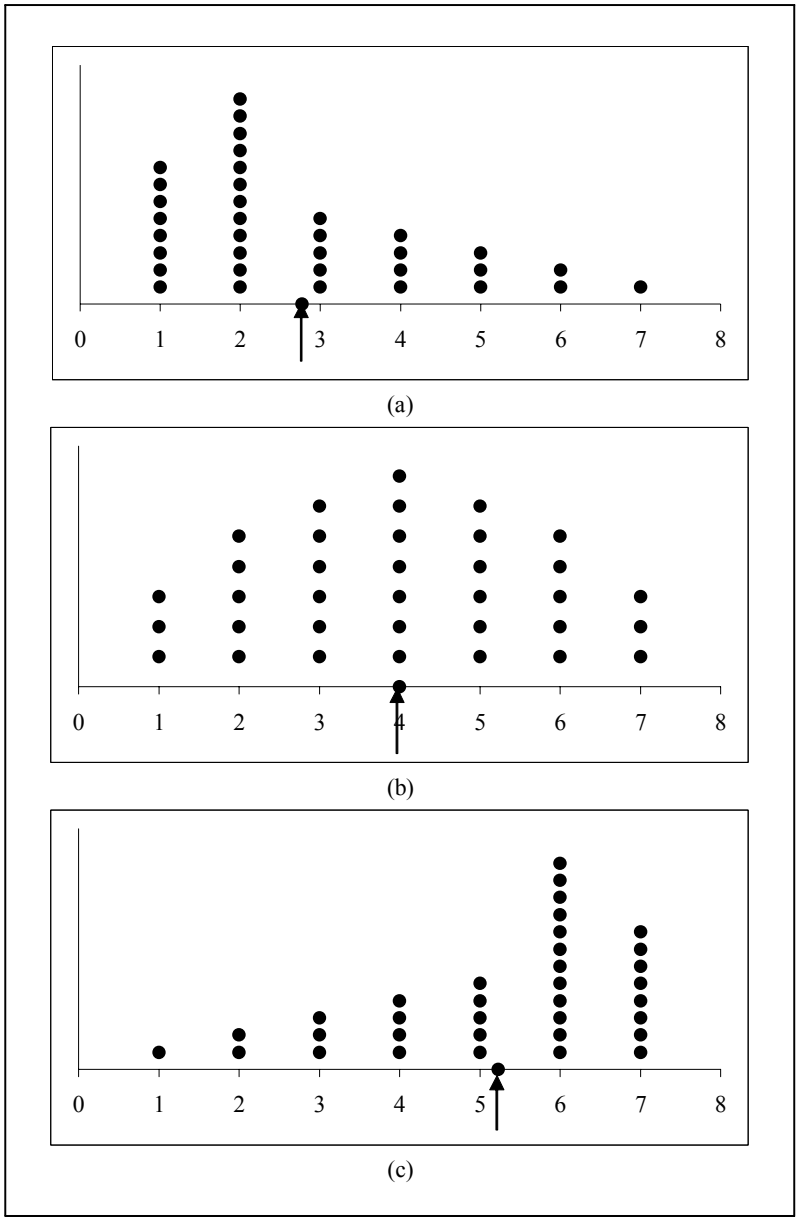


Figura 2.17: Diagramas de pontos de distribuições com diferentes tipos de assimetria

No diagrama (a), a assimetria é tal que há maior concentração na cauda inferior, enquanto no diagrama (c), a concentração é maior na cauda superior. Visto de outra maneira, no diagrama (a) os dados se estendem para o lado positivo da escala, enquanto no diagrama (c), os dados se estendem para o lado negativo da escala. Dizemos que a distribuição ilustrada no diagrama (a) apresenta uma *assimetria à direita*, enquanto a do diagrama (c) apresenta uma *assimetria à esquerda.* No diagrama (b) temos uma *simetria* perfeita ou *assimetria nula*.

2.5.1 Ceoficiente de assimetria de Pearson

Esses três tipos de assimetria podem ser caracterizados pela posição da moda com relação à média dos dados. No primeiro tipo, a moda tende a estar à esquerda da média, enquanto no terceiro tipo, a moda tende a estar à direita da média (lembre-se que a média é o centro de gravidade ou ponto de equilíbrio da distribuição). Para distribuições simétricas, a moda coincide com a média. Definem-se, assim, os três tipos de assimetria:

- se a média é maior que a moda ($\overline{x} > x^*$), dizemos que a distribuição é *assimétrica à direita* ou tem *assimetria positiva* [diagrama (a) da Figura 2.17];
- se a média é igual à moda ($\overline{x} = x^*$), dizemos que a distribuição é *simétrica* ou tem *assimetria nula* [diagrama (b) da Figura 2.17];
- se a média é menor que a moda ($\overline{x} < x^*$), dizemos que a distribuição é *assimétrica à esquerda* ou tem *assimetria negativa* [diagrama (c) da Figura 2.17].

Essas definições, no entanto, não permitem "medir" diferentes graus de assimetria. Por exemplo, considere os diagramas de pontos (a) e (b) dados na Figura 2.18, ambos assimétricos à direita.

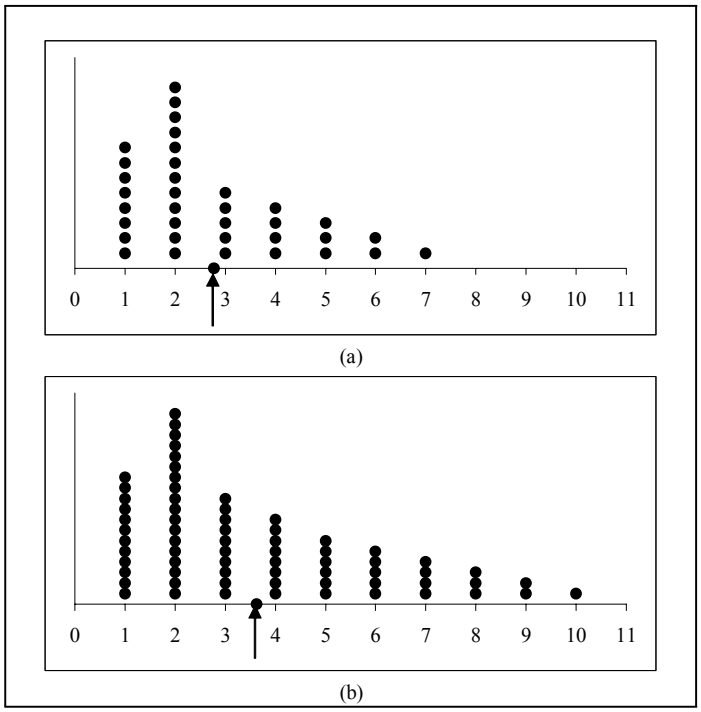


Figura 2.18: Duas distribuições assimétricas à direita

Uma forma de medirmos essas diferentes assimetrias é através da distância $\overline{x} - x^*$ entre a média e a moda, mas como as distribuições podem ter graus de dispersão diferentes, é importante que consideremos a diferença acima na mesma escala. Assim, define-se um dos coeficientes de assimetria (definição devida a Karl Pearson) como:

$$e = \frac{\overline{x} - x^*}{\sigma} \tag{2.25}$$

Se o coeficiente é negativo, temos assimetria negativa; se é positivo, tem-se assimetria positiva e se é nulo, tem-se uma distribuição simétrica. Ao dividirmos pelo desvio padrão, tiramos o efeito de escalas diferentes, o que resulta na *adimensionalidade* do coeficiente.

Para os dados do diagrama (a) da Figura 2.18, temos que $x^* = 2$, $\overline{x} = 2,7714$ e $\sigma = 1,6228$; logo,

$$e = \frac{2,7714 - 2}{1,6228} = 0,475351$$

Para os dados do diagrama (b) da Figura 2.18, $x^* = 2$, $\overline{x} = 3,6232$ e $\sigma = 2,3350$; logo,

$$e = \frac{3,6232 - 2}{2,3350} = 0,6952$$

o que indica uma assimetria mais acentuada.

É interessante observar que existem outros coeficientes de assimetria; o que apresentamos é o menos utilizado, mas é o mais intuitivo.

2.5.2 Coeficiente de assimetria de Bowley

Analisando a Figura 2.5, podemos ver que entre Q_1 e Q_2 e entre Q_2 e Q_3 há sempre 25% dos dados. Então, a diferença entre as distâncias $Q_2 - Q_1$ e $Q_3 - Q_2$ nos dá informação sobre a assimetria da distribuição. Se $Q_2 - Q_1 < Q_3 - Q_2$, isso significa que "andamos mais rápido" para cobrir os 25% inferiores do que os 25% superiores, ou seja, a distribuição "se arrasta" para a direita. Analogamente, se $Q_2 - Q_1 > Q_3 - Q_2$, isso significa que "andamos mais devagar" para cobrir os 25% inferiores do que os 25% superiores, ou seja, a distribuição "se arrasta" para a esquerda. De forma mais precisa, temos o seguinte resultado:

$$Q_2 - Q_1 < Q_3 - Q_2 \Longrightarrow \text{assimetria positiva}$$

$$Q_2 - Q_1 > Q_3 - Q_2 \Longrightarrow \text{assimetria negativa}$$

$$Q_2 - Q_1 = Q_3 - Q_2 \Longrightarrow \text{simetria ou assimetria nula}$$

Para tirar o efeito de escala, temos que dividir por uma medida de dispersão - lembre-se que dividimos pelo desvio padrão quando trabalhamos com as diferenças $\overline{x} - x^*$. Aqui, para não termos efeito dos valores discrepantes, usaremos o intervalo interquartil para gerar a seguinte medida de assimetria, que é chamada *coeficiente de assimetria de Bowley*:

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

que pode ser reescrito como

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

Analisando essa expressão, podemos ver que quanto mais assimétrica à direita for uma distribuição, mais próximos serão Q_1 e Q_2 e, portanto, B se aproxima de $+1$. Analogamente, quanto mais assimétrica à esquerda, mais próximos serão Q_2 e Q_3 e, portanto, B se aproxima de -1.

Exercício 2.17 *Considere novamente as notas de 50 alunos, cujo ramos e folhas é dado a seguir. Calcule o coeficiente de assimetria de Bowley para essa distribuição.*

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

2.6 O *boxplot*

A partir dos quartis constrói-se um gráfico chamado *boxplot* ou *gráfico de caixas*, que ilustra os principais aspectos da distribuição e é também muito útil na comparação de distribuições.

O *boxplot* é formado basicamente por um retângulo vertical (ou horizontal). O comprimento do lado vertical (ou horizontal) é dado pelo intervalo interquartil [Figura 2.19(a), onde estamos trabalhando com um retângulo vertical]. O tamanho do outro lado é indiferente, sugerindo-se apenas uma escala razoável. Na altura da mediana, traça-se uma linha, dividindo o retângulo em duas partes [Figura 2.19(b)].

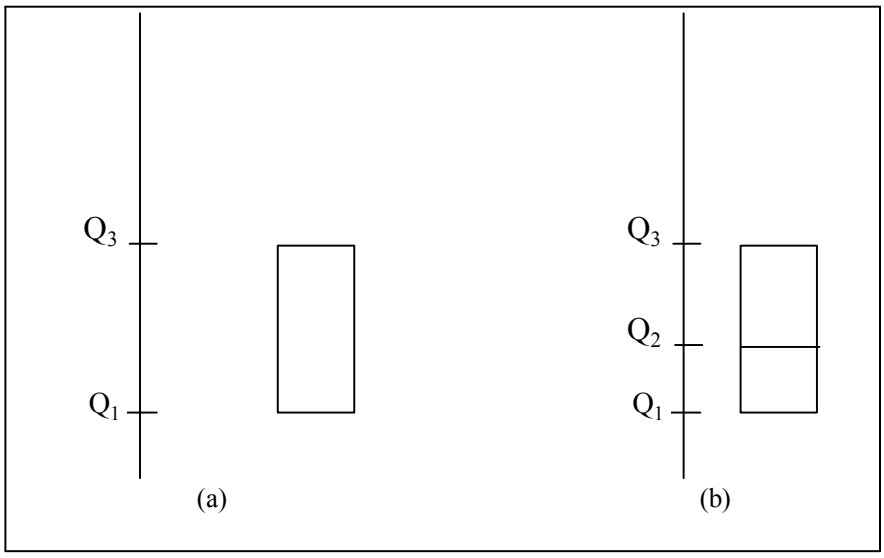


Figura 2.19: Construção do boxplot - Etapa 1

Note que aí já temos representados 50% da distribuição e também já temos idéia da assimetria da mesma - nessa figura temos uma leve assimetria à direita, já que $Q_2 - Q_1 < Q_3 - Q_2$. Para representar os 25% restantes em cada cauda da distribuição temos que cuidar primeiro da presença de possíveis *outliers* ou valores discrepantes, que, como já dito, são valores que se distanciam dos demais.

Regra 4 ***Regra de valores discrepantes*** *para o boxplot*

Um valor x será considerado valor discrepante ou outlier, dentro do seu conjunto, se

$$x < Q_1 - 1,5\,IQ$$

ou

$$x > Q_3 + 1,5\,IQ$$

Veja a Figura 2.20(a). Qualquer valor para fora das linhas pontilhadas é considerado um valor discrepante. Para representar o domínio de variação dos dados na cauda inferior que não são *outliers*, traça-se, a partir do lado do retângulo definido por Q_1, uma linha para baixo até o menor valor que não seja *outlier.* Da mesma forma, na cauda superior, traça-se, a partir do lado do retângulo definido por Q_3, uma linha para cima até o maior valor que não seja *outlier.* [Figura 2.20(b)]. Esses pontos são chamados *juntas.* Dito de outra forma, as juntas são os valores mínimo e máximo do conjunto de dados formado pelos valores não discrepantes. No gráfico são marcadas as juntas; os valores $Q_1 - 1,5\,IQ$ e $Q_3 + 1,5\,IQ$ só servem para definir os outliers, não devendo ser marcados no gráfico.

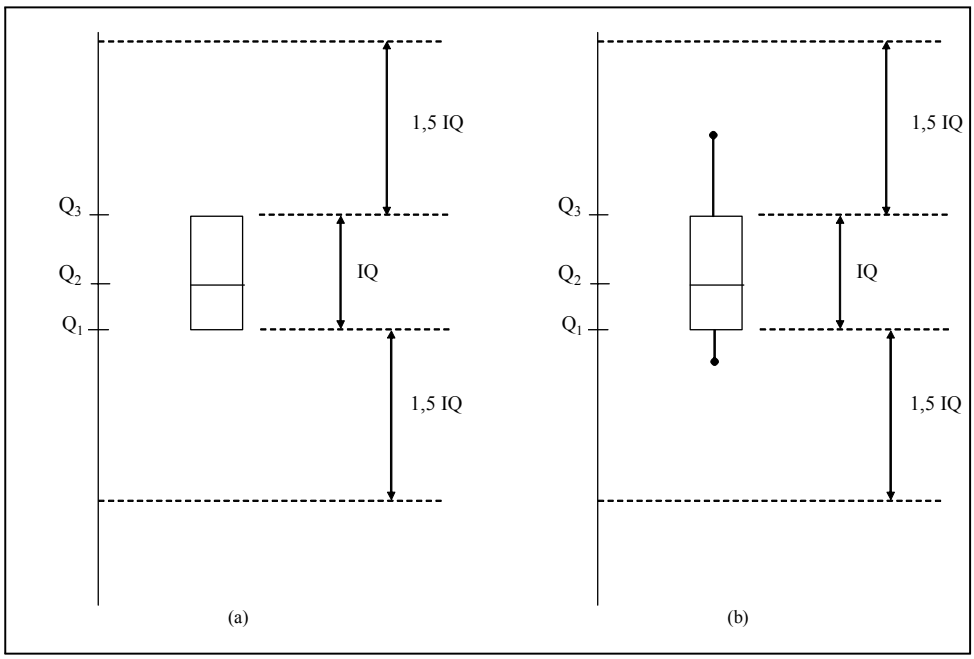


Figura 2.20: Construção do boxplot - Etapa 2

Quanto aos *outliers*, eles são representados individualmente por um X (ou algum outro carácter), explicitando-se, de preferência, os seus valores, mas com uma possível quebra de escala no eixo (Figura 2.21).

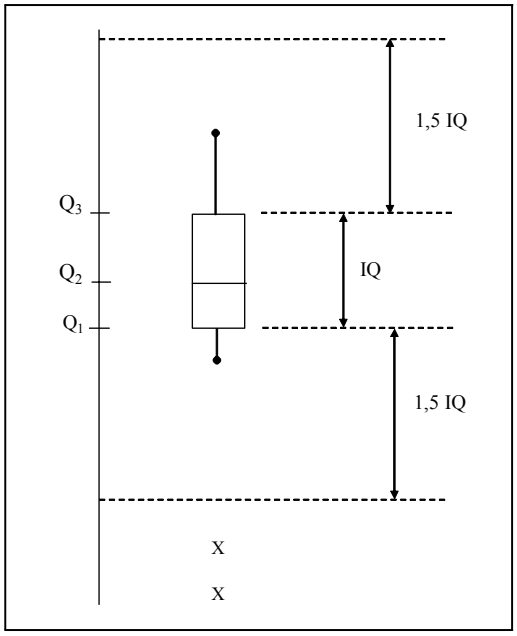


Figura 2.21: Construção do boxplot - Etapa 3

Note que a construção do *boxplot* é toda baseada nos quartis, que são medidas resistentes contra valores discrepantes.

2.6.1 Exemplo

Consideremos novamente as notas de 50 alunos, representadas no gráfico ramos e folhas da Figura 2.22.

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

Figura 2.22: Notas de 50 alunos

A mediana divide o conjunto de dados em duas partes com 25 observações de cada lado (parte sombreada de cinza e a outra). Como o número de observações é par, a mediana é a média dos valores

centrais, que estão circundados por uma borda, um na parte inferior e outro na parte superior.

$$Q_2 = \frac{x_{\left(\frac{50}{2}\right)} + x_{\left(\frac{50}{2}+1\right)}}{2} = \frac{x_{(25)} + +x_{(26)}}{2} = \frac{73+74}{2} = 73,5$$

O primeiro quartil é a mediana da parte inferior, que é o valor circundado por uma borda na parte sombreada de cinza e o terceiro quartil é a mediana da parte superior, que é o valor circundado por uma borda na parte superior, não sombreada.

$$\begin{aligned} Q_1 &= 63 \\ Q_3 &= 82 \\ IQ &= 82 - 63 = 19 \end{aligned}$$

Para estudarmos os *outliers*, temos que calcular

$$\begin{aligned} Q_1 - 1,5IQ &= 63 - 1,5 \times 19 = 34,5 \\ Q_3 + 1,5IQ &= 82 + 1,5 \times 19 = 110,5 \end{aligned}$$

Como a maior nota é 97, não há *outliers* na cauda superior, mas na cauda inferior, temos a nota 29 que é menor que 34,5 e, portanto, um *outlier* inferior. Excluído esse *outlier*, o menor valor que não é discrepante é 37 e o maior valor é 97; logo, as juntas são 37 e 97. Na Figura 2.23 temos o *boxplot* resultante.

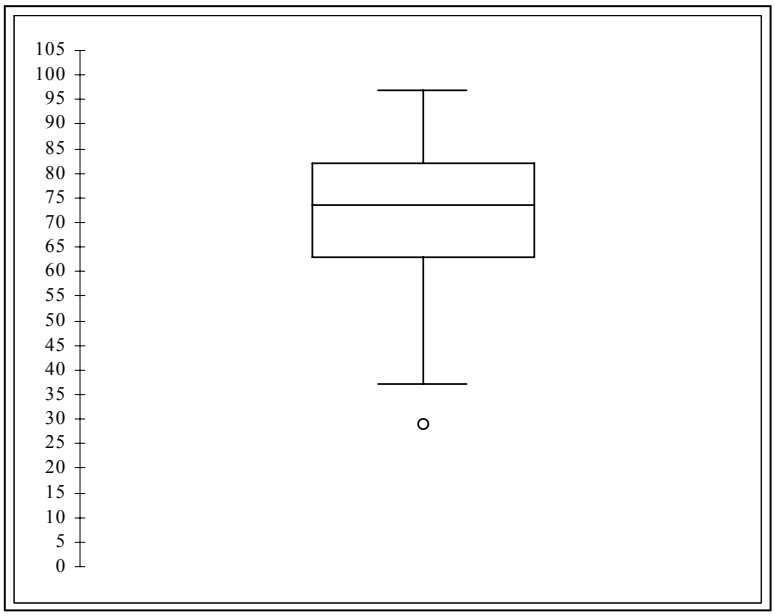


Figura 2.23: Boxplot das 50 notas de alunos

Note que no gráfico final não marcamos os valores 34,5 e 110,5; eles são usados apenas para delimitar os *outliers*. São as juntas que são exibidas no gráfico.

Tabela 2.4: População urbana e rural das UFs brasileiras (em 1000 hab.)

UF	População			UF	População		
	Urbana	Rural	Total		Urbana	Rural	Total
RO	885	496	1381	MG	14672	3220	17892
AC	371	188	559	ES	2464	635	3099
AM	2108	706	2814	RJ	13822	570	14392
RR	248	78	326	SP	34593	2440	37033
PA	4121	2072	6193	PR	7787	1778	9565
AP	425	53	478	SC	4218	1139	5357
TO	860	298	1158	RS	8318	1870	10188
MA	3365	2288	5653	MS	1748	331	2079
PI	1789	1055	2844	MT	1988	517	2505
CE	5316	2116	7432	GO	4397	607	5004
RN	2037	741	2778	DF	1962	90	2052
PB	2448	997	3445				
PE	6059	1861	7920				
AL	1920	903	2823				
SE	1274	512	1786				
BA	8773	4298	13071				

Fonte: IBGE - Censo Demográfico 2000

2.6.2 Exemplo

Considere os dados apresentados na Tabela 2.4, onde temos as populações urbana, rural e total, em 1000 habitantes, dos estados brasileiros. Vamos, inicialmente, construir o *boxplot* para a população total e, em seguida, um *boxplot* comparativo das populações urbana e rural. Na tabela a seguir temos as estatísticas necessárias para a construção desses gráficos.

Estatística	Total	Urbana	Rural
Q_1	2052 (DF)	1748 (MS)	496 (RO)
Q_2	3099 (ES)	2448 (PB)	741 (RN)
Q_3	7920 (PE)	6059 (PE)	1870 (RS)
IQ	5868	4311	1374
$Q_1 - 1,5IQ$	-6750	-4718,5	-1565
$Q_3 + 1,5IQ$	16722	12525,5	3931
Junta inferior	326 (RR)	248 (RR)	53 (AP)
Junta superior	1439 (RJ)	8733 (BA)	3220 (MG)
Outliers	17892 (MG)	13822 (RJ)	4298 (BA)
	37033 (SP)	14672 (MG)	
		34593 (SP)	

Na Figura 2.24 temos o *boxplot* para a população total; vemos aí que as populações de São Paulo e Minas Gerais são *outliers* e a distribuição apresenta uma forte assimetria à direita, ou seja, muitos estados têm população pequena enquanto alguns poucos têm população bem grande.

Na Figura 2.25 temos um *boxplot* comparativo das populações urbana e rural. Podemos ver que a população urbana apresenta maior variabilidade e também uma forte assimetria positiva. Há 3

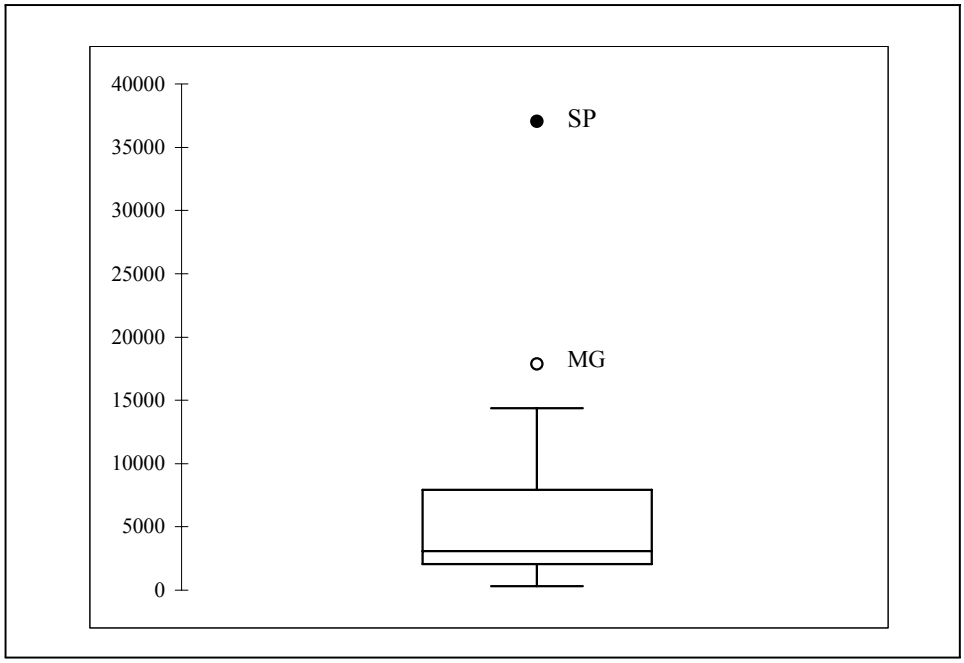


Figura 2.24: População total (em 1000 hab) das Unidades da Federação brasileiras

UFs que são discrepantes: São Paulo, Minas Gerais e Rio de Janeiro. Em termos da população rural, a Bahia é o único *outlier* e a distribuição também é assimétrica à direita.

Exercício 2.18 *Construa o boxplot para os salários dos funcionários do Departamento de Recursos Humanos, cujos valores em reais são 6300, 5700, 4500, 3800, 3200, 7300, 7100, 5600, 6400, 7000, 3700, 6500, 4000, 5100, 4500.*

2.7 Exercícios Complementares

Exercício 2.19 *Quatro amigos trabalham em um supermercado em tempo parcial com os seguintes salários horários:*

Pedro:	*R$ 3,50*	*João:*	*R$ 2,60*
Marcos:	*R$ 3,80*	*Luiz:*	*R$ 2,20*

Se Pedro trabalha 10 horas por semana, João 12 horas, Marcos 15 horas e Luiz 8 horas, qual é o salário horário médio desses quatro amigos?

Exercício 2.20 *Na UFF, o coeficiente de rendimento (CR) semestral dos alunos é calculado como uma média das notas finais nas disciplinas cursadas, levando em conta a carga horária (ou crédito) das disciplinas, de modo que disciplinas com maior carga horária têm maior peso no CR. Suponha que um aluno tenha cursado 5 disciplinas em um semestre, obtendo médias finais de 7,5; 6,1; 8,3; 6,5; 7,5. As três primeiras disciplinas tinham carga horária de 4 horas semanais, a quarta, carga horária de 6 horas e a última, 2 horas semanais. Calcule o CR do aluno nesse semestre.*

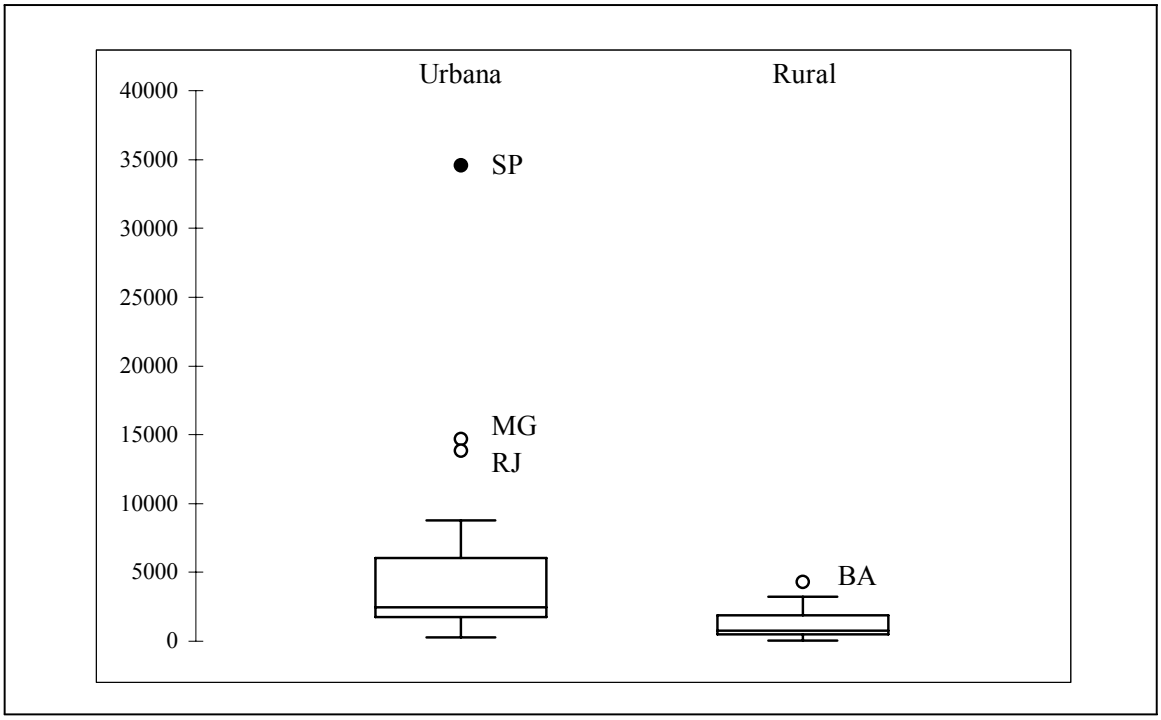


Figura 2.25: População urbana e rural das UFs brasileiras (em 1000 hab)

Exercício 2.21 *Em uma pesquisa sobre atividades de lazer realizada com uma amostra de 20 alunos de um campus universitário, perguntou-se o número de horas que os alunos gastaram "navegando" na Internet na semana anterior. Os resultados obtidos foram os seguintes:*

15	*24*	*18*	*8*	*10*	*12*	*15*	*14*	*12*	*10*
18	*12*	*6*	*20*	*18*	*16*	*10*	*12*	*15*	*9*

1. *Calcule a média, a moda e a mediana desses dados, especificando as respectivas unidades.*

2. *Calcule a amplitude, o desvio médio absoluto e o desvio padrão desses dados, especificando as respectivas unidades.*

Exercício 2.22 *No final do ano 2005, o dono de um pequeno escritório de administração deu a seus 8 funcionários uma gratificação de 250 reais, paga junto com o salário de dezembro.*

1. *Se em novembro o salário médio desses funcionários era de 920 reais, qual o salário médio em dezembro? Que propriedades você utilizou para chegar a esse resultado?*

2. *Se em novembro o desvio padrão dos salários desses funcionários era de 180 reais, qual o desvio padrão dos salários em dezembro? Que propriedades você utilizou para chegar a esse resultado?*

Exercício 2.23 *No mês de dissídio de determinada categoria trabalhista, os funcionários de uma empresa tiveram reajuste salarial de 8,9%.*

1. *Se no mês anterior ao dissídio o salário médio desses funcionários era de 580 reais, qual o valor do salário médio depois do reajuste? Que propriedades você utilizou para chegar a esse resultado?*

2. *Se no mês anterior ao dissídio o desvio padrão dos salários desses funcionários era de 220 reais, qual o valor do desvio padrão dos salários depois do reajuste? Que propriedades você utilizou para chegar a esse resultado?*

Exercício 2.24 *O número médio de empregados das empresas industriais do setor de fabricação de bebidas em determinado momento era de 117 empregados, enquanto o número mediano era de 27. Dê uma explicação para a diferença entre essas medidas de tendência central.*

Exercício 2.25 *Na tabela a seguir temos o número de empresas por faixa de pessoal ocupado (PO) do setor de fabricação de bebidas em determinado momento.*

1. *Calcule a média, a mediana e a moda (King e Czuber) dessa distribuição, especificando as respectivas unidades.*

Classe de PO	*Número de empresas*
$[10, 30)$	*489*
$[30, 100)$	*269*
$[100, 500)$	*117*
$[500, 1000)$	*15*
$[1000, 2000)$	*9*
$[2000, 4000)$	*7*

2. *Calcule o desvio médio absoluto e o desvio padrão dessa distribuição, especificando as respectivas unidades.*

3. *Calcule Q_1, Q_3 e P_{90}.*

Exercício 2.26 *Os dados a seguir representam o número de apólices de seguro que um corretor conseguiu vender em cada um de seus 20 primeiros dias em um emprego novo: 2, 4, 6, 3, 2, 1, 4, 3, 5, 2, 1, 1, 4, 0, 2, 2, 5, 2, 2, 1. Analise a assimetria da distribuição, utilizando os coeficientes de Pearson e de Bowley.*

Exercício 2.27 *Em sua política de fidelização de clientes, determinado supermercado tem uma promoção de dar descontos especiais diferenciados no mês do aniversário do cliente. O desconto básico é de 5%, mas clientes especiais - aqueles com pontuação alta - podem receber prêmios adicionais, que variam a cada mês e de filial para filial. A seguir você tem os pontos dos clientes aniversariantes de determinado mês em uma das filiais do supermercado.*

77	*69*	*72*	*73*	*71*	*75*	*75*	*74*	*71*	*72*	*74*	*73*	*75*	*71*	*74*
73	*78*	*77*	*74*	*75*	*69*	*76*	*76*	*80*	*74*	*85*	*74*	*73*	*72*	*74*

1. *Construa o gráfico ramo-e-folhas e comente suas principais características.*

2. *Calcule a mediana e o intervalo interquartil IQ.*

3. *Construa o boxplot e comente suas principais características.*

4. *Essa filial dá uma garrafa de champagne para seus clientes especiais, segundo a seguinte regra: a cada mês, os clientes com pontuação acima do terceiro quartil por 1,5 vezes o intervalo interquartil serão premiados. Algum cliente ganhará a garrafa de champagne nesse mês?*

Exercício 2.28 *Em uma granja foi observada a distribuição dos frangos com relação ao peso apresentada na Tabela 2.5.*

Tabela 2.5: Peso de frangos para o exercício 2.28

Peso (gramas)	n_i
$960 \vdash 980$	60
$980 \vdash 1000$	160
$1000 \vdash 1020$	280
$1020 \vdash 1040$	260
$1040 \vdash 1060$	160
$1060 \vdash 1080$	80

1. *Qual é a média da distribuição?Qual é a variância da distribuição?*

2. *Queremos dividir os frangos em 4 categorias, com relação ao peso, de modo que*

   - *os 20% mais leves sejam da categoria D;*
   - *os 30% seguintes sejam da categoria C;*
   - *os 30% seguintes sejam da categoria B;*
   - *os 20%restantes sejam da categoria A.*

   Quais os limites de peso entre as categorias A, B, C, D?

3. *O granjeiro decide separar deste lote os animais com peso inferior a dois desvios padrões abaixo da média para receberem ração reforçada e também separar os animais com peso superior a um e meio desvio padrão acima da média para usá-los como reprodutores. Qual a porcentagem de animais que serão separados em cada caso?*

.

PÁGINA EM BRANCO

Capítulo 3

Outras Medidas Estatísticas

3.1 Média geométrica

Definição 12 *A média geométrica de n valores positivos $x_1, x_2, \ldots, x_n$ é definida como*

$$\overline{x}_g = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \prod_{i=1}^{n} (x_i)^{1/n} \tag{3.1}$$

Em Demografia, a média geométrica pode ser usada para se estimar a população de uma determinada localidade num ano t_x, desde que se suponha que a taxa de crescimento entre os 2 censos seja constante. Sejam P_0 a população no 1º censo, realizado na data t_0, P_N a população do 2º censo realizado na data t_N e P_x a população que se quer estimar na data t_x ($t_0 < t_x < t_N$). O crescimento da população entre os dois censos é igual a $\frac{P_N}{P_0}$; se a taxa de crescimento é constante igual a r, isso significa que ao fim do primeiro período a população é igual a

$$P_1 = P_0 + P_0 \times r = P_0 \times (1 + r)$$

Ao final do segundo período,

$$P_2 = P_1 + P_1 \times r = P_1 \times (1 + r) = P_0 \times (1 + r) \times (1 + r) = P_0 \times (1 + r)^2$$

Ao final do último período,

$$P_N = P_0 \times (1 + r)^N$$

Logo,

$$\frac{P_N}{P_0} = (1 + r)^N \qquad \Rightarrow \qquad r = \sqrt[N]{\frac{P_N}{P_0}} - 1$$

A população em qualquer período x entre os censos, então, é dada por

$$P_x = P_0 \times (1 + r)^x = P_0 \times \left(\sqrt[N]{\frac{P_N}{P_0}} \right)^x$$

Lembrando que

$$\sqrt[n]{x} = x^{\frac{1}{n}}$$

então podemos escrever

$$P_x = P_0 \times \frac{P_N^{\frac{x}{N}}}{P_0^{\frac{x}{N}}} = (P_0)^{1-\frac{x}{N}} \times (P_N)^{\frac{x}{N}} = \sqrt[N]{(P_0)^{N-x}\,(P_N)^x}$$

Vê-se, então, que P_x é uma média geométrica de $N-x$ valores iguais a P_0 e de x valores iguais a P_N. Em particular, se o instante de tempo x é o período central, isto é, $x = \frac{N}{2}$, então

$$P_x = \sqrt[N]{(P_o)^{\frac{N}{2}}\,(P_N)^{\frac{N}{2}}} = \left[(P_o)^{\frac{N}{2}}\,(P_N)^{\frac{N}{2}}\right]^N = \sqrt{P_0 \times P_N}$$

a média geométrica de P_0 e P_N. De acordo com os Censos Demográficos realizados pelo IBGE, a população (recenseada) do estado do Rio de Janeiro em 1/9/1980 era de 11.489.797 habitantes e em 1/9/1991 de 12.783.761. Admitindo um crescimento geométrico constante, uma estimativa para a população desse estado em 1985 pode ser calculada como

$$\begin{aligned} P_{85} &= \sqrt[11]{(11.489.797)^6\,(12.783.761)^5} = 11.489.797 \times \left(\sqrt[11]{\frac{12.783.761}{11.489.797}}\right)^5 = \\ &= 11.489.797 \times (1,009748691)^5 = 12.060.876 \end{aligned}$$

3.1.1 Exemplo - Matemática Financeira

Vamos fazer uma comparação entre as médias aritmética e geométrica através de um exemplo de matemática financeira elementar. Para isso, considere as seguintes taxas de juros mensais: $i_1 = 2,5\% = 0,025$; $i_2 = 3,8\% = 0,038$; $i_3 = 4,5\% = 0,045$; $i_4 = 4,9\% = 0,049$; $i_5 = 6,2\% = 0,062$ e $i_6 = 7,8\% = 0,078$; suponha também que uma pessoa tenha um capital inicial de $C_0 = 150$ u.m. (unidades monetárias).

No regime de capitalização simples (juros simples), apenas o capital inicial rende juros. Já no regime de capitalização composta (juros compostos), os rendimentos incorporados ao capital inicial, em cada período, também rendem juros no período seguinte. Vamos analisar os resultados da aplicação acima sob os dois regimes de capitalização.

Capitalização Simples:

Como os juros só incidem sobre o capital inicial, em cada mês o valor dos juros J_t (em u.m.) é calculado como

$$J_t = C_0 \times i_t$$

em que i_t está na forma decimal; ao final do período o montante é

$$C_t = C_{t-1} + J_t$$

Então, para o primeiro mês temos

$$J_1 = C_0 \times i_1 = 150 \times 0,025 = 3,75 \qquad C_1 = C_0 + J_1 = 150 + 3,75 = 153,75$$

Para o segundo mês,

$$J_2 = C_0 \times i_2 = 150 \times 0,038 = 5,7 \qquad C_2 = C_1 + J_2 = 153,75 + 5,7 = 159,45$$

Continuando com esses cálculos, obtemos para o último mês, conforme ilustrado na Tabela 3.1,

$$J_6 = C_0 \times i_6 = 150 \times 0,078 = 11,7 \qquad C_6 = C_5 + J_6 = 182,85 + 11,7 = 194,55$$

Tabela 3.1: Cálculo dos juros no regime de capitalização simples

Mês	Taxa de juros (%)	Valor dos juros (u.m.)	Montante (u.m.)
1	2,5	3,750	153,750
2	3,8	5,700	159,450
3	4,5	6,750	166,200
4	4,9	7,350	173,550
5	6,2	9,300	182,850
6	7,8	11,700	194,550
Média aritmética	**4,95**	**7,425**	

Note que o capital final C_6 é obtido como

$$\begin{aligned} C_6 &= C_5 + J_6 = J_6 + C_4 + J_5 = J_6 + J_5 + C_3 + J_4 = J_6 + J_5 + J_4 + C_2 + J_3 = \\ &= J_6 + J_5 + J_4 + J_3 + C_1 + J_2 = J_6 + J_5 + J_4 + J_3 + J_2 + C_0 + J_1 = \\ &= J_6 + J_5 + J_4 + J_3 + J_2 + J_1 + C_0 = \\ &= C_0 \times i_6 + C_0 \times i_5 + C_0 \times i_4 + C_0 \times i_3 + C_0 \times i_2 + C_0 \times i_1 + C_0 = \\ &= (i_6 + i_5 + i_4 + i_3 + i_2 + i_1) \times C_0 + C_0 \qquad (3.2) \\ &= C_0 + C_0 \times \sum_{k=1}^{6} i_k \end{aligned}$$

Para obtermos o mesmo montante ao final do sexo mês, mas a uma taxa de juros constante, temos que ter

$$i_6 + i_5 + i_4 + i_3 + i_2 + i_1 = i + i + i + i + i + i$$

ou

$$i = \frac{i_1 + i_2 + i_3 + i_4 + i_5 + i_6}{6} = \frac{\sum_{k=1}^{6} i_k}{6}$$

ou seja, a taxa de juros comum tem que ser a média aritmética das taxas mensais. No nosso exemplo,

$$i = \frac{0,025 + 0,038 + 0,045 + 0,049 + 0,062 + 0,078}{6} = 0,0495 \text{ ou } 4,95\%$$

Na Tabela 3.2 temos o resultado de uma aplicação a uma taxa constante de 4,95%. Note que o montante final é o mesmo obtido anteriormente.

Tabela 3.2: Regime de capitailização simples - taxa de juros constante

Taxa de juros (%)	Valor dos juros (u.m.)	Montante (u.m.)
4,950	7,425	157,425
4,950	7,425	164,850
4,950	7,425	172,275
4,950	7,425	179,700
4,950	7,425	187,125
4,950	7,425	194,550

Capitalização Composta:
Como os juros incidem também sobre rendimentos, o valor dos juros para cada mês é dado por

$$J_t = C_{t-1} \times i_t$$

e o montante é

$$C_t = C_{t-1} + J_t$$

Então, para o primeiro mês temos

$$J_1 = i_1 \times C_0 = 0,025 \times 150 = 3,75 \qquad C_1 = C_0 + J_1 = 150 + 3,75 = 153,75$$

Para o segundo mês,

$$J_2 = i_2 \times C_1 = 0,038 \times 153,75 = 5,8425 \qquad C_2 = C_1 + J_2 = 153,75 + 5,8425 = 159,5925$$

Conforme ilustrado na Tabela 3.3, os valores para o sexto mês são:

$$\begin{aligned} J_6 &= i_6 \times C_5 = 0,078 \times 185,792754443175 = 14,491834846568 \\ C_6 &= C_5 + J_6 = 185,792754443175 + 14,491834846568 = 200,284589289743 \end{aligned}$$

Tabela 3.3: Cálculo dos juros no regime de capitalização composta

Mês	Taxa de juros (%)	Taxa de juros índice	Valor dos juros (u.m.)	Montante (u.m.)
1	2,5	1,025	3,750000000	153,75000000
2	3,8	1,038	5,842500000	159,59250000
3	4,5	1,045	7,181662500	166,77416250
4	4,9	1,049	8,171933960	174,94609646
5	6,2	1,062	10,846657981	185,79275444
6	7,8	1,078	14,49183485	200,28458929
Média geométrica		**1,04936372**		

Note que o montante final é calculado como

$$\begin{aligned} & C_6 = i_6 \times C_5 + C_5 = (1+i_6) \times C_5 = (1+i_6) \times (C_4 + C_4 \times i_5) = \\ = \; & (1+i_6) \times (1+i_5) \times C_4 = (1+i_6) \times (1+i_5) \times (C_3 + C_3 \times i_4) = \\ = \; & (1+i_6) \times (1+i_5) \times (1+i_4) \times C_3 = \\ = \; & (1+i_6) \times (1+i_5) \times (1+i_4) \times (C_2 + C_2 \times i_3) = \\ = \; & (1+i_6) \times (1+i_5) \times (1+i_4) \times (1+i_3) \times C_2 = \\ = \; & (1+i_6) \times (1+i_5) \times (1+i_4) \times (1+i_3) \times (C_1 + C_1 \times i_2) = \end{aligned}$$

$$\begin{aligned} = \; & (1+i_6) \times (1+i_5) \times (1+i_4) \times (1+i_3) \times C_1 = \\ = \; & (1+i_6) \times (1+i_5) \times (1+i_4) \times (1+i_3) \times (C_0 + C_0 \times i_1) = \end{aligned}$$

$$= (1+i_6) \times (1+i_5) \times \cdots \times (1+i_1) \times C_0 = C_0 \times \prod_{k=1}^{6} (1+i_k) \tag{3.3}$$

Para obtermos o mesmo capital final, mas a uma taxa constante i, temos que ter

$$\begin{aligned} & (1+i_6)\times(1+i_5)\times(1+i_4)\times(1+i_3)\times(1+i_2)\times(1+i_1) \\ = \; & (1+i)\times(1+i)\times(1+i)\times(1+i)\times(1+i)\times(1+i) \end{aligned}$$

ou seja,

$$(1+i)^6 = (1+i_6)\times(1+i_5)\times(1+i_4)\times(1+i_3)\times(1+i_2)\times(1+i_1)$$

ou ainda

$$(1+i) = \sqrt[6]{(1+i_6)\times(1+i_5)\times(1+i_4)\times(1+i_3)\times(1+i_2)\times(1+i_1)}$$

Então, a taxa comum é calculada como uma média geométrica, não das taxas mensais, mas dos valores $1+i_t$. O "1" aparece exatamente por que os juros incidem sobre o capital do mês anterior. Logo, a taxa comum, em forma percentual, é

$$i = \left(\sqrt[6]{(1+i_6)\times(1+i_5)\times(1+i_4)\times(1+i_3)\times(1+i_2)\times(1+i_1)} - 1\right)\times 100$$

No nosso exemplo, essa taxa comum é

$$\begin{aligned} i &= 100\times\left(\sqrt[6]{1,025\times1,038\times1,045\times1,049\times1,062\times1,078} - 1\right) = \\ &= 100\times\left(\sqrt[6]{1,335230595265} - 1\right) = 4,93637217932303 \end{aligned}$$

Uma aplicação a essa taxa constante, sob o regime de capitalização composta, resulta no mesmo montante final, conforme ilustrado na Tabela 3.4.

Tabela 3.4: Regime de capitalização composta - taxa de juros constante

Taxa de juros (%)	Valor dos juros (u.m.)	Montante (u.m.)
4,93637218	7,40455827	157,40455827
4,93637218	7,77007482	165,17463309
4,93637218	8,15363463	173,32826773
4,93637218	8,55612839	181,88439611
4,93637218	8,97849073	190,86288684
4,93637218	9,42170245	200,28458929

Aplicação 1

Um capital inicial de 1200 u.m. foi aplicado em um regime de capitalização composta, rendendo ao final de um trimestre (3 meses) juros de 126,52. Qual foi a taxa média mensal?

Solução:

Note que da equação (3.3) obtemos

$$\frac{C_t}{C_0} = (1+i_1)\times\cdots\times(1+i_t)$$

Em termos da taxa média comum,

$$\frac{C_t}{C_0} = (1+i)^t \Rightarrow \sqrt[t]{\frac{C_t}{C_0}} = (1+i) \Rightarrow i = \left(\sqrt[t]{\frac{C_t}{C_0}} - 1\right)$$

No exercício, o capital final é 1326,52 e, portanto, a variação nos três meses é

$$\frac{C_3}{C_0} = \frac{1326,52}{1200,00} = 1,105433333$$

Logo, a taxa média mensal é

$$i = \left(\sqrt[3]{1,105433333} - 1\right) = 0,033976937 \text{ ou } 3,398\%$$

Aplicação 2

Resolva o exercício anterior para um regime de capitalização simples.

Solução:

Da equação (3.2), obtemos que

$$\frac{C_t}{C_0} = 1 + \frac{i_1 + \cdots + i_t}{100}$$

Em termos da média comum,

$$\frac{C_t}{C_0} - 1 = \frac{i + \cdots + i}{100}$$

ou seja,

$$i = 100 \times \frac{\left(\frac{C_t}{C_0} - 1\right)}{t} = 100 \times \frac{\frac{C_t - C_0}{C_0}}{t}$$

Note que $\frac{C_t - C_0}{C_0}$ é a variação relativa; dividindo pelo número de períodos, obtemos a variação média. No nosso exercício,

$$i = 100 \times \frac{\left(\frac{1326,52}{1200,00} - 1\right)}{3} = 100 \times \frac{0,105433333}{3} = 3,5144444$$

3.2 Média harmônica

Considere o seguinte exemplo: uma pessoa viaja num fim de semana do Rio de Janeiro para São Paulo, dirigindo seu próprio carro. Na ida, ela desenvolve uma velocidade média de 70km/h mas, na volta, por estar o tráfego na via Dutra mais tranqüilo, ela desenvolve uma velocidade média de 90km/h. Qual a velocidade média para a viagem completa? Para responder esta pergunta, temos que lembrar que a velocidade média é dada pela razão entre a distância percorrida e o tempo gasto para percorrê-la. Para simplificar, suponhamos que a distância entre as duas cidades seja de 450 km. Então, a distância total percorrida é de $2 \times 450 = 900$km. Por outro lado, o tempo gasto na ida foi de $\frac{450}{70}$h e na volta, $\frac{450}{90}$h. Logo, a velocidade média para a viagem completa é de

$$\overline{x}_h = \frac{2 \times 450}{\frac{450}{70} + \frac{450}{90}} = \frac{2}{\frac{1}{70} + \frac{1}{90}} = \frac{1}{\frac{\frac{1}{70} + \frac{1}{90}}{2}}.$$

Essa última expressão nos leva à definição de média harmônica: a *média harmônica* de um conjunto de valores $x_1, x_2, \ldots, x_n$ é o inverso da média aritmética dos inversos dos valores, isto é:

$$\boxed{\overline{x}_h = \frac{1}{\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}{n}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}}. \tag{3.4}$$

Analisando essa expressão, conclui-se que a velocidade média para a viagem completa é a média harmônica das velocidades médias desenvolvidas na ida e na volta. Note que a distância entre as cidades é irrelevante.

Exercício 3.1 *Durante 3 anos uma escola estadual teve uma verba fixa de R$500,00 para comprar folhas de cartolina. A compra é sempre feita na primeira semana de janeiro e os preços de cada folha de cartolina estão na tabela 3.5 abaixo: Qual o preço médio da folha de cartolina pago pela*

Tabela 3.5: Preço da cartolina para o exercício 3.1

Mês	Preço (R$)
jan/98	0,35
jan/99	0,45
jan/00	0,50

escola nos anos de 1998 a 2000?

Observação:

Embora não muito usual, o cálculo das médias geométrica e harmônica para dados agrupados será apresentado principalmente por aspectos didáticos, visando sua aplicação no estudo de números índices.

Suponhamos, então, que temos n_1 valores iguais a x_1, n_2 iguais a x_2 , ..., n_k iguais a x_k. Os valores x_i podem ou não ser pontos médios das classes de uma tabela de freqüências; o que importa é a repetição de cada um deles. Seja $n = n_1 + n_2 + \cdots + n_k$ o número total de observações.

A média geométrica, por definição, é:

$$\begin{aligned}\overline{x}_g &= \sqrt[n]{x_1 \times \cdots \times x_1 \times x_2 \times \cdots \times x_2 \times \cdots \times x_k \times \cdots \times x_k} = \sqrt[n]{x_1^{n_1} \times x_2^{n_2} \times \cdots \times x_k^{n_k}} = \\ &= \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} = \prod_{i=1}^{k} x_i^{f_i}\end{aligned}$$

Essa expressão será útil quando apresentarmos o índice de Divisia.

Para a média harmônica, temos que:

$$\begin{aligned}\overline{x}_h &= \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_2} + \cdots + \frac{1}{x_k} + \cdots + \frac{1}{x_k}} = \\ &= \frac{n}{n_1 \times \frac{1}{x_1} + n_2 \times \frac{1}{x_2} + \cdots + n_k \times \frac{1}{x_k}} = \\ &= \frac{1}{\frac{n_1}{n} \times \frac{1}{x_1} + \frac{n_2}{n} \times \frac{1}{x_2} + \cdots + \frac{n_k}{n} \times \frac{1}{x_k}} \Rightarrow\end{aligned}$$

$$\overline{x}_h = \frac{1}{f_1 \times \frac{1}{x_1} + f_2 \times \frac{1}{x_2} + \cdots + f_k \times \frac{1}{x_k}} = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

Essa última expressão será muito útil quando for apresentado o índice de Paasche.

Exercício 3.2 *Considere a situação do Exercício 3.1 com a seguinte modificação: as quantias gastas nos três anos são 500, 550 e 620. Nesse caso, qual é o preço médio da folha de cartolina? Qual é a diferença nas duas situações?*

3.3 Coeficiente de variação

Considere a seguinte situação: uma fábrica de ervilhas comercializa seu produto em embalagens de 300 gramas e em embalagens de um quilo ou 1000 gramas. Para efeitos de controle do processo de enchimento das embalagens, sorteia-se uma amostra de 10 embalagens de cada uma das máquinas, obtendo-se os seguintes resultados:

$$300g \longrightarrow \begin{cases} \overline{x} = 295g \\ \sigma = 5g \end{cases}$$
$$1000g \longrightarrow \begin{cases} \overline{x} = 995g \\ \sigma = 5g \end{cases}$$

Vamos interpretar esses números. Na primeira máquina, as embalagens deveriam estar fornecendo peso de 300g mas, devido a erros de ajuste da máquina de preenchimento, o peso médio das 10 embalagens é de apenas 295g. O desvio padrão de 5g significa que, em média, os pesos das embalagens estão 5 gramas abaixo ou acima do peso médio das 10 latas. Uma interpretação análoga vale para a segunda máquina.

Em qual das duas situações a variabilidade parece ser maior? Ou seja, em qual das duas máquinas parece haver um problema mais sério? Note que, em ambos os casos, há uma dispersão de 5g em torno da média, mas 5g em 1000g é menos preocupante que 5g em 300g.

Como um exemplo mais extremo, um desvio padrão de 10 unidades em um conjunto cuja observação típica é 100 é muito diferente de um desvio padrão de 10 unidades em um conjunto cuja observação típica é 10000. Surge, assim, a necessidade de uma medida de *dispersão relativa*, que permita comparar, por exemplo, esses dois conjuntos. Uma dessas medidas é o *coeficiente de variação.*

Definição 13 *Dado um conjunto de observações $x_1, x_2, \ldots, x_n$, o* ***coeficiente de variação*** *(CV) é definido como a razão entre o desvio padrão dos dados e sua média, ou seja:*

$$CV = \frac{\sigma}{\overline{x}}\,. \tag{3.5}$$

Note que, como o desvio padrão e a média são ambos medidos na mesma unidade dos dados originais, o coeficiente de variação é *adimensional*. Este fato permite comparações entre conjuntos de dados diferentes, medidos em unidades diferentes. Em geral, o CV é apresentado em forma percentual, isto é, multiplicado por 100.

No exemplo das latas de ervilha, os coeficientes de variação para as embalagens oriundas das 2 máquinas são

$$300g \longrightarrow CV = \frac{5}{300} \times 100 = 1,67\%$$
$$1000g \longrightarrow CV = \frac{5}{1000} \times 100 = 0,5\%$$

o que confirma a nossa observação anterior: a variabilidade na máquina de 300g é relativamente maior.

Exercício 3.3 *Faça uma análise comparativa do desempenho dos alunos e alunas de uma turma de Estatística, segundo as notas dadas a seguir. Para isso, calcule a média, o desvio padrão e o coeficiente de variação, comentando os resultados.*

Homens	*4,5*	*6,1*	*3,2*	*6,9*	*7,1*	*8,2*	*3,3*	*2,5*	*5,6*	*7,2*	*3,4*
Mulheres	*6,3*	*6,8*	*5,9*	*6,0*	*4,9*	*6,1*	*6,3*	*7,5*	*7,7*	*6,5*	

3.4 Escores padronizados

Considere os dois conjuntos de dados abaixo, que representam as notas em Estatística e Cálculo dos alunos de uma determinada turma.

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	3	5	5	7
Cálculo	6	8	9	10	7	7	8	9	5

As notas médias nas duas disciplinas são:

$$\begin{aligned} \overline{x}_E &= \frac{6+4+5+7+8+3+5+5+7}{9} = \frac{50}{9} = 5,5556 \\ \overline{x}_C &= \frac{6+8+9+10+7+7+8+9+5}{9} = \frac{69}{9} = 7,6667 \end{aligned}$$

As variâncias são:

$$\begin{aligned} \sigma_E^2 &= \frac{6^2+4^2+5^2+7^2+8^2+3^2+5^2+5^2+7^2}{9} - \left(\frac{50}{9}\right)^2 = \frac{298}{9} - \frac{2500}{81} = \\ &= \frac{298 \times 9 - 2500}{81} = \frac{182}{81} = 2,246914 \end{aligned}$$

$$\begin{aligned} \sigma_C^2 &= \frac{6^2+8^2+9^2+10^2+7^2+7^2+8^2+9^2+5^2}{9} - \left(\frac{69}{9}\right)^2 = \frac{549}{9} - \frac{4761}{81} = \\ &= \frac{549 \times 9 - 4761}{81} = \frac{180}{81} = 2,222222 \end{aligned}$$

Os desvios padrões são:

$$\begin{aligned} \sigma_E &= \sqrt{\frac{182}{81}} = 1,498971 \\ \sigma_C &= \sqrt{\frac{180}{81}} = 1,490712 \end{aligned}$$

Analisando os dois conjuntos de notas, pode-se ver que o aluno 1 tirou 6 em Estatística e em Cálculo. No entanto, a nota média em Estatística foi 5,56, enquanto que em Cálculo a nota média foi 7,67. Assim, o 6 em Estatística "vale mais" que o 6 em Cálculo, no sentido de que ele está mais próximo da nota média. Uma forma de medir tal fato é considerar a posição relativa de cada aluno

no grupo. Para isso, o primeiro passo consiste em comparar a nota do aluno com a média do grupo, considerando o seu desvio em torno da média. Se x_i é a nota do aluno, passamos a trabalhar com $x_i - \overline{x}$. Dessa forma vemos que a nota 6 em Estatística gera um desvio de 0,44, enquanto a nota 6 em Cálculo gera um desvio de -1,67, o que significa que o aluno 1 tirou nota acima da média em Estatística e nota abaixo da média em Cálculo.

Um outro problema que surge na comparação do desempenho nas 2 disciplinas é o fato de o desvio padrão ser diferente nas 2 matérias. A variabilidade em Estatística foi um pouco maior que em Cálculo. Assim, o segundo passo consiste em padronizar a escala. Essa padronização da escala se faz dividindo os desvios em torno da média pelo desvio padrão do conjunto, o que nos dá o escore padronizado:

$$z_i = \frac{x_i - \overline{x}}{\sigma_x}. \tag{3.6}$$

O desvio padrão das notas de Estatística é $\sigma_E = 1,49897$ e das notas de Cálculo é $\sigma_C = 1,49071$. Na tabela a seguir temos os escores padronizados; podemos ver aí que o escore relativo à nota 6 em Estatística é maior que o escore da nota 6 em Cálculo, indicando que a primeira "vale mais" que a segunda.

Aluno	1	2	3	4	5	6	7	8	9
Estatística	$0,297$	$-1,038$	$-0,371$	$0,964$	$1,631$	$-1,705$	$-0,371$	$-0,371$	$0,964$
Cálculo	$-1,118$	$0,224$	$0,894$	$1,565$	$-0,447$	$-0,447$	$0,224$	$0,894$	$-1,789$

Da mesma forma, o 5 em Estatística do aluno 7 vale mais que o 5 em Cálculo do aluno 9: ambos estão abaixo da média, mas o 7 em Estatística está "mais próximo" da média.

Ao padronizarmos os dados, a nossa escala passa a ser definida em termos de desvio padrão. Ou seja, passamos a dizer que tal observação está abaixo (ou acima) da média por determinado número de desvios padrões. Com isso, tira-se o efeito de as médias e as variabilidades serem diferentes.

Podemos escrever o escore padronizado como

$$z_i = \frac{1}{\sigma_x} x_i - \frac{\overline{x}}{\sigma_x}$$

e daí vemos que esse escore é obtido a partir dos dados originais por uma transformação linear: somamos uma constante $\left(-\frac{\overline{x}}{\sigma_x}\right)$ e multiplicamos por outra constante $\left(\frac{1}{\sigma_x}\right)$. Das propriedades da média e do desvio padrão vistas nas aulas anteriores, resulta que a média e o desvio padrão dos escores padronizados podem ser obtidos a partir da média e do desvio padrão dos dados originais:

$$\begin{aligned} \overline{z} &= \frac{1}{\sigma_x}\overline{x} - \frac{\overline{x}}{\sigma_x} = 0 \\ \sigma_z^2 &= \frac{1}{\sigma_x^2}\sigma_x^2 = 1 \end{aligned}$$

Logo, os escores padronizados têm sempre média zero e desvio padrão (ou variância) 1.

3.4.1 Teorema de Chebyshev e valores discrepantes

Os escores padronizados podem ser usados para se detectarem valores discrepantes ou muito afastados do conjunto de dados, graças ao Teorema de Chebyshev.

Teorema 1 ***Teorema de Chebyshev***

Para qualquer distribuição de dados, pelo menos $\left(1 - 1/z^2\right)$ *dos dados estão dentro de* z *desvios padrões da média, onde* z *é qualquer valor maior que 1. Dito de outra forma, pelo menos* $\left(1 - 1/z^2\right)$ *dos dados estão no intervalo* $[\overline{x} - z\sigma; \overline{x} + z\sigma]$.

Vamos analisar esse teorema em termos dos escores padronizados. Suponha que x' seja um valor do conjunto de dados dentro do intervalo $[\overline{x} - z\sigma; \overline{x} + z\sigma]$. Isso significa que

$$\overline{x} - z\sigma < x' < \overline{x} + z\sigma$$

Subtraindo $\overline{x}$ e dividindo por σ todos os termos dessa desigualdade obtemos que

$$\begin{aligned} \frac{\overline{x} - z\sigma - \overline{x}}{\sigma} &< \frac{x' - \overline{x}}{\sigma} < \frac{\overline{x} + z\sigma - \overline{x}}{\sigma} \Rightarrow \\ -z &< \frac{x' - \overline{x}}{\sigma} < +z \end{aligned}$$

O termo do meio nada mais é que o escore padronizado da observação x'. Assim, o teorema de Chebyshev pode ser estabelecido em termos dos escores padronizados como: para pelo menos $\left(1 - 1/z^2\right)$ dos dados, os respectivos escores padronizados estão no intevalo $(-z, +z)$, onde z é qualquer valor maior que 1.

O fato interessante desse teorema é que ele vale para qualquer distribuição de dados. Vamos ver alguns exemplos numéricos.

- $z = 2$

  Nesse caso, $1 - 1/z^2 = 3/4$, ou seja, para pelo menos 75% dos dados, os escores padronizados estão no intervalo $(-2, +2)$.

- $z = 3$

  Nesse caso, $1 - 1/z^2 = 8/9 = 0,889$, ou seja, para aproximadamente 89% dos dados, os escores padronizados estão no intervalo $(-3, +3)$.

- $z = 4$

  Nesse caso, $1 - 1/z^2 = 15/16 = 0,9375$, ou seja, para 93,75% dos dados, os escores padronizados estão no intervalo $(-4, +4)$.

Como regra de detecção de valores discrepantes, pode-se usar o Teorema de Chebyshev para se estabelecer, por exemplo, que dados cujos escores padronizados estejam fora do intervalo $(-3, +3)$ são valores discrepantes e, portanto, devem ser verificados cuidadosamente para se identificar a causa de tal discrepância. Algumas vezes, tais valores podem ser resultados de erros, mas muitas vezes eles são valores legítimos e a presença deles requer alguns cuidados na análise estatística.

Exercício 3.4 *Considere os dados da Tabela 3.6 sobre a densidade populacional das unidades da federação brasileira. Calcule os escores padronizados e determine se alguma UF pode ser considerada valor discrepante com relação a essa variável.*

Tabela 3.6: Densidade populacional dos estados brasileiros, para a Atividade 5.2

UF	Densidade Populacional (hab/km^2)	UF	Densidade Populacional (hab/km^2)
RO	6	SE	81
AC	4	BA	24
AM	2	MG	31
RR	2	ES	68
PA	5	RJ	328
AP	4	SP	149
TO	5	PR	48
MA	17	SC	57
PI	12	RS	37
CE	51	MS	6
RN	53	MT	3
PB	61	GO	15
PE	81	DF	353
AL	102		

Fonte: IBGE - Censo Demográfico 2000

3.5 Exercícios Complementares

Exercício 3.5 *A contagem de bactérias em uma cultura aumentou de 2500 para 9200 em três dias. Qual o acréscimo percentual diário médio?*

Exercício 3.6 *Na tabela 3.7 temos as variações mensais do IPCA (Índice de Preços ao Consumidor Amplo) calculadas pelo IBGE para o ano de 1999. Segundo previsões feitas pelo então secretário-adjunto de Política Econômica (Folha de São Paulo, 11/12/1999), o IPCA no ano de 1999 deveria ficar abaixo de 9%. Para que as previsões do secretário se confirmassem, qual deveria ter sido a taxa máxima do IPCA em dezembro?*

Tabela 3.7: IPCA 1999 para o exercício 3.6

Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov
0,70	1,05	1,10	0,56	0,30	0,19	1,09	0,56	0,31	1,19	0,95

Fonte: IBGE

Exercício 3.7 *Para se estudar o desempenho de 2 companhias corretoras de ações, selecionou-se de cada uma delas amostras das ações negociadas. Para cada ação selecionada, computou-se a porcentagem de lucro apresentada durante um período fixado de tempo, obtendo-se os dados abaixo. Com base nos coeficientes de variação, qual companhia teve melhor desempenho?*

Corretora A

38 45 48 48 54 54 55 55 55 55 56 59 60 60 62 64 65 70

Corretora B

50 50 51 52 52 53 54 55 55 55 56 56 57 57 57 58 58 59 59 59 61

Capítulo 4

Análise Bidimensional

4.1 Introdução

Até o momento, vimos como organizar e resumir informações referentes a uma única variável. No entanto, é bastante freqüente depararmos com situações onde há interesse em estudar conjuntamente duas ou mais variáveis. Num estudo sobre mortalidade infantil, por exemplo, é importante acompanhar também o tratamento pré-natal da mãe; espera-se, neste caso, que haja uma diminuição da taxa de mortalidade infantil com o aumento dos cuidados durante a gravidez.

Neste capítulo nos deteremos no estudo de distribuições bidimensionais, dando ênfase à forma de representação tabular e gráfica e às medidas de associação entre variáveis quantitativas. Seguindo convenção usual, denotaremos por uma letra maiúscula a variável em estudo e pela letra minúscula correspondente o valor observado da variável.

4.2 Variáveis qualitativas

Consideremos inicialmente o caso de duas variáveis *qualitativas*. Como exemplo, vamos trabalhar com os dados apresentados na Tabela 4.1, onde temos o grau de instrução, que representaremos pela letra Y, e o sexo, que representaremos pela letra X, de 20 funcionários de uma empresa. Cada funcionário pesquisado dá origem a um par de valores (x_i, y_i).

4.2.1 Distribuição conjunta de freqüências

Uma forma alternativa de representar essas informações é através de uma distribuição ou tabela conjunta de freqüências, como a apresentada na Tabela 4.2. Como temos duas variáveis de interesse X e Y, precisamos das duas dimensões, linha e coluna, para representar as informações disponíveis, que serão apresentadas em forma de contagem ou freqüência. A escolha da variável linha e da variável coluna depende do objetivo do estudo. Se existe entre as variáveis uma relação do tipo dependente/explanatória, isto é, se queremos usar uma das variáveis para "explicar" a outra, então é costume colocar a variável explanatória na coluna e denotá-la variável X. Caso contrário, qualquer uma das duas pode ser a variável coluna. No exemplo, estamos interessados em analisar a diferença no grau de instrução entre os sexos; sendo assim, o sexo é a variável explanatória X e o grau de instrução é a variável explicada ou dependente Y.

Em cada cela temos o número de indivíduos que pertencem simultaneamente às respectivas categorias. Assim, podemos ver que há 2 homens com 1º grau, 4 mulheres com 2º grau e assim por

Tabela 4.1: Grau de instrução e sexo de 20 funcionários de uma empresa

Funcionário	Grau de instrução	Sexo
1	1º grau	Masculino
2	2º grau	Feminino
3	2º grau	Feminino
4	3º grau	Masculino
5	3º grau	Feminino
6	1º grau	Feminino
7	2º grau	Masculino
8	2º grau	Feminino
9	3º grau	Feminino
10	1º grau	Feminino
11	3º grau	Masculino
12	2º grau	Masculino
13	1º grau	Feminino
14	2º grau	Feminino
15	2º grau	Masculino
16	3º grau	Feminino
17	1º grau	Masculino
18	3º grau	Feminino
19	2º grau	Masculino
20	3º grau	Feminino

Fonte: Dados hipotéticos

diante. Como já visto no caso univariado, essa forma de apresentação é mais interessante, uma vez que não estamos interessados na observação individual e, sim, no comportamento dos grupos.

Tabela 4.2: Distribuição conjunta do grau de instrução e sexo dos funcionários de uma empresa

	Sexo		
Grau de instrução	Feminino	Masculino	Total
1º grau	3	2	5
2º grau	4	4	8
3º grau	5	2	7
Total	12	8	20

Além das contagens em cada cela, acrescentamos também a linha e a coluna com os respectivos totais. Os totais das linhas, então, nos dizem que há 5 funcionários com 1º grau, 8 com 2º grau e 7 com 3º grau. Já os totais das colunas nos dizem que há 8 funcionários do sexo masculino e 12 do sexo feminino. O total de funcionários (20) pode ser obtido somando-se os totais das linhas (grau de instrução): $5 + 8 + 7 = 20$ ou das colunas (sexo): $8 + 12 = 20$.

Na construção de tabelas de freqüências univariadas, foi acrescentada à tabela a coluna de freqüências relativas, que davam a proporção de elementos em cada classe com relação ao número total de elementos. Um procedimento análogo pode ser feito para as tabelas bidimensionais; a diferença é que, neste caso, existem três possibilidades de expressarmos as proporções de cada cela: (i) com relação ao total geral; (ii) com relação ao total de cada linha e (iii) com relação ao total de cada coluna. A escolha entre essas três possibilidades deverá ser feita de acordo com o objetivo da

análise. Nas Tabelas 4.3 a 4.5 temos as três versões para os dados da Tabela 4.2 usando freqüências relativas.

Tabela 4.3: Distribuição conjunta relativa do grau de instrução e sexo de 20 funcionários de uma empresa

	Sexo		
Grau de instrução	Feminino	Masculino	Total
1º grau	0,150	0,100	0,250
2º grau	0,200	0,200	0,400
3º grau	0,250	0,100	0,350
Total	0,600	0,400	1,000

Tabela 4.4: Distribuição condicional do grau de instrução, dado o sexo, de 20 funcionários de uma empresa

	Sexo		
Grau de instrução	Feminino	Masculino	Total
1º grau	0,250	0,250	0,250
2º grau	0,333	0,500	0,400
3º grau	0,417	0,250	0,350
Total	1,000	1,000	1,000

Tabela 4.5: Distribuição condicional do sexo, dado o grau de instrução, de 20 funcionários de uma empresa

	Sexo		
Grau de instrução	Feminino	Masculino	Total
1º grau	0,600	0,400	1,000
2º grau	0,500	0,500	1,000
3º grau	0,714	0,286	1,000
Total	0,600	0,400	1,000

Da Tabela 4.3 podemos concluir que 15% dos funcionários da empresa são do sexo Feminino e têm 1º grau, enquanto 10% são do sexo masculino e têm 3º grau. Essa é a *tabela da distribuição conjunta relativa*; em cada cela temos a freqüência relativa dos indivíduos que pertencem simultaneamente às duas categorias em questão.

Da Tabela 4.4 podemos ver que 25% das mulheres têm 1º grau, enquanto 50% dos homens têm 2º grau. Essa tabela nos dá a *distribuição condicional do grau de instrução (variável linha), dado o sexo (variável coluna).* Na coluna Total temos a distribuição do grau de instrução (variável linha) na população completa. Esse total, obviamente, coincide com os totais das colunas na Tabela 4.3. Essa é a tabela apropriada para a análise desejada, de comparar os sexos segundo o grau de instrução. Daí podemos ver, por exemplo, que a porcentagem de mulheres com 3º grau é maior, havendo uma inversão para o 2º grau.

Finalmente, da Tabela 4.5 conclui-se, por exemplo, que, dos funcionários com 1º grau, 60% são mulheres e 40% são homens, enquanto que dos funcionários com 3º grau, 71,4% são mulheres e 28,6% são homens. Essa é a *distribuição condicional do sexo (variável coluna) dado o grau de instrução*

(variável linha). Na linha Total temos a distribuição por sexo na população completa, que coincide com os totais das linhas da Tabela 4.3.

Um ponto importante a salientar é que, na construção de tabelas com freqüências relativas, um cuidado especial deve ser tomado com relação ao arredondamento dos números. Arredondamentos excessivos podem fazer com que os totais de linhas e/ou colunas não somem 1!

Para formalizar o procedimento de construção das tabelas acima, suponhamos que as variáveis de interesse sejam Y e X, que assumem, respectivamente, os valores $\{y_1, \ldots, y_I\}$ e $\{x_1, \ldots, x_J\}$. Os valores dessas variáveis são colocados nas linhas e colunas de uma tabela de dupla entrada. Como já dito, a escolha de qual variável será a variável coluna X depende, em geral, do objetivo do estudo. Olhando os dados como vindos de uma grande população, as categorias das duas variáveis podem ser vistas como subpopulações. Normalmente escolhe-se a variável coluna X como sendo aquela que define as subpopulações que queremos comparar. É como se fôssemos usar a variável coluna X para "explicar" a variável linha Y (lembre a notação usual de função: $y = f(x)$). No exemplo anterior, essas subpopulações de interesse eram definidas pelo sexo do funcionário, ou seja, o interesse era comparar a instrução **por** sexo, numa tentativa de explicar o grau de instrução pelo sexo do funcionário.

Vamos denotar por n_{ij} a contagem simples de cada cela, referente à categoria i da primeira variável e à categoria j da segunda variável. Como no caso univariado, seja n o número total de observações; então:

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \quad . \tag{4.1}$$

Os totais por linha serão denotados por $n_{i\cdot}$ e os totais por coluna por $n_{\cdot j}$, de modo que

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij} \qquad i = 1, \ldots, I \tag{4.2}$$

e

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij} \qquad j = 1, \ldots, J \tag{4.3}$$

(O $\cdot$ indica qual a variável, linha ou coluna, que está sendo totalizada.) Então, a tabela de freqüências genérica, análoga à Tabela 4.2, é

		X				
		x_1	x_2	$\cdots$	x_J	Total
	y_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
	y_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
Y	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	y_I	n_{I1}	$\cdots$	$\cdots$	n_{IJ}	$n_{I\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	n

A tabela de freqüências conjuntas (Tabela 4.3) é:

		X				
		x_1	x_2	$\cdots$	x_J	Total
	y_1	f_{11}	f_{12}	$\cdots$	f_{1J}	$f_{1\cdot}$
	y_2	f_{21}	f_{22}	$\cdots$	f_{2J}	$f_{2\cdot}$
Y	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	y_I	f_{I1}	f_{I2}	$\cdots$	f_{IJ}	$f_{I\cdot}$
	Total	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot J}$	n

onde

$$f_{ij} = \frac{n_{ij}}{n} \qquad i = 1, \ldots, I;\ j = 1, \ldots, J$$

$$f_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad i = 1, \ldots, I$$

$$f_{\cdot j} = \frac{n_{\cdot j}}{n} \qquad j = 1, \ldots, J$$

A notação para a distribuição condicional da variável linha dada a variável coluna (Tabela 4.4) é um pouco mais complicada. Cada cela da j-ésima coluna, $j = 1, \ldots, J$, é calculada como o percentual da cela com relação ao total da coluna j. Vamos denotar essas freqüências relativas por $f_{i|j}$ (lê-se freqüência de i dado j). Note que esses valores são as freqüências para a categoria i da variável linha *dado* o valor j da variável coluna. Então, a tabela da distribuição condicional da variável linha *dada* a variável coluna é:

		X				
		x_1	x_2	$\cdots$	x_J	Total
	y_1	$f_{1\|1}$	$f_{1\|2}$	$\cdots$	$f_{1\|J}$	$f_{1\cdot}$
	y_2	$f_{2\|1}$	$f_{2\|2}$	$\cdots$	$f_{2\|J}$	$f_{2\cdot}$
Y	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	y_I	$f_{I\|1}$	$f_{I\|2}$	$\cdots$	$f_{I\|J}$	$f_{I\cdot}$
	Total	$1,00$	$1,00$	$\cdots$	$1,00$	$1,00$

onde

$$f_{i|j} = \frac{n_{ij}}{n_{\cdot j}} \qquad i = 1, \ldots, I;\ j = 1, \ldots, J$$

A coluna Total nessa tabela é a freqüência da categoria i daa variável linha relativa ao total geral:

$$f_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad i = 1, \ldots, I$$

Analogamente, a tabela da distribuição condicional da variável coluna *dada* a variável linha (Tabela 4.5) é:

		X				
		x_1	x_2	$\cdots$	x_J	Total
	y_1	$f_{1\|1}$	$f_{2\|1}$	$\cdots$	$f_{J\|1}$	$1,00$
	y_2	$f_{1\|2}$	$f_{2\|2}$	$\cdots$	$f_{J\|2}$	$1,00$
Y	$\vdots$	$\vdots$		$\ddots$	$\vdots$	$\vdots$
	y_I	$f_{1\|I}$	$f_{2\|I}$	$\cdots$	$f_{J\|I}$	$1,00$
	Total	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot J}$	$1,00$

onde

$$f_{j|i} = \frac{n_{ij}}{n_{i\cdot}} \qquad i = 1, \ldots, I;\ j = 1, \ldots, J$$

$$f_{\cdot j} = \frac{n_{\cdot j}}{n} \qquad j = 1, \ldots, J$$

Alguns pacotes estatísticos têm a opção de apresentar todos os três tipos de freqüências relativas em uma única tabela. Na Tabela 4.6 apresentamos, como exemplo, a saída do programa SAS[1]

[1] Statistical Analysis System

referente aos dados da Tabela 4.1. A legenda para cada cela está na cela do canto superior esquerdo da tabela: FREQUENCY indica a freqüência absoluta, PERCENT indica o percentual com relação ao total geral (equivalente à Tabela 4.3), ROW PCT indica os percentuais com relação ao total das linhas (equivalente à Tabela 4.5) e COL PCT indica os percentuais com relação ao total das colunas (equivalente à Tabela 4.4).

Tabela 4.6: Tabela de freqüências bidimensional gerada pelo SAS com base nos dados da tabela 4.1

INST (grau de instrucao)			
	SEXO		
Frequency			
Percent			
Row Pct			
Col Pct	Feminino	Masculino	Total
1o. grau	3	2	5
	15.00	10.00	25.00
	60.00	40.00	
	25.00	25.00	
2o. grau	4	4	8
	20.00	20.00	40.00
	50.00	50.00	
	33.33	50.00	
3o. grau	5	2	7
	25.00	10.00	35.00
	71.43	28.57	
	41.67	25.00	
Total	12	8	20
	60.00	40.00	100.00

Exercício 4.1 *Na tabela abaixo temos dados sobre hábitos de fumar de uma amostra de moradores de uma pequena cidade (dados fictícios).*

Hábitos	*Idade*		
de fumo	< 20	$[20, 30)$	≥ 30
Fumante	*143*	*171*	*40*
Ex-fumante	*11*	*152*	*140*
Nunca fumou	*66*	*57*	*20*

Responda às seguintes perguntas:

1. *Nesse grupo, quantas pessoas com 30 anos ou mais são fumantes?*
2. *Quantos ex-fumantes há nesse grupo?*
3. *Qual é o número total de pessoas nesse grupo?*
4. *Dentre as pessoas que nunca fumaram, qual é o percentual de jovens, isto é, pessoas com menos de 20 anos?*

5. *Construa um gráfico apropriado para ver se existe diferença entre os grupos etários com relação ao hábito de fumar.*

4.3 Variáveis quantitativas

4.3.1 Diagramas de dispersão

Quando as variáveis envolvidas em uma análise bidimensional são do tipo *quantitativo* (salário, idade, altura, etc.), é possível construir tabelas de dupla entrada de modo análogo ao visto para variáveis qualitativas. Algumas vezes, para evitar um grande número de entradas, pode ser necessário agrupar os dados em classes, da mesma forma que fizemos para tabelas unidimensionais. No entanto, pela sua própria natureza, as variáveis quantitativas são passíveis de métodos de análise mais refinados.

Um instrumento de análise bastante útil é o diagrama de dispersão, que é um gráfico bidmensional onde os valores das variáveis envolvidas são representados como pares ordenados no plano cartesiano. Na Tabelas 4.7 a 4.9 temos três conjuntos de dados, cujos diagramas de dispersão se encontram nas Figuras 4.1 a 4.3. Nesses gráficos, as linhas pontilhadas estão passando pelo ponto central do conjunto, isto é, pelo ponto $(\overline{x}, \overline{y})$.

Tabela 4.7: Variação diária das Bolsas de Valores - Junho 1993

Dia	Variação percentual		Dia	Variação percentual	
	Bovespa	BVRJ		Bovespa	BVRJ
1	4,9935	6,9773	17	-4,6706	-6,2360
2	5,5899	6,1085	18	0,6629	2,6259
3	3,8520	2,4847	21	1,1651	0,8728
4	0,9984	-0,1044	22	3,2213	4,8243
7	2,4872	2,4942	23	-2,7226	-4,7266
8	0,0142	0,1239	24	1,2508	-0,4985
9	-1,7535	-0,4221	25	7,1845	6,6798
11	8,1764	9,5148	28	2,5674	1,2299
14	0,6956	-1,7350	29	-1,3235	-3,0375
15	1,6164	2,2749	30	1,6685	1,2303
16	7,5829	15,4173			

Fonte: Folha de São Paulo (índice de fechamento)

Analisando esses gráficos , pode-se ver que as relações entre as variáveis envolvidas mudam; na Figura 4.1 existe uma tendência crescente entre as variáveis, isto é, quando o índice da Bovespa aumenta, o índice da BVRJ também tende a aumentar. Na Figura 4.2 essa relação se inverte, ou seja, aumentando a latitude, a temperatura tende a diminuir. Já na Figura 4.3 não é possível estabelecer nenhuma relação entre as variáveis, contrariando a superstição de que linhas da vida longas indicam maior longevidade.

4.3.2 Covariância

Vamos estudar, agora, uma medida de associação entre variáveis, que está relacionada ao tipo mais simples de associação: a linear. Então, tal medida irá representar o quanto a "nuvem" de dados em um diagrama de dispersão se aproxima de uma reta.

Tabela 4.8: Latitude e temperatura média de 15 cidades dos EUA

Latitude	Temperatura (^{o}F)
34	56,4
32	51,0
39	36,7
39	37,8
41	36,7
45	18,2
41	30,1
33	55,9
34	46,6
47	13,3
44	34,0
39	36,3
41	34,0
32	49,1
40	34,5

Fonte: Dunn e Clark (1974) p. 250

Tabela 4.9: Idade ao morrer e comprimento da "linha da vida"

Idade (anos)	Linha da vida (cm)	Idade (anos)	Linha da vida (cm)	Idade (anos)	Linha da da vida (cm)
19	9,75	65	8,85	74	8,85
40	9,00	65	9,75	74	9,60
42	9,60	66	8,85	75	6,45
42	9,75	66	9,15	75	9,76
47	11,25	66	10,20	75	10,20
49	9,45	67	9,15	76	6,00
50	11,25	68	7,95	77	8,85
54	9,00	68	8,85	80	9,00
56	7,95	68	9,00	82	9,75
56	12,00	69	7,80	82	10,65
57	8,10	69	10,05	82	13,20
57	10,20	70	10,50	83	7,95
58	8,55	71	9,15	86	7,95
61	7,20	71	9,45	88	9,15
62	7,95	71	9,45	88	9,75
62	8,85	72	9,45	94	9,00
65	8,25	73	8,10		

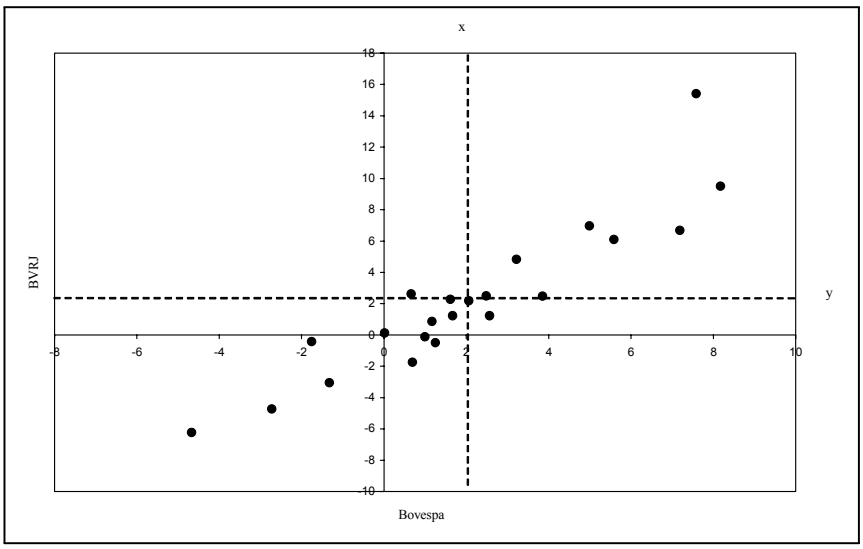


Figura 4.1: Variação diária das Bolsas de Valores - dados originais

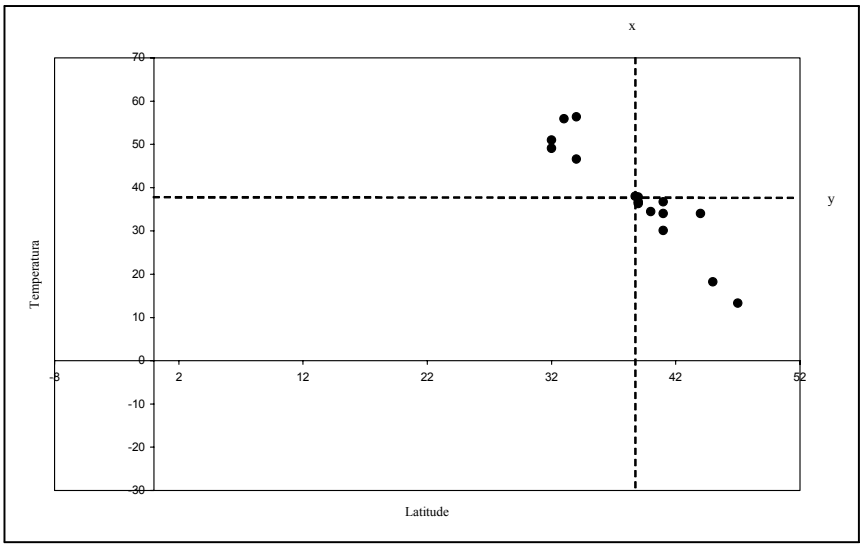


Figura 4.2: Latitude e temperatura média de 15 cidades dos EUA - dados originais

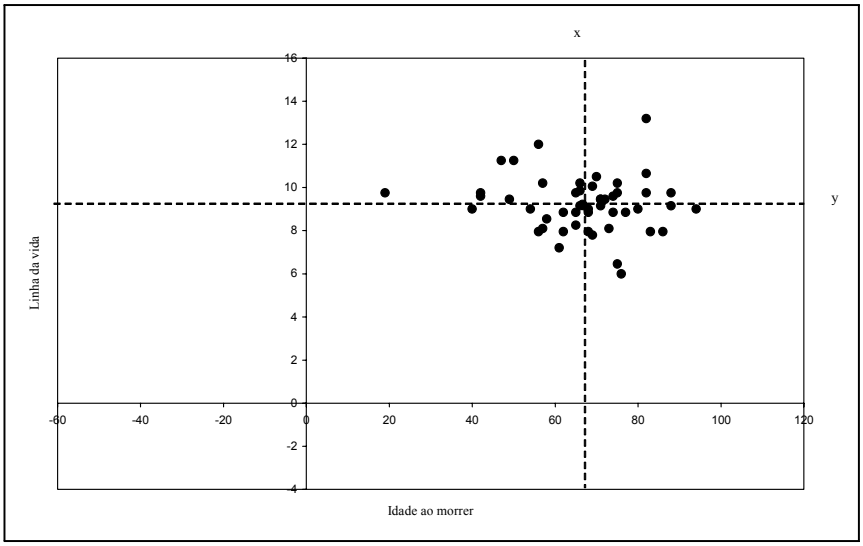


Figura 4.3: Diagrama de dispersão do comprimento da linha da vida e idade ao morrer - dados originais

Para diferenciar as três situações ilustradas nos gráficos anteriores, um primeiro ponto que devemos observar é o fato de as três "nuvens" de pontos estarem centradas em pontos diferentes, representados pela interseção dos eixos em linha pontilhada; note que esse é o ponto $(\overline{x}, \overline{y})$. Para facilitar comparações, é interessante uniformizar a origem, colocando as três nuvens centradas na origem $(0, 0)$. Lembrando as propriedades da média aritmética, sabe-se que a transformação $x_i - \overline{x}$ resulta em um conjunto de dados com média zero. Então, para quantificar as diferenças entre os gráficos anteriores, o primeiro ponto a considerar é a centralização da nuvem: em vez de trabalharmos com os dados originais (x_i, y_i),vamos trabalhar com os dados transformados $(x_i - \overline{x}, y_i - \overline{y})$. Nas Figuras 4.4 a 4.6 estão representados os diagramas de dispersão para essas variáveis transformadas.

Analisando esses três últimos gráficos, pode-se ver que, para o primeiro conjunto de dados, onde a tendência entre as variáveis é crescente, a maioria dos pontos está no primeiro e terceiro quadrante, enquanto que, no segundo gráfico, onde a relação é decrescente, a maioria dos pontos está no segundo e quarto quadrantes.

O primeiro e terceiro quarantes se caracterizam pelo fato de as abcissas e ordenadas terem o mesmo sinal e, portanto, seu produto é positivo; já no segundo e quarto quadrantes, as abcissas e ordenadas têm sinais opostos e, portanto, seu produto é negativo. Então, para diferenciar esses gráficos, podemos usar uma medida baseada no produto das coordenadas $x_i - \overline{x}$ e $y_i - \overline{y}$. Como no caso da variância ou desvio médio absoluto, para considerar todos os pares possíveis e descontar o número de observações, vamos tomar o valor médio desses produtos; define-se, assim, a *covariância entre as variáveis* X *e* Y por

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \tag{4.4}$$

onde x_i e y_i são os valores observados. No gráfico 4.6, os pontos estão espalhados nos quatro quadrantes e, assim, essa média tende a ser nula, ou melhor, próxima de zero.

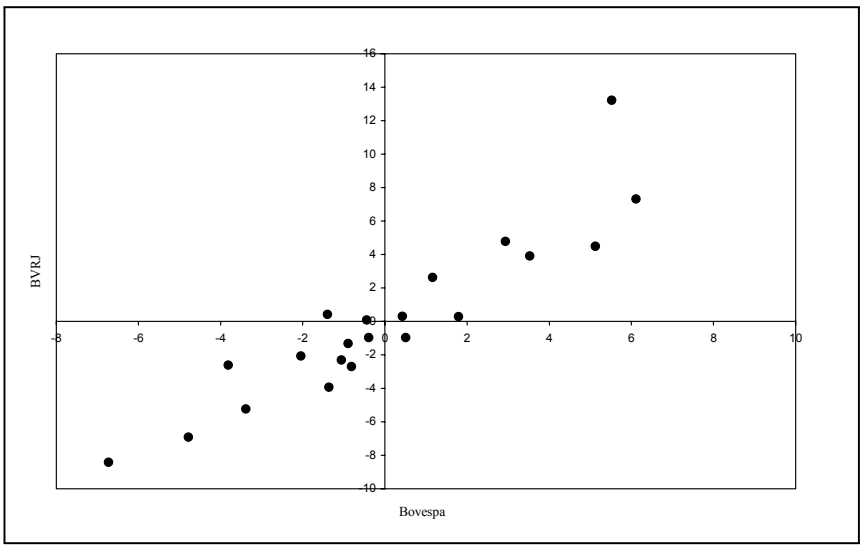


Figura 4.4: Variação diária das Bolsas de Valores - dados centrados na média

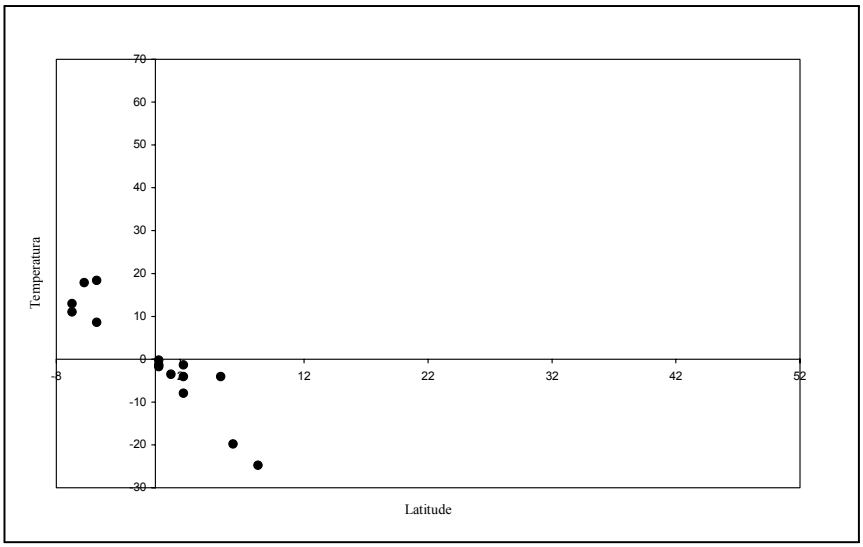


Figura 4.5: Latitude e temperatura média de 15 cidades dos EUA - dados centrados na média

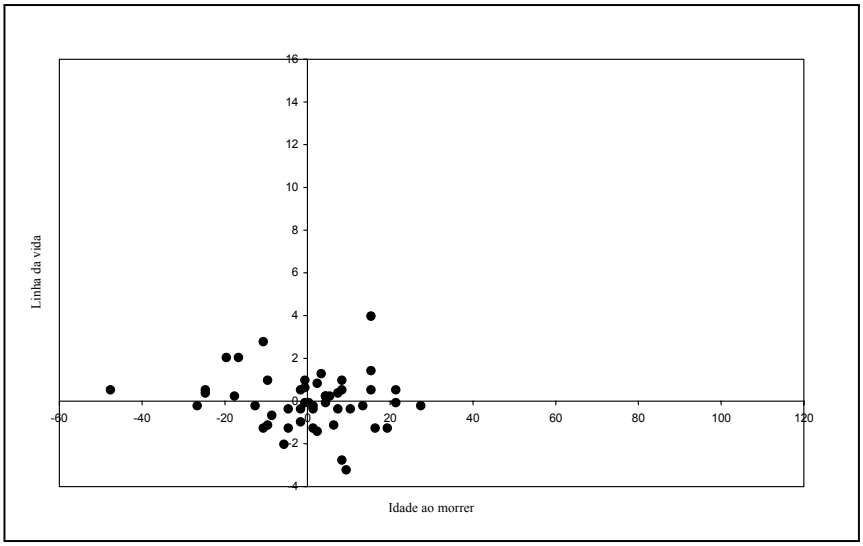


Figura 4.6: Diagrama de dispersão do comprimento da linha da vida e idade ao morrer - dados centrados na média

De maneira análoga à desenvolvida para a variância, a fórmula acima não é conveniente para fazer cálculos em máquinas de calcular mais simples. Assim, vamos desenvolver uma expressão alternativa. Note que:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &= \sum_{i=1}^{n}(x_iy_i - x_i\overline{y} - y_i\overline{x} + \overline{x}\,\overline{y}) = \\
&= \sum_{i=1}^{n} x_iy_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\overline{x\,y} = \\
&= \sum_{i=1}^{n} x_iy_i - \overline{y}n\overline{x} - \overline{x}n\overline{y} + n\overline{x\,y} = \\
&= \sum_{i=1}^{n} x_iy_i - n\overline{x\,y}
\end{aligned}$$

Logo,

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\left(\sum_{i=1}^{n} x_iy_i - n\overline{x\,y}\right) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \overline{x\,y} \tag{4.5}$$

Da Equação (4.5) podemos ver que a covariância é a *média do produto menos o produto das médias.* Resulta também que a covariância entre X e X é a variância de X, isto é: $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

É bastante importante salientar a interpretação da covariância: ela mede o grau de *associação linear* entre variáveis. Considerando o diagrama de dispersão da Figura 4.7, pode-se ver que existe uma associação quadrática perfeita entre as variáveis; no entanto, a covariância entre elas é nula!

4.3.3 Coeficiente de correlação

Um dos problemas da covariância é a sua dependência da escala dos dados, o que faz com que seus valores possam variar de $-\infty$ a $+\infty$. Note que sua unidade de medida é dada pelo produto das unidades de medida das variáveis X e Y envolvidas. Então, fica difícil comparar situações como as ilustradas nos gráficos das Figuras 4.8 e 4.9; para a primeira, temos que $\mathrm{Cov}(X,Y) = 304,51$ e

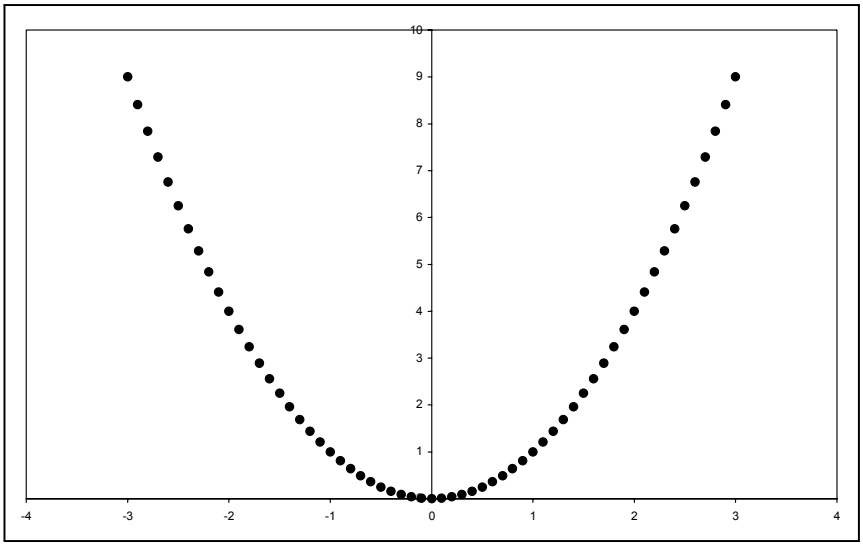


Figura 4.7: Conjunto de dados com covariância nula

para a segunda, $\text{Cov}(X, Y) = 609,02$. No entanto, os valores de X no primeiro conjunto variam de $-4,6706$ a $8,1764$ com um desvio padrão de $3,2757$ e no segundo conjunto de dados, variam de $-9,3412$ a $16,3528$, com um desvio padrão de $6,5514$.

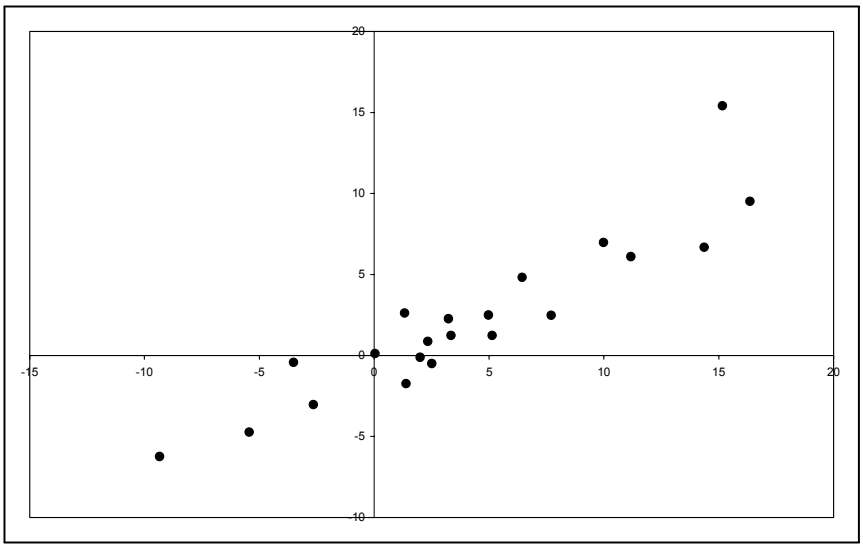


Figura 4.8: Influência da escala na covariância - parte (a)

Para uniformizar as escalas dos dados, o artifício será trabalhar com as variáveis padronizadas, isto é, $\dfrac{x_i - \overline{x}}{\sigma_X}$ e $\dfrac{y_i - \overline{y}}{\sigma_Y}$. Como já visto no estudo dos escores padronizados, cada um dos conjuntos de dados assim transformados tem desvio padrão igual a 1. Nas Figuras 4.10 a 4.12 temos o diagrama de dispersão para os dados transformados.

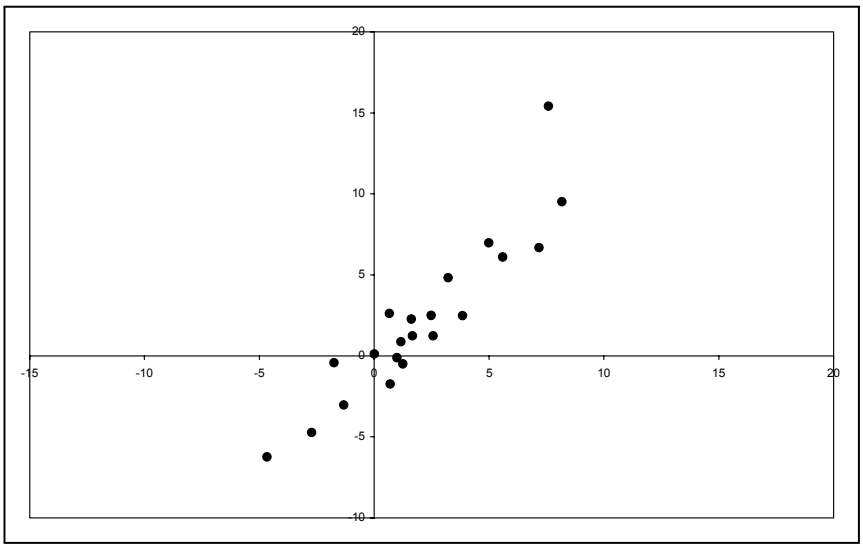


Figura 4.9: Influência da escala na covariância - parte (b)

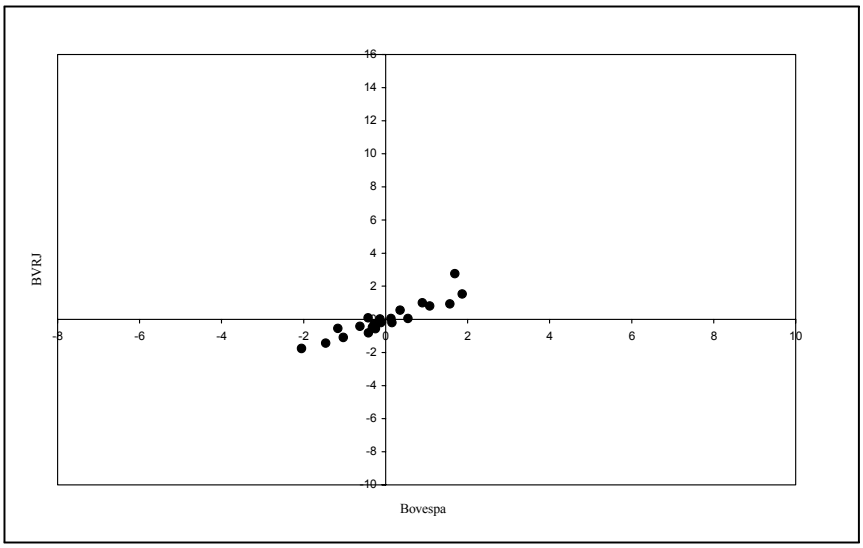


Figura 4.10: Variação diária nas Bolsas de Valores - dados padronizados

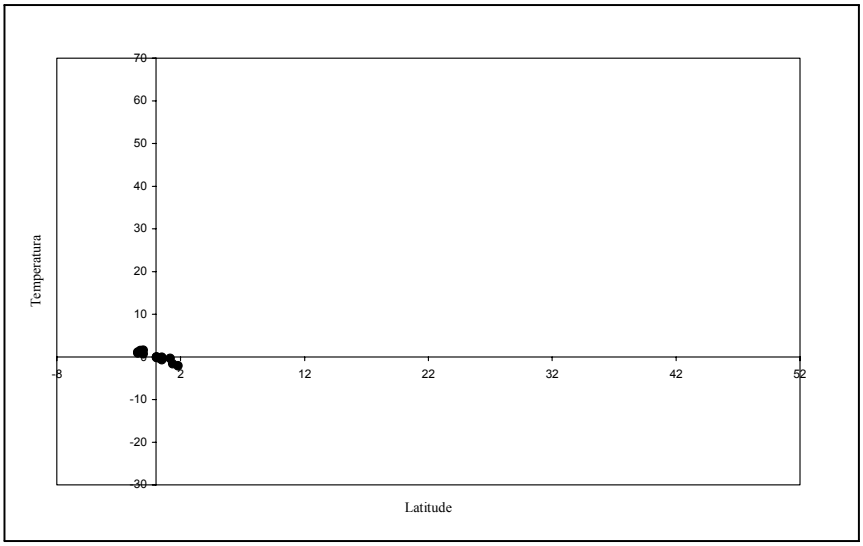


Figura 4.11: Latitude e temperatura média de 15 cidades dos EUA - dados padronizados

Define-se, então, o *coeficiente de correlação entre as variáveis* X *e* Y como sendo

$$\text{Corr}(X,Y) = \rho(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma_X}\right)\left(\frac{y_i - \overline{y}}{\sigma_Y}\right) = \frac{\text{Cov}(X,Y)}{\sigma_X\,\sigma_Y} \tag{4.6}$$

Para os dois conjuntos de dados das Figuras 4.8 e 4.9, o coeficiente de correlação é $0,9229$ e ambos têm o mesmo diagrama de dispersão, apresentado na Figura 4.10.

4.3.4 Propriedades da covariância e do coeficiente de correlação

Note que o coeficiente de correlação é adimensional! Além disso, ele tem uma propriedade bastante interessante, que é a seguinte:

$$-1 \leq \rho(X,Y) \leq 1 \tag{4.7}$$

Assim, valores do coeficiente de correlação próximos de 1 indicam uma forte associação linear crescente entre as variáveis, enquanto valores próximos de -1 indicam uma forte associação linear decrescente. Já valores próximos de zero indicam fraca associação linear (isso não significa que não exista algum outro tipo de associação; veja o caso da Figura 4.7).

Vamos ver agora o que acontece com a covariância e o coeficiente de correlação quando somamos uma constante aos dados e/ou multiplicamos os dados por uma constante. Vamos mostrar que

$$\text{Cov}(aX+b, cY+d) = ac\,\text{Cov}(X,Y) \tag{4.8}$$

e

$$\text{Corr}(aX+b, cY+d) = \frac{ac}{|ac|}\,\text{Corr}(X,Y) \tag{4.9}$$

De fato: fazendo $U = aX + b$ e $V = cY + d$, sabemos que $\overline{u} = a\overline{x} + b$ e $\overline{v} = c\overline{y} + d$ e $\sigma_U = |a|\,\sigma_X$ e

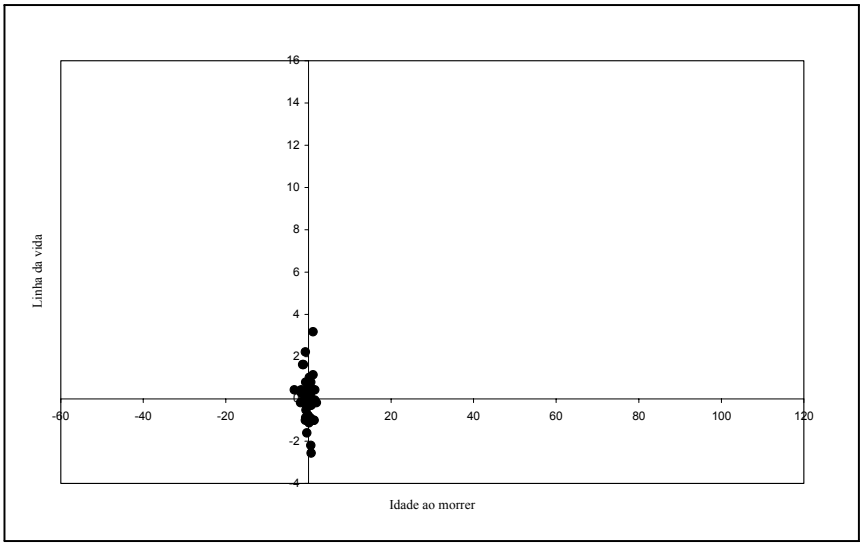


Figura 4.12: Diagrama de dispersão da idade ao morrer e comprimento da "linha da vida" - dados padronizados

$\sigma_V = |c|\,\sigma_Y$ Logo,

$$\begin{aligned}
\mathrm{Cov}(U,V) &= \frac{1}{n}\sum_{i=1}^{n}(u_i-\overline{u})(v_i-\overline{v}) = \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i+b-a\overline{x}-b)(cy_i+d-c\overline{y}-d) = \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i-a\overline{x})(cy_i-c\overline{y}) = \\
&= \frac{ac}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \\
&= ac\,\mathrm{Cov}(X,Y)
\end{aligned}$$

Para o coeficiente de correlação, temos que

$$\begin{aligned}
\mathrm{Corr}(aX+b,cY+d) &= \mathrm{Corr}(U,V) = \frac{\mathrm{Cov}(U,V)}{\sigma_U\sigma_V} = \\
&= \frac{ac\,\mathrm{Cov}(X,Y)}{|c|\,\sigma_X.\,|d|\,\sigma_Y} = \frac{ac}{|ac|}\,\mathrm{Corr}(X,Y)
\end{aligned}$$

Logo,

$$\mathrm{Corr}(aX+b,cY+d) = \begin{cases} \mathrm{Corr}(X,Y) & \text{se } ac>0 \\ -\mathrm{Corr}(X,Y) & \text{se } ac<0 \end{cases}.$$

4.3.5 Exemplo

Na Tabela 4.10 temos o consumo de cigarros per capita (X) em 1930 e as mortes por 1.000.000 habitantes em 1950, causadas por câncer de pulmão em 11 países.

O diagrama de dispersão para esses dados está na Figura 4.13 abaixo.

Tabela 4.10: Consumo de cigarros e morte por câncer de pulmão

País	X	Y	País	X	Y
Islândia	240	63	Holanda	490	250
Noruega	255	100	Suiça	180	180
Suécia	340	140	Finlândia	1125	360
Dinamarca	375	175	Grã-Bretanha	1150	470
Canadá	510	160	Estados Unidos	1275	200
Austrália	490	180			

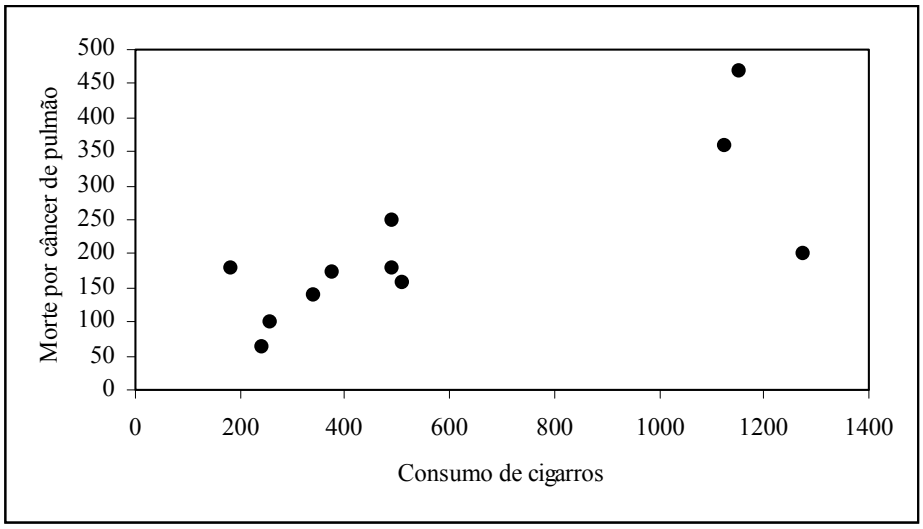


Figura 4.13: Conumo de cigarros e mortes por câncer de pulmão em 11 países

Na tabela a seguir temos os detalhes dos cálculos a serem feitos no caso de se estar utilizando uma calculadora mais simples.

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	240	63	57600	3969	15120
	255	100	65025	10000	25500
	340	140	115600	19600	47600
	375	175	140625	30625	65625
	510	160	260100	25600	81600
	490	180	240100	32400	88200
	490	250	240100	62500	122500
	180	180	32400	32400	32400
	1125	360	1265625	129600	405000
	1150	470	1322500	220900	540500
	1275	200	1625625	40000	255000
Soma	6430	2278	5365300	607594	1679045

A covariância de X e Y é a *média do produto menos o produto das médias*, ou seja:

$$\text{Cov}\,(X,Y) = \frac{1679045}{11} - \frac{6430}{11} \times \frac{2278}{11} = \frac{18469495 - 14647540}{121} = \frac{3821955}{121} = 31586,404959$$

A variância de cada variável é a *média dos quadrados menos o quadrado da média*, ou seja:

$$\text{Var}(X) = \frac{5365300}{11} - \left(\frac{6430}{11}\right)^2 = \frac{59018300 - 41344900}{121} = \frac{17673400}{121} = 146061,157025$$

$$\text{Var}(Y) = \frac{607594}{11} - \left(\frac{2278}{11}\right)^2 = \frac{6683534 - 5189284}{121} = \frac{1494250}{121} = 12349,173554$$

Os desvios padrão são:

$$\sigma_X = 382,179483 \qquad \sigma_Y = 111,126835$$

e, assim, o coeficiente de correlação é:

$$\rho(X,Y) = \frac{31586,404959}{382,179483 \times 111,1268354} = 0,743728$$

Essa correlação parece indicar que há um aumento no número de mortes por câncer do pulmão à medida que aumenta o número de cigarros consumidos.

Note como os cálculos foram feitos! Trabalhando com o denominador comum, reduz-se o número de divisões nos cálculos!

Exercício 4.2 *Considere os dados sobre a produção de ovos nos 50 estados dos Estados Unidos, da Tabela 4.11. Calcule o coeficiente de correlação entre o preço e a quantidade de ovos.*

Tabela 4.11: Produção de ovos nos Estados Unidos em 1990

Estado	Quant. (milhões)	Preço/dz (cents)	Estado	Quant. (milhões)	Preço/dz (cents)	Estado	Quant. (milhões)	Preço/dz. (cents)
AK	0,7	151,0	MA	235,0	105,0	OR	652,0	77,0
AL	2206,0	92,7	MD	885,0	76,6	PA	4976,0	61,0
AR	3620,0	86,3	ME	1069,0	101,0	RI	53,0	102,0
AZ	73,0	61,0	MI	1406,0	58,0	SC	1422,0	70,1
CA	7472,0	63,4	MN	2499,0	57,7	SD	435,0	48,0
CO	788,0	77,8	MO	1580,0	55,4	TN	277,0	71,0
CT	1029,0	106,0	MS	1434,0	87,8	TX	3317,0	76,7
DE	168,0	117,0	MT	172,0	68,0	UT	456,0	64,0
FL	2586,0	62,0	NC	3033,0	82,8	VA	943,0	86,3
GA	4302,0	80,6	ND	51,0	55,2	VT	31,0	106,0
HI	227,5	85,0	NE	1202,0	50,3	WA	1287,0	74,1
IA	2151,0	56,5	NH	43,0	109,0	WI	910,0	60,1
ID	187,0	79,1	NJ	442,0	85,0	WV	136,0	104,0
IL	793,0	65,0	NM	283,0	74,0	WY	1,7	83,0
IN	5445,0	62,7	NV	2,2	53,9			
KS	404,0	54,5	NY	975,0	68,1			
KY	412,0	67,7	OH	4667,0	59,1			
LA	273,0	115,0	OK	869,0	101,0			

Fonte: Gujarati (1995) Basic Econometrics - McGraw-Hill, 3ª ed. - Tabela 1.1

4.4 Exercícios complementares

Exercício 4.3 *Em uma pesquisa realizada em uma cidade, entrevistou-se uma amostra de moradores. Dentre as variáveis pesquisadas estava a classe de renda e o jornal preferido, dentre os três maiores da cidade. Os dados constam da tabela abaixo. Utilize um gráfico para ilustrar esses dados, levando em conta o tipo de relaçao entre as variáveis. Quantas pessoas há na classe alta? Quantas pessoas lêem o jornal A? Dentre as pessoas da classe pobre, qual é o prcentual de leitores do jornal C?*

Jornal	*Classe social*			
	Pobre	*Média inferior*	*Média*	*Alta*
A	*15*	*27*	*44*	*22*
B	*20*	*27*	*26*	*11*
C	*13*	*18*	*14*	*3*

Exercício 4.4 *Calcule o coeficiente de correlação entre o preço de venda e a área das casas, cujos dados encontram-se na Tabela 4.12.*

Tabela 4.12: Vendas de casas em Boulder, Colorado (1995)

Preço (Y) (1000 US$)	Área (X) (m^2)	Preço (Y) (1000 US$)	Área (X) (m^2)	Preço (Y) (1000 US$)	Área (X) (m^2)
113	126	163	227	186	228
114	158	168	228	187	219
120	126	168	249	187	222
120	126	169	244	188	279
122	158	169	263	188	249
123	126	170	234	190	317
129	229	171	283	192	304
137	196	172	286	193	195
140	262	173	268	195	217
142	272	175	223	195	232
143	189	175	270	200	234
146	158	175	231	200	322
146	218	176	249	200	304
148	276	177	285	207	300
149	218	178	243	270	252
152	302	178	251	290	322
153	168	180	279	300	353
157	302	180	189	320	349
157	289	181	153	328	388
160	277	185	316		

Exercício 4.5 *Os 4 conjuntos de dados apresentados na Tabela 4.13 constam de. Anscombe(1973). Para cada um deles construa o diagrama de dispersão e calcule a média, o desvio padrão e o coeficiente de correlação. Comente os resultados obtidos.*

Exercício 4.6 *Muitas vezes a determinação da capacidade de produção instalada para certo tipo de indústria em certos tipos de localidades é um processo difícil e custoso. Como alternativa, pode-se*

Tabela 4.13: Dados de Anscombe para o coeficiente de correlação

Conjunto 1		Conjunto 2		Conjunto 3		Conjunto 4	
X	Y	X	Y	X	Y	X	Y
10,0	9,14	8,0	6,58	10	8,04	10,0	7,46
8,0	8,14	8,0	5,76	8	6,95	8,0	6,77
13,0	8,74	8,0	7,71	13	7,58	13,0	12,74
9,0	8,77	8,0	8,84	9	8,81	9,0	7,11
11,0	9,26	8,0	8,47	11	8,33	11,0	7,81
14,0	8,10	8,0	7,04	14	9,96	14,0	8,84
6,0	6,13	8,0	5,25	6	7,24	6,0	6,08
4,0	3,10	19,0	12,50	4	4,26	4,0	5,39
12,0	9,13	8,0	5,56	12	10,84	12,0	8,15
7,0	7,26	8,0	7,91	7	4,82	7,0	6,42
5,0	4,74	8,0	6,89	5	5,68	5,0	5,73

estimar a capacidade de produção através de uma outra variável de medida mais fácil, que esteja linearmente relacionada com ela. Suponha que foram observados os valores, dados na Tabela 4.14, para as variáveis capacidade de produção instalada, potência instalada e área construída. Com base num critério estatístico, qual das variáveis você escolheria para estimar a capacidade de produção instalada?

Tabela 4.14: Dados de capacidade da produção instalada

X: capacidade de produção instalada (ton)									
4	5	4	5	8	9	10	11	12	12
Y: potência instalada (1000 kW)									
1	1	2	3	3	5	5	6	6	6
Z: área construída (100 m^2)									
6	7	10	10	11	9	12	10	11	14

Para facilitar a solução do exercício, você tem os seguintes resultados:

$$\begin{aligned} \sum X_i &= 80 \qquad \sum X_i^2 = 736 \\ \sum Y_i &= 38 \qquad \sum Y_i^2 = 182 \\ \sum Z_i &= 100 \qquad \sum Z_i^2 = 1048 \\ \sum X_i Y_i &= 361 \qquad \sum X_i Z_i = 848 \qquad \sum Y_i Z_i = 411 \end{aligned}$$

Exercício 4.7 *Na Tabela 4.15 são apresentados dados de despesas com alimentação e renda para um conjunto de 40 domicílios. Qual o sinal esperado para o coeficiente de correlação? Você espera que ele seja muito grande? Calcule o coeficiente de correlação.*

Tabela 4.15: Despesas com alimentação e renda

	Despesas com Alimentação	Renda Mensal		Despesas com Alimentação	Renda Mensal
1	52,25	258,3	21	98,14	719,80
2	58,32	343,1	22	123,94	720,00
3	81,79	425,00	23	126,31	722,30
4	119,9	467,50	24	146,47	722,30
5	125,8	482,90	25	115,98	734,40
6	100,46	487,70	26	207,23	742,50
7	121,51	496,50	27	119,80	747,70
8	100,08	519,40	28	151,33	763,30
9	127,75	543,30	29	169,51	810,20
10	104,94	548,70	30	108,03	818,50
11	107,48	564,60	31	168,90	825,60
12	98,48	588,30	32	227,11	833,30
13	181,21	591,30	33	84,94	834,00
14	122,23	607,30	34	98,70	918,10
15	129,57	611,20	35	141,06	918,10
16	92,84	631,00	36	215,40	929,60
17	117,92	659,60	37	112,89	951,70
18	82,13	664,00	38	166,25	1014,00
19	182,28	704,20	39	115,43	1141,30
20	139,13	704,80	40	269,03	1154,60

.

.

.

.

.

PÁGINA EM BRANCO

Capítulo 5

Solução dos Exercícios

5.1 Capítulo 1

1. É possível identificar as seguintes variáveis:

   - Condição do domicílio - variável qualitativa
   - Condição da rua - variável qualitativa
   - Tipo de imóvel - variável qualitativa
   - Renda - pode ser qualitativa se for perguntada a faixa ou quantitativa, se for perguntada a renda exata; a primeira opção é a mais provável para esse tipo de pesquisa.
   - Classificação econômica - variável qualitativa
   - Número de pessoas - variável quantitativa
   - Presença de crianças - variável qualitativa
   - Número de crianças - variável quantitativa
   - Presença de adolescentes - variável qualitativa
   - Número de adolescentes - variável quantitativa
   - Idade do chefe e da dona-de-casa - pode ser quantitativa, caso se pergunte a idade exata, ou qualitativa, caso se identifique a faixa etária
   - Grau de instrução do chefe e da dona-de-casa - variável qualitativa
   - Condição de atividade do chefe - variável qualitativa
   - Presença de geladeira, máquina de lavar, etc - variáveis qualitativas do tipo Sim/Não
   - Acesso a serviços de mídia - variáveis qualitativas do tipo Sim/Não

2. A distribuição é dada na tabela a seguir.

Sexo	Freqüência Simples	
	Absoluta	Relativa %
Masculino	14	60,87
Feminino	9	39,13
Total	23	100,00

3. Veja a Figura 5.1.

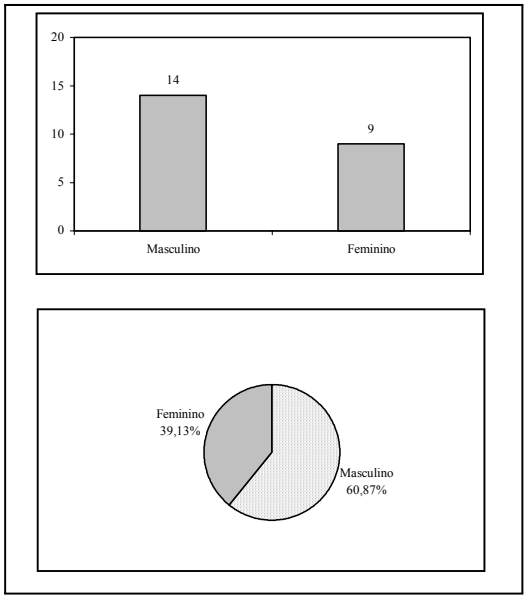


Figura 5.1: Solução do Exercício 3

Tabela 5.1: Distribuição do número de dependentes dos funcionários do Depto de RH

Número de	Freqüência Simples		Freqüência Acumulada	
dependentes	Absoluta	Relativa %	Absoluta	Relativa %
0	5	33,33	5	33,33
1	3	20,00	8	53,33
2	3	20,00	11	73,33
3	3	20,00	14	93,33
4	1	6,67	15	100,00
Total	15	100,00		

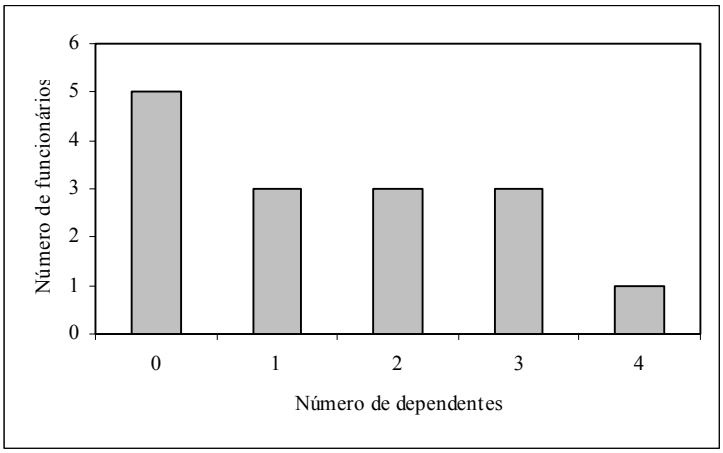


Figura 5.2: Distribuição do número de dependentes dos funcionários do Depto de RH

4. Veja Tabela 5.1

5. Veja Figura 5.2.

6. Veja a Tabela 5.2.

Tabela 5.2: Distribuição de freqüência da idade dos funcionários do Depto de RH

Faixa	Freqüência Simples		Freqüência Acumulada	
Etária	Absoluta	Relativa %	Absoluta	Relativa %
$24-28$	4	26,67	4	26,67
$29-33$	3	20,00	7	46,67
$34-38$	4	26,67	11	73,33
$39-43$	1	6,67	12	80,00
$44-48$	1	6,67	13	86,67
$49-53$	2	13,33	15	100,0
Total	15	100,00		

7. O valor mínimo é 3200 e o valor máximo é 7300. Dessa forma, a amplitude exata é $7300-3200 = 4100$ e o próximo múltiplo de 5 é 4105. Logo, o comprimento de cada classe é $\frac{4105}{5} = 821$. Obtém-se a Tabela 5.3.

Tabela 5.3: Distribuição dos salários dos funcionários do Depto de RH

Faixa	Freqüência Simples		Freqüência Acumulada	
salarial	Absoluta	Relativa %	Absoluta	Relativa %
$3200 \vdash 4021$	4	26,67	4	26,67
$4021 \vdash 4842$	2	13,33	6	40,00
$4842 \vdash 5663$	2	13,33	8	53,33
$5663 \vdash 6484$	3	20,00	11	73,33
$6484 \vdash 7305$	4	26,67	15	100,00
Total	15	100,00		

8. Veja a Tabela 5.4 e os gráficos nas Figuras **??** e 5.4.

Notas	Freqüência Simples		Freqüência Acumulada	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)
$2 \vdash 3$	1	2,0	1	2,0
$3 \vdash 4$	2	4,0	3	6,0
$4 \vdash 5$	2	4,0	5	10,0
$5 \vdash 6$	3	6,0	8	16,0
$6 \vdash 7$	12	24,0	20	40,0
$7 \vdash 8$	14	28,0	34	68,0
$8 \vdash 9$	12	24,0	46	92,0
$9 \vdash 10$	4	8,0	50	100,0
Total	50	100,0		

Tabela 5.4: Distribuição de freqüência das notas de 50 alunos

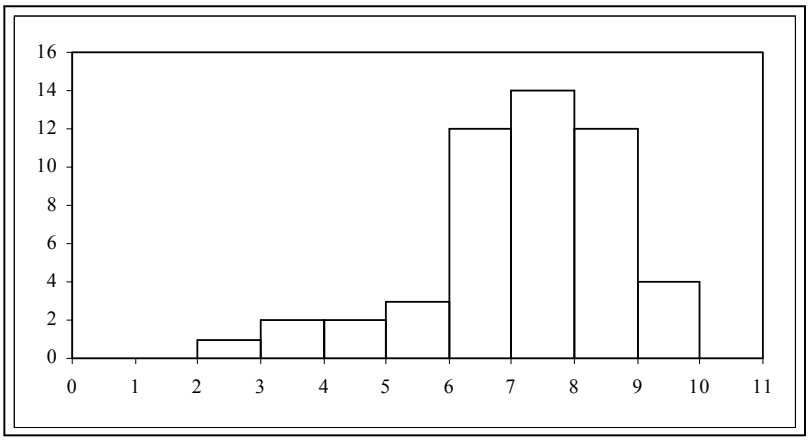


Figura 5.3: Histograma da distribuição das notas de 50 alunos

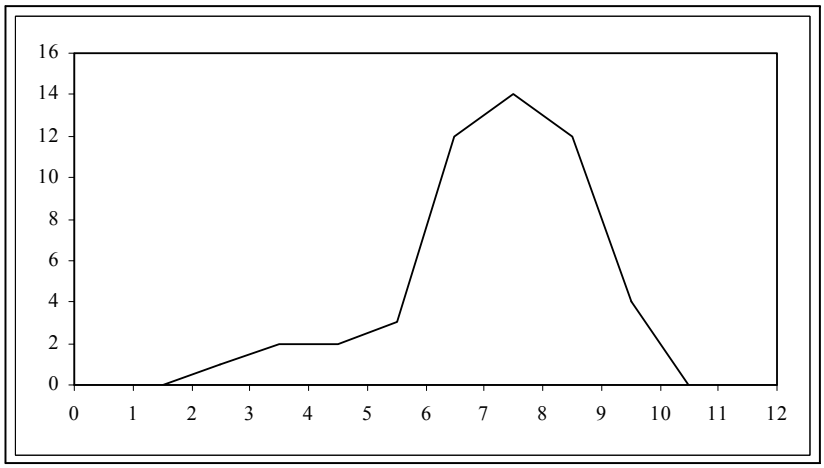


Figura 5.4: Polígono de freqüência da distribuição das notas de 50 alunos

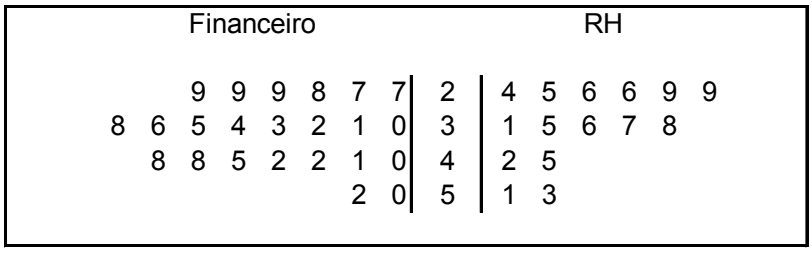


Figura 5.5: Idades de Funcionários de Dois Departamentos

9. Veja a Figura 5.5.

10. Variáveis qualitativas: Sexo e matéria predileta

    Variável quantitativa discreta: nota - némero de questões certas

    Veja as Figuras 5.6, 5.7, 5.8 com as tabelas e gráficos para essas variáveis.

Freqüência Simples	
Absoluta	Relativa %
21	50
21	50
42	100

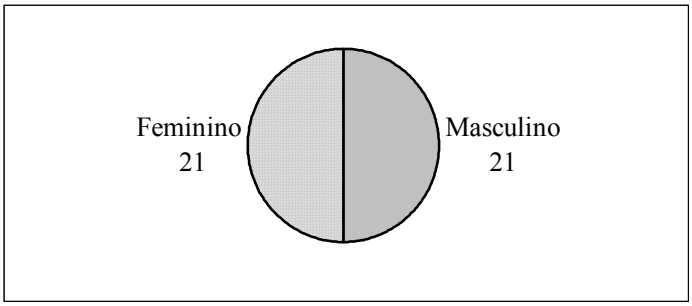


Figura 5.6: Distribuição dos alunos de Economia por sexo

Matéria Predileta	Freqüência Simples	
	Absoluta	Relativa %
Ciências	3	7,14
Geografia	8	19,05
História	7	16,67
Matemática	14	33,33
Português	10	23,81
Total	42	100,00

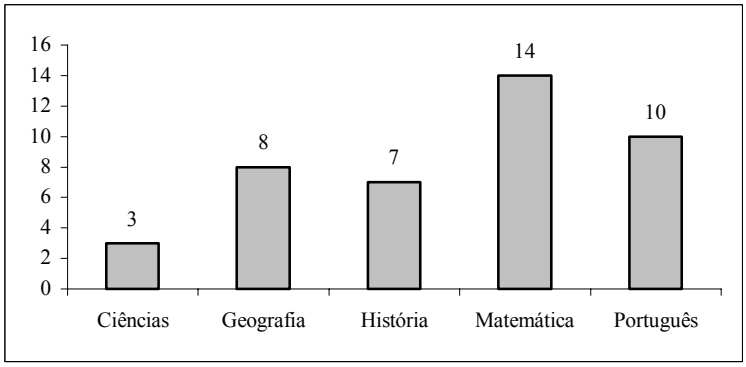


Figura 5.7: Distribuição dos alunos de Economia por matéria predileta

11. Veja a Figura 5.9.

12. No exercício são dadas as freqüências acumuladas simples, que vamos representar pela letra F. Para obtermos as freqüências absolutas simples, que vamos representar pela letra f, devemos notar o seguinte: para o menor valor (zero), a freqüência simples é igual à acumulada, ou seja:

$$f_1 = F_1 = 2913$$

Para o segundo valor, temos:

$$f_1 + f_2 = F_2 \Rightarrow f_2 = F_2 - F_1 \Rightarrow f_2 = 4500 - 2913 = 1587$$

Nota	Freqüência Simples		Freqüência Acumulada	
	Absoluta	Relativa %	Absoluta	Relativa %
1	1	2,38	1	2,38
2	2	4,76	3	7,14
3	1	2,38	4	9,52
4	3	7,14	7	16,67
5	11	26,19	18	42,86
6	7	16,67	25	59,52
7	5	11,90	30	71,43
8	8	19,05	38	90,48
9	4	9,52	42	100,00
Total	42	100,00		

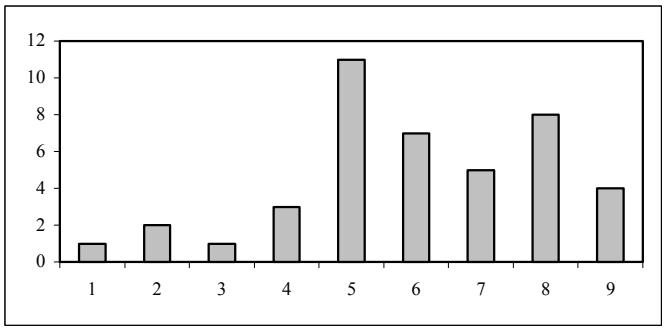


Figura 5.8: Distribuiçãod as notas dos alunos de Economia

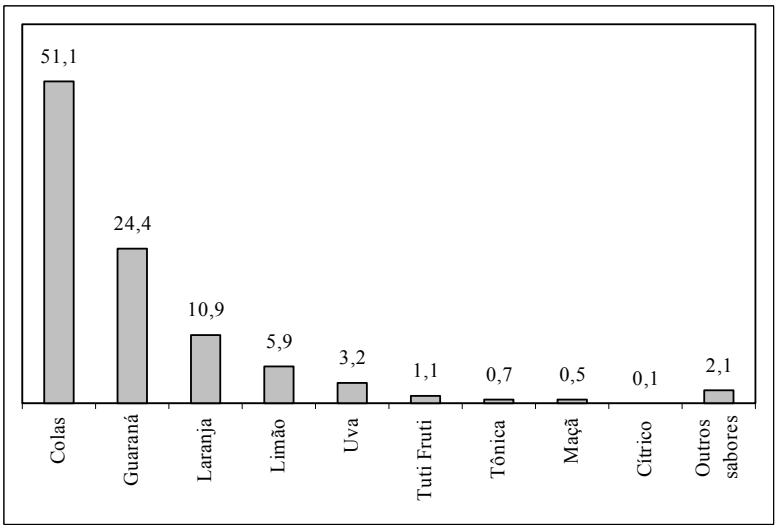


Figura 5.9: Distribuição da preferência de sabor de refrigerantes

Para o terceiro valor, temos:

$$f_1 + f_2 + f_3 = F_3 \Rightarrow F_2 + f_3 = F_3 \Rightarrow f_3 = F_3 - F_2 \Rightarrow f_3 = 4826 - 4500 = 326$$

De forma anloga, obtemos que

$$f_4 = F_4 - F_3 = 4928 - 4826 = 102 f_5 = F_5 - F_4 = 5000 - 4928 = 72$$

Obtemos, então, a seguinte tabela:

Número de sinistros	Número de apólices			
	Freqüência Simples		Freqüência Acumulada	
	Absoluta	Relativa	Absoluta	Relativa
0	2913	58,26	2913	58,26
1	1587	31,74	4500	90,00
2	326	6,52	4826	96,52
3	102	2,04	4928	98,56
4	72	1,44	5000	100,00
Total	5000	100,00		

13. O gráfico apresentado na Figura **??** é um gráfico de colunas. Havendo disponibilidade de se usar o recurso de cores, é possível usar o o gráfico de setores também.

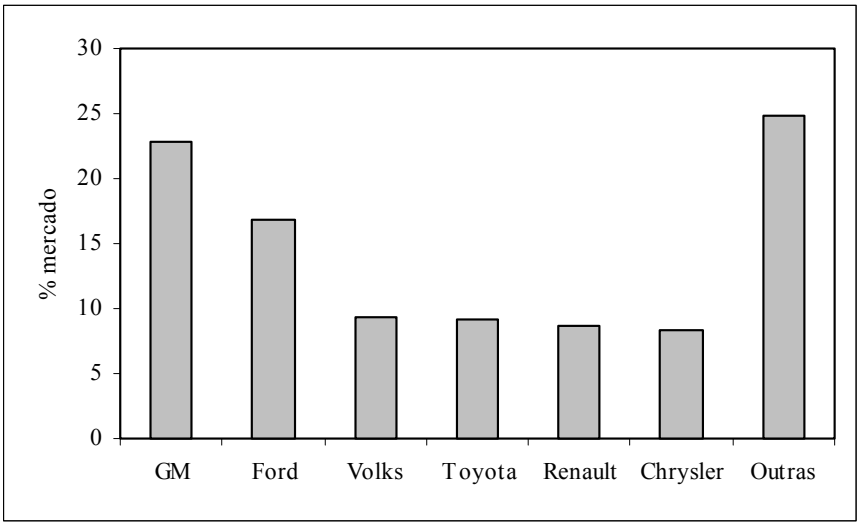


Figura 5.10: Concentração do mercado automobilístico

14. Os valores mínimo e máximo são, respectivamente, 1815 e 118800, o que fornece uma amplitude exata de 116985. Tomando o próximo múltiplo de 5, a maplitude efetiva passa a ser 116990, o que dá um comprimento de classe $\frac{116990}{5} = 23398$. Veja a Tabela **??** e os gráficos na Figura 5.11.

15. Veja a Tabela 5.6.

16. Veja a Figura 5.12.

17. Veja a Figura 5.13.

Número de	Freqüência Simples		Freqüência Acumulada	
Horas Trabalhadas	Absoluta	Relativa (%)	Absoluta	Relativa (%)
1815 ⊢ 25213	63	63,0	63	63,0
25213 ⊢ 48611	17	17,0	80	80,0
48611 ⊢ 72009	9	9,0	89	89,0
72009 ⊢ 95407	8	8,0	97	97,0
95407 ⊢ 118805	3	3,0	100	100,0
Total	100	100,0		

Tabela 5.5: Distribuição de freqüência do número de horas trabalhadas

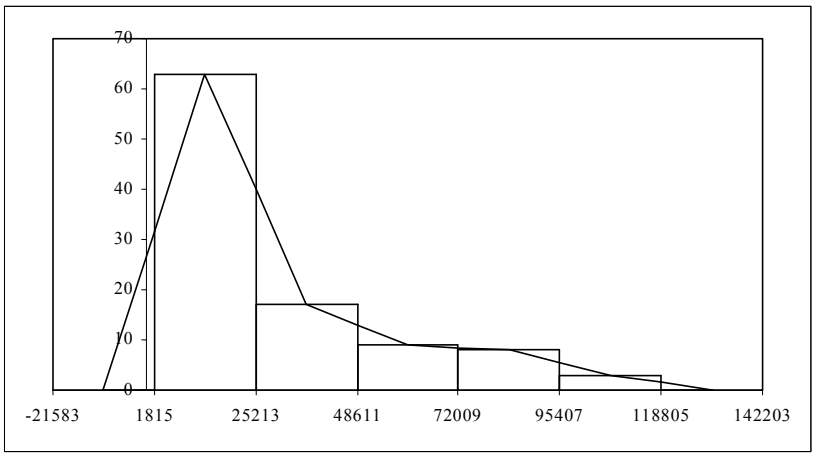


Figura 5.11: Distribuição da jornada de trabalho de empregados de empresas alimentares

População	Freqüência Simples		Freqüência Acumulada	
(em 1000 hab	Absoluta	Relativa (%)	Absoluta	Relativa (%)
50 ⊢ 60	7	11,67	7	11,67
60 ⊢ 70	12	20,00	19	31,67
70 ⊢ 80	11	18,33	30	50,00
80 ⊢ 90	3	5,00	33	55,00
90 ⊢ 100	4	6,67	37	61,67
100 ⊢ 200	13	21,67	50	83,33
200 ⊢ 500	7	11,67	57	95,00
500 ou mais	3	5,00	60	100,00
Total	60	100,0		

Tabela 5.6: População dos municípios mineiros com mais de 50000 habitantes

0	2 2 3 4 4 5 5 6 6
1	2 5 7
2	4
3	1 7
4	8
5	1 3 7
6	1 8
7	
8	1 1
9	
10	2
11	
12	
13	
14	9
32	8
35	3

Figura 5.12: Densidade populacional das UFs brasileiras

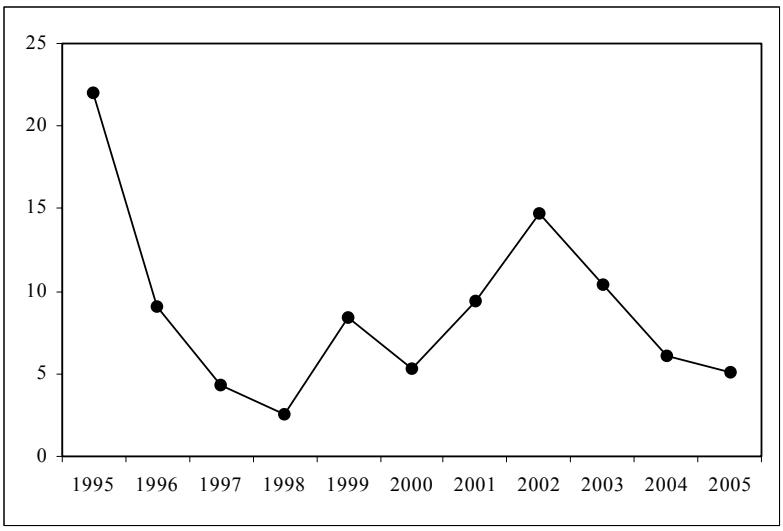


Figura 5.13: Inflação brasileira anual - INPC

5.2 Capítulo 2

1. Temos 15 funcionários. Os dados ordenados são os seguintes: 3200, 3780, 3800, 4000, 4500, 4500, 5100, 5600, 5700, 6300, 6400, 6500, 7000, 7100, 7300. A média é

$$\overline{x} = \frac{3200 + 3780 + \cdots + 7300}{15} = \frac{80700}{15} = 5380$$

A moda é

$$x^* = 4500$$

e a mediana é a observação de posição $\frac{15+1}{2} = 8$, ou seja,

$$Q_2 = x_{(8)} = 5600$$

Todas essas medidas estão em R$.

2. Note que os dados já estão ordenados; caso não estivessem, uma boa opção para ajudar na solução do exercício seria construir o diagrama de ramos e folhas. Temos 50 notas. Logo,

$$\overline{x} = \frac{2,9 + 3,7 + \cdots + 9,7}{50} = \frac{357,1}{50} = 7,142$$

A nota modal é $x^* = 6,3$, que aparece 3 vezes. Como o número de observações é par ($n = 50$), a mediana é a média das 2 observações centrais, cujas posições são $\frac{50}{2}$ e $\frac{50}{2} + 1$, ou seja, a mediana é a média da 25^a e da 26^a observações:

$$Q_2 = \frac{7,3 + 7,4}{2} = 7,35$$

3. Temos o seguinte:

$$\sum_{i=1}^{6} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 10 + 11 + 15 + 19 + 21 + 26 = 102$$

$$\sum_{i=1}^{6} f_i = f_1 + f_2 + f_3 + f_4 + f_5 + f_6 = 3 + 5 + 9 + 10 + 2 + 1 = 30$$

$$\begin{aligned} \sum_{i=1}^{6} f_i x_i &= f_1x_1 + f_2x_2 + f_3x_3 + f_4x_4 + f_5x_5 + f_6x_6 = \\ &= 3 \times 10 + 5 \times 11 + 9 \times 15 + 10 \times 19 + 2 \times 21 + 1 \times 26 = \\ &= 478 \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{6} f_i x_i^2 &= f_1x_1^2 + f_2x_2^2 + f_3x_3^2 + f_4x_4^2 + f_5x_5^2 + f_6x_6^2 = \\ &= 3 \times 10^2 + 5 \times 11^2 + 9 \times 15^2 + 10 \times 19^2 + 2 \times 21^2 + 1 \times 26^2 = \\ &= 8098 \end{aligned}$$

4. Vamos denotar por x_1 e x_2 as notas na primeira e segunda provas. Então, a média final é calculada como

$$\overline{x}_p = \frac{2x_1 + 3x_2}{2+3}$$

Para aprovação direta, sem prova final, temos que ter $\overline{x}_p \geq 6$. Logo,

$$\overline{x}_p \geq 6 \Leftrightarrow \frac{2x_1 + 3x_2}{2+3} \geq 6 \Leftrightarrow 2 \times 5,5 + 3x_2 \geq 30 \Leftrightarrow 3x_2 \geq 19 \Leftrightarrow x_2 \geq \frac{19}{3} = 6,3\overline{3}$$

Se fosse média simples, teríamos que ter

$$\overline{x} \geq 6 \Leftrightarrow \frac{x_1 + x_2}{2} \geq 6 \Leftrightarrow 5,5 + x_2 \geq 12 \Leftrightarrow x_2 \geq 6,5$$

5. A mesma relação que se aplica às temperaturas individuais se aplica também à temperatura média, ou seja, a temperatura média em graus Celsius é

$$\overline{C} = \frac{5}{9}(\overline{F} - 32) = \frac{5}{9}(45 - 32) = 7,22°C$$

6. Não é necessário recalcular a média em milhares de reais; basta dividir a média por 1000, ou seja, o lucro médio é de 1035,42 milhares de reais.

7. A média dos dados é $\overline{x} = \frac{49}{8} = 6,125$.

$$\begin{aligned} & \sum_{i=1}^{8}(x_i - \overline{x}) \\ = \; & (2 - 6,125) + (4 - 6,125) + (5 - 6,125) + (6 - 6,125) + (7 - 6,125) \\ & +2 \times (8 - 6,125) + (9 - 6,125) \\ = \; & -4,125 - 2,125 - 1,125 - 0,125 + 0,875 + 2 \times 1,875 + 2,875 \\ = \; & -7,5 + 7,5 = 0 \end{aligned}$$

$$\begin{aligned} DMA &= \frac{1}{8}\sum_{i=1}^{8}|x_i - \overline{x}| \\ &= \frac{1}{8}\left[\begin{array}{c} |2 - 6,125| + |4 - 6,125| + |5 - 6,125| + |6 - 6,125| + |7 - 6,125| + \\ 2 \times |8 - 6,125| + |9 - 6,125| \end{array}\right] \\ &= \frac{1}{8}(4,125 + 2,125 + 1,125 + 0,125 + 0,875 + 2 \times 1,875 + 2,875) \\ &= \frac{1}{8}(7,5 + 7,5) = 1,875 \end{aligned}$$

8. .

$$\begin{aligned} \sigma^2 &= \frac{1}{8}\left[\begin{array}{c} (2 - 6,125)^2 + (4 - 6,125)^2 + (5 - 6,125)^2 + (6 - 6,125)^2 + (7 - 6,125)^2 + \\ 2 \times (8 - 6,125)^2 + (9 - 6,125)^2 \end{array}\right] \\ &= \frac{1}{8}\left(4,125^2 + 2,125^2 + 1,125^2 + 0,125^2 + 0,875^2 + 2 \times 1,875^2 + 2,875^2\right) \\ &= \frac{38,875}{8} = 4,859375 \end{aligned}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{4.859375} = 2,204399$$

9. .

$$\sigma^2 = \frac{2^2+4^2+5^2+6^2+7^2+8^2+8^2+9^2}{8} - 6,125^2 = \frac{339}{8} - 37,515625 = 4,859375$$

10. Note que podemos escrever

$$C = \frac{5}{9}F - \frac{160}{9}$$

Como visto, somar uma constante aos dados não altera o desvio padrão; logo, o termo $-\frac{160}{9}$ não tem influência sobre o resultado. Mas quando multiplicamos por uma constante, o desvio padrão fica multiplicado pelo módulo da constante. Logo,

$$\sigma_C = \frac{5}{9}\sigma_F \Rightarrow \sigma_C = \frac{5}{9} \times 5,2^\circ F = 2,8889^\circ F$$

11. A amplitude é $\Delta = 52 - 27 = 25$. Pela fórmula simplificada, a variância é

$$\begin{aligned}\sigma^2 &= \frac{2\times 27^2 + 28 + 2\times 29^2 + \cdots + 50^2 + 52^2}{23} - \left(\frac{846}{23}\right)^2 \\ &= \frac{32506}{23} - \left(\frac{846}{23}\right)^2 = \frac{747638 - 715716}{23^2} = 60,344 \Rightarrow \sigma = 7,7681\end{aligned}$$

$$\begin{aligned}DMA &= \frac{2\times\left|27-\frac{846}{23}\right| + \left|28-\frac{846}{23}\right| + \cdots + \left|52-\frac{846}{23}\right|}{23} \\ &= \frac{156,3478261}{23} = 6,7977\end{aligned}$$

$$\begin{aligned}Q_1 &= x_{(6)} = 29 \\ Q_3 &= x_{(12+6)} = x_{(18)} = 42 \\ IQ &= 42 - 29 = 13\end{aligned}$$

12. A distribuição de freqüências completa é a seguinte:

Classes	Ponto	Freq. Simples		Freq. Acumulada	
	Médio	Absoluta	Relativa	Absoluta	Relativa
$4 \vdash 6$	5	10	0,20	10	0,20
$6 \vdash 8$	7	12	0,24	22	0,44
$8 \vdash 10$	9	18	0,36	40	0,80
$10 \vdash 12$	11	6	0,12	46	0,92
$12 \vdash 14$	13	4	0,08	50	1,00
Total		50	1,00		

A média é

$$\overline{x} = 5\times 0,20 + 7\times 0,24 + 9\times 0,36 + 11\times 0,12 + 13\times 0,08 = 8,28$$

O desvio médio absoluto é

$$\begin{aligned}DMA &= 0,20\times|5-8,28| + 0,24\times|7-8,28| + 0,36\times|9-8,28| \\ &\quad +0,12\times|11-8,28| + 0,08\times|13-8,28| \\ &= 1,9264\end{aligned}$$

Usando a fórmula alternativa, temos que

$$\begin{aligned}\sigma^2 &= 0,20 \times 5^2 + 0,24 \times 7^2 + 0,36 \times 9^2 + 0,12 \times 11^2 + 0,08 \times 13^2 - 8,28^2 \\ &= 73,96 - 68,5584 = 5,4016\end{aligned}$$

13. A mediana está na classe $8 \vdash 10$. Abaixo desta classe temos 44% das observações. Assim, para completar 50% ficam faltando 6% - veja a Figura 5.14 a seguir. A regra de proporcionalidade é

$$\frac{Q_2 - 8}{6} = \frac{10 - 8}{36} \Rightarrow Q_2 - 8 = \frac{12}{36} \Rightarrow Q_2 = 8,33$$

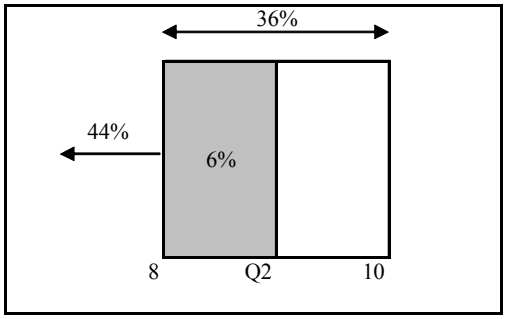


Figura 5.14: Solução do Exercício 12 - Cálculo da mediana

14. A mediana está na segunda classe e na primeira classe temos $(101/407) \times 100 = 24,816\%$ dos dados. Logo, faltam $50 - 24,816 = 25,184\%$ e isso nos leva à seguinte regra de três:

$$\frac{Q_2 - 20}{25,184} = \frac{30 - 20}{30,713} \Rightarrow Q_2 = 28,1998$$

Note que 30,713 é a freqüência relativa da segunda classe!

O primeiro quartil também está na segunda classe e na primeira classe temos $(101/407) \times 100 = 24,816\%$ dos dados. Logo, faltam $25 - 24,816 = 0,184\%$ e isso nos leva à seguinte regra de três:

$$\frac{Q_2 - 20}{0,184} = \frac{30 - 20}{30,713} \Rightarrow Q_2 = 20,0599$$

O terceiro quartil está na terceira classe e nas duas primeiras classes temos $(226/407) \times 100 = 55,528\%$ dos dados. Logo, faltam $75 - 55,528 = 19,472\%$ e isso nos leva à seguinte regra de três:

$$\frac{Q_3 - 30}{19,472} = \frac{40 - 30}{23,3415} \Rightarrow Q_3 = 38,3422$$

Note que 23,3415 é a freqüência relativa da terceira classe!

15. A moda está na classe classe: $20 \vdash 30$, cuja freqüência absoluta é 125; as freqüências das classes vizinhas inferior e superior são 101 e 95, respectivamente.

$$\text{King:} \qquad \frac{x^* - 20}{30 - x^*} = \frac{95}{101} \Rightarrow x^* = 24,8469$$

$$\text{Czuber:} \qquad \frac{x^* - 20}{30 - x^*} = \frac{125 - 101}{125 - 95} \Rightarrow x^* = 24,4444$$

16. Os pontos médios das classes são 14, 18, 22, 26, 30. Logo,

$$\overline{x} = \frac{14 \times 10 + 18 \times 18 + 22 \times 29 + 26 \times 10 + 30 \times 3}{70} = 20,743$$

$$\sigma^2 = \frac{14^2 \times 10 + 18^2 \times 18 + 22^2 \times 29 + 26^2 \times 10 + 30^2 \times 3}{70} - 20,743^2 = 16,699$$

A classe modal é $20 \vdash 24$:

$$\text{King} \quad : \quad \frac{x^* - 20}{24 - x^*} = \frac{10}{18} \Rightarrow 18x^* - 360 = 240 - 10x^* \Rightarrow x^* = 21,4286$$

$$\text{Czuber} \quad : \quad \frac{x^* - 20}{24 - x^*} = \frac{29 - 18}{29 - 10} \Rightarrow 19x^* - 380 = 264 - 11x^* \Rightarrow x^* = 21,4667$$

As freqüências relativas simples e acumuladas são exibidas na tabela a seguir:

	Freqüência relativa	
	Simples	Acumulada
$12 \vdash 16$	0,14286	0,14286
$16 \vdash 20$	0,25714	0,40000
$20 \vdash 24$	0,41429	0,81429
$24 \vdash 28$	0,14286	0,95714
$28 \vdash 32$	0,04286	1,00000

O primeiro quartil está na segunda classe; a mediana e o terceiro quartil estão ambos na terceira classe:

$$\frac{Q_1 - 16}{20 - 16} = \frac{0.25 - 0.14286}{0.25714} \Rightarrow Q_1 = 17,6667$$

$$\frac{Q_2 - 20}{24 - 20} = \frac{0.50 - 0.40}{0.41429} \Rightarrow Q_2 = 20,9655$$

$$\frac{Q_3 - 20}{24 - 20} = \frac{0.75 - 0.40}{0.41429} \Rightarrow Q_1 = 23,3793$$

17. Para esses dados temos

$$\begin{aligned} Q_2 &= \frac{x_{(25)} + x_{(26)}}{2} = 7,35 \\ Q_1 &= x_{(13)} = 6,3 \\ Q_3 &= x_{(38)} = 8,2 \end{aligned}$$

Logo

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(8,2 - 7,35) - (7,35 - 63)}{8,2 - 6,3} = -0,1053$$

18. Os quartis para esse conjunto de dados são $Q_2 = x_{(8)} = 5600; Q_1 = x_{(4)} = 4000; Q_3 = x_{(12)} = 6500$. O intervalo interquartil é $Q_3 - Q_1 = 6500 - 4000 = 2500$ e a regra para *outliers* é

$$\begin{aligned} x &< Q_1 - 1,5IQ = 4000 - 1,5 \times 2500 = 250 \\ x &> Q_3 + 1,5IQ = 6500 + 1,5 \times 2500 = 10250 \end{aligned}$$

Como o menor salário é 3200 e o maior salário é 7300, não há salários discrepantes. O *boxplot* é dado na Figura 5.15.

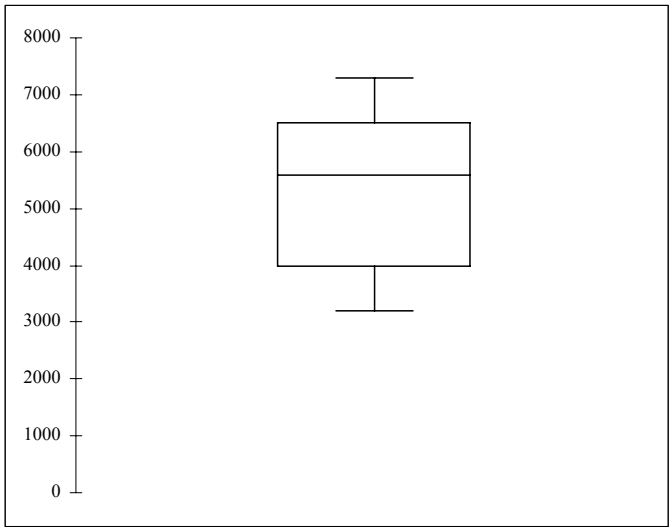


Figura 5.15: Solução do Exercício 16

19. Para calcular o salário horário médio, temos que dividir o total dos vencimentos pelo total de horas trabalhadas pelos 4 amigos.

$$\begin{aligned}\overline{x} &= \frac{10\times 3,50 + 12\times 2,6 + 15\times 3,80 + 8\times 2,20}{10+12+15+8}\\ &= \frac{10\times 3,50 + 12\times 2,6 + 15\times 3,80 + 8\times 2,20}{45}\\ &= \frac{10}{45}\times 3,50 + \frac{12}{45}\times 2,6 + \frac{15}{45}\times 3,80 + \frac{8}{45}\times 2,20\\ &= \frac{140,8}{45} = 3,1289\end{aligned}$$

Note que o selário médio é uma média ponderada dos salários individuais, com o peso sendo definido pelo número de horas de trabalho.

20. A carga horária semanal total é $4+4+4+6+2=20$. Logo, o CR do aluno é

$$CR = \frac{4}{20}\times 7,5 + \frac{4}{20}\times 6,1 + \frac{4}{20}\times 8,3 + \frac{6}{20}\times 6,5 + \frac{2}{20}\times 7,5 = \frac{141,6}{20} = 7,08$$

21. O diagrama de ramos e folhas é o seguinte:

0	6	8	9												
1	0	0	0	2	2	2	2	4	5	5	5	6	8	8	8
2	0	4													

(a) A média é

$$\overline{x} = \frac{6+8+9+\cdots+20+24}{20} = \frac{274}{20} = 13,7$$

A moda é $x^* = 12$ (4 repetições) e a mediana é a média dos valores centrais:

$$Q_2 = \frac{x_{(10)} + x_{(11)}}{3} = \frac{12+14}{2} = 13$$

Todos esses resultados estão medidos em horas por semana.

(b) A amplitude é $\Delta = 24 - 6 = 18$ horas semanais. O desvio médio absoluto, também em horas semanais, é

$$\begin{aligned} DMA &= \frac{1}{20} \times \left[\begin{array}{l} |6-13,7| + |8-13,7| + |9-13,7| + 3 \times |10-13,7| + \\ 4 \times |12-13,7| + |14-13,7| + 3 \times |15-13,7| + \\ |16-13,7| + 3 \times |18-13,7| + |20-13,7| + |24-13,7| \end{array} \right] \\ &= 3,6 \end{aligned}$$

A variância, pela fórmula simplificada, é

$$\begin{aligned} \sigma^2 &= \frac{6^2 + 8^2 + 9^2 + 3 \times 10^2 + 4 \times 12^2 + 14^2 + 3 \times 15^2 + 16^2 + 3 \times 18^2 + 20^2 + 24^2}{20} - 13,7^2 \\ &= \frac{4132}{20} - 187,69 = 206,6 - 187,69 = 18,91 \Rightarrow \sigma = \sqrt{18.91} = 4,3486 \text{ horas semanais} \end{aligned}$$

: 4. 348 6

22. .

(a) Todos os salários ficaram aumentados em 250 reais. Se chamamos de x_i o salário do funcionário i no mês de novembro e de y_i o salário desse mesmo funcionário em dezembro, então $y_i = x_i + 250$. De acordo com a Propriedade 2, temos que o salário médio em dezembro é $\overline{y} = \overline{x} + 250 = 920 + 250 = 1170$ reais.

(b) Os novos salários são $y_i = x_i + 250$. Como visto na Propriedade 2, somar uma constante não altera as medidas de dispersão; logo, os novos salários têm o mesmo desvio padrão dos salários de novembro, ou seja, 180 reais.

23. .

(a) Seja x_i o salário do funcionário i no mês anterior ao dissídio. Depois do aumento, seu salário passa a ser $y_i = x_i + 0,089x_i = 1,089x_i$. Logo, todos os salários ficam multiplicados por 1,089 e, pela Propriedade 3, a média também fica multiplicada por este valor, ou seja, depois do dissídio o salário médio passa a ser $\overline{y} = 1,089\overline{x} = 1,089 \times 580 = 631,62$ reais.

(b) Como visto, os novos salários são $y_i = 1,089x_i$. Logo, pela Propriedade 3,

$$\sigma_y = 1,089\sigma_x = 1,089 \times 220 = 239,58$$

24. A diferença se deve à existência de grandes empresas no setor de bebidas, com muitos empregados. Como vimos, a média é bastante influenciada pelos valores discrepantes.

25. .

(a) Completando a tabela obtemos

Classe de PO	Ponto	Freq. Simples		Freq. Acumulada	
	médio	Absoluta	Relativa (%)	Absoluta	Relativa (%)
$[10, 30)$	20	489	53,9735	489	53,9735
$[30, 100)$	65	269	29,6909	758	83,6645
$[100, 500)$	300	117	12,9139	875	96,5784
$[500, 1000)$	750	15	1,6556	890	98,2340
$[1000, 2000)$	1500	9	0,9934	899	99,2274
$[2000, 4000)$	3000	7	0,7726	906	100,0000
TOTAL		906	100,0000		

Como as freqüências relativas estão em forma percentual, temos que dividir o resultado por 100, ou seja:

$$\begin{aligned}\overline{x} &= \begin{array}{c}20\times 0,539735+65\times 0,296909+300\times 0,129139+750\times 0,016556\\ +1500\times 0,009934+3000\times 0,007726\end{array}\\ &= 119,3 \text{ empregados}\end{aligned}$$

A mediana está na classe $10 \vdash 30$. A freqüência abaixo desta classe é nula. Logo, a regra de três é

$$\frac{Q_2-10}{50}=\frac{30-10}{53.9735}\Rightarrow Q_2-10=\frac{1000}{53.9735}\Rightarrow Q_2=28,528 \text{ empregados}$$

Note a diferença da média para a mediana, resultado da presença de empresas com muitos empregados - muitas empresas têm poucos empregados, mas poucas empresas têm muitos empregados, o que "puxa" a média para cima.

A classe modal é a primeira classe; logo, a freqüência da classe vizinha inferior é 0.

$$\text{King:}\qquad \frac{30-x^*}{x^*-10}=\frac{0}{269}\Rightarrow 30-x^*=0\Rightarrow x^*=30$$

Como a classe inferior não tem "força" para puxar a moda, a classe superior "puxa" até onde é possível, ou seja, até o limite superior da classe modal.

$$\text{Czuber:}\qquad \frac{30-x^*}{x^*-10}=\frac{489-269}{489-0}\Rightarrow 14670-489x^*=220x^*-2200\Rightarrow x^*=23,794$$

(b)

$$\begin{aligned}DMA &= \begin{array}{l}0,539735\times|20-119,3322|+0,296909\times|65-119,3322|\\ +0,129139\times|300-119,3322|+0,016556\times|750-119,3322|\\ +0,009934\times|1500-119,3322|+0,007726\times|3000-119,3322|\end{array}\\ &= 139,489691 \text{ empregados}\end{aligned}$$

$$\begin{aligned}\sigma^2 &= \begin{array}{l}0,539735\times 20^2+0,296909\times 65^2+0,129139\times 300^2+0,016556\times 750^2\\ +0,009934\times 1500^2+0,007726\times 3000^2-(119,3322)^2\end{array}\\ &= 114293,1843-14240,18102=100053,0033\end{aligned}$$

$$\sigma=\sqrt{100053,0033}=316,31 \text{ empregados}$$

26. A média dos dados é $\overline{x}=2,6$, com desvio padrão $\sigma=1,5620$. A moda é $x^*=2$. Os quartis são $Q_1=\frac{x_{(5)}+x_{(6)}}{2}=1,5; Q_2=\frac{x_{(10)}+x_{(11)}}{2}=2; Q_3=\frac{x_{(15)}+x_{(16)}}{2}=4$. Com esses valores obtemos os coeficientes de assimetria:

$$e=\frac{\overline{x}-x^*}{\sigma}=\frac{2,6-2}{1,5620}=0,3841$$

$$B=\frac{(Q_3-Q_2)-(Q_2-Q_1)}{Q_3-Q_1}=\frac{(4-2)-(2-1,5)}{4-1,5}=\frac{1,5}{3,5}=0,4286$$

Existe, assim, uma assimetria positiva nos dados; veja o diagrama de pontos na Figura 5.16.

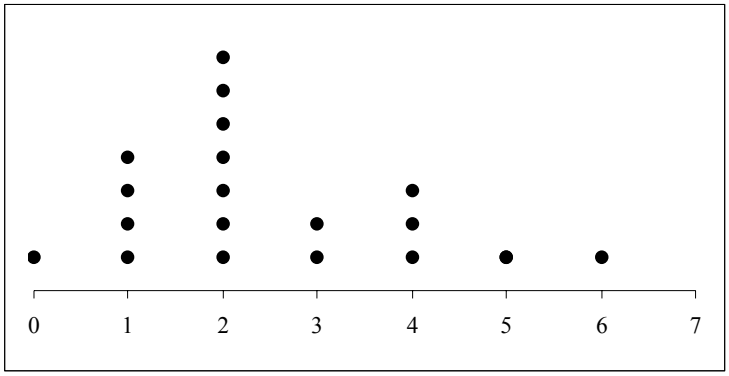


Figura 5.16: Número de apólices vendidas

27. .

(a) Há uma grande concentração de folhas no ramo 7. Nesses casos é usual "quebrar" o ramo em dois: no ramo superior ficam as folhas de 0 a 4 e no ramo inferior, as folhas de 5 a 9. Com isso fica mais saliente a maior concentração de clientes com pontos entre 70 e 74.

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

(b) Temos 30 clientes. Logo,

$$\begin{aligned} Q_2 &= \frac{x_{(15)} + x_{(16)}}{4} = 74 \\ Q_1 &= x_{(8)} = 72 \\ Q_3 &= x_{(23)} = 75 \\ IQ &= Q_3 - Q_1 = 75 - 72 = 3 \end{aligned}$$

(c) Veja a Figura 5.17. É visível a presença de dois valores discrepantes. Excluindo esses dois valores, a distribuição apresenta uma leve assimetria às esquerda - note que Q_2 está mais próximo de Q_3 do que de Q_1.

(d) A regra para premiação especial é a regra de valores discrepantes; assim, dois clientes ganharão a garrafa de champagne.

28. .

(a) $\overline{x} = 1020,8\text{g}$

(b) $\sigma^2 = 691,36 \qquad \sigma = 26,2937\text{g}$

(c) O limite superior da classe D é o 20º percentil; o da classe C é o 50º percentil, o da classe B é o 80º percentil e, obviamente, o da classe A é o valor máximo, 1080.

O 20º percentil está na classe $980 \vdash 1000$, onde se acumulam 22% da distribuição e a regra de proporcionalidade que o define é:

$$\frac{1000 - P_{20}}{0,22 - 0,20} = \frac{1000 - 980}{0,16} \Rightarrow P_{20} = 997,2$$

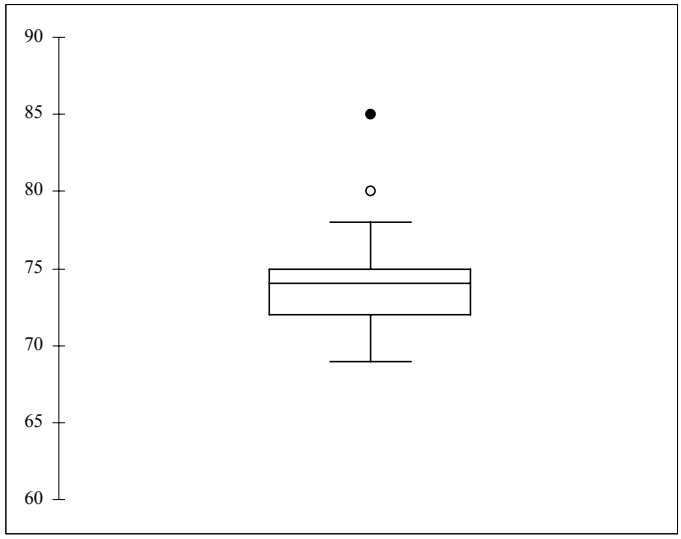


Figura 5.17: Exercício sobre promoção do supermercado

O 50º percentil (mediana) é o limite superior da terceira classe (note que nessa classe temos 50% da distribuição acumulada). O 80º percentil está na classe $1040 \vdash 1060$, onde se acumulam 92% da distribuição:

$$\frac{1060 - P_{80}}{0,92 - 0,80} = \frac{1000 - 980}{0,16} \Rightarrow P_{80} = 1045$$

As classes de peso são, pois: [960, 997,5); [997,5; 1020); [1020; 1045); >=1045.

(d) Ração reforçada: $\overline{x} - 2\sigma = 1020,8 - 2 \times 26,2937 = 968,2125$. Podemos estimar a porcentagem de frangos por uma regra de três análoga à utilizada para determinar qualquer separatriz. A diferençca é que agora temos a separatriz e queremos a freqüência, ou seja, a área. Veja a Figura 5.18

$$\frac{968,2125 - 960}{x} = \frac{980 - 960}{6} \Rightarrow x = 2,46\%$$

Reprodutores: $\overline{x} + 1,5 \times \sigma = 1020,8 + 1,5 \times 26,2937 = 1060,2406$.Veja a Figura 5.19

$$\frac{1080 - 1060,2406}{x} = \frac{1080 - 1060}{8} \Rightarrow x = 7,90\%$$

5.3 Capítulo 3

1. O número de folhas de cartolina compradas em cada ano é:

$$98 : \frac{500}{0,35} \qquad 99 : \frac{500}{0,45} \qquad 00 : \frac{500}{0,50}$$

e o valor total gasto é 3×500. Logo, o preço médio por folha de cartolina é

$$p_m = \frac{3 \times 500}{\frac{500}{0,35} + \frac{500}{0,45} + \frac{500}{0,50}} = \frac{3}{\frac{1}{0,35} + \frac{1}{0,45} + \frac{1}{0,50}} = 0,424$$

Note que o preço médio é a média harmônica dos preços pagps em cada ano.

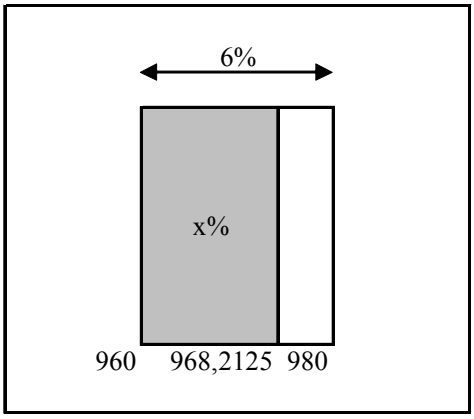


Figura 5.18: Percentual com ração reforçada

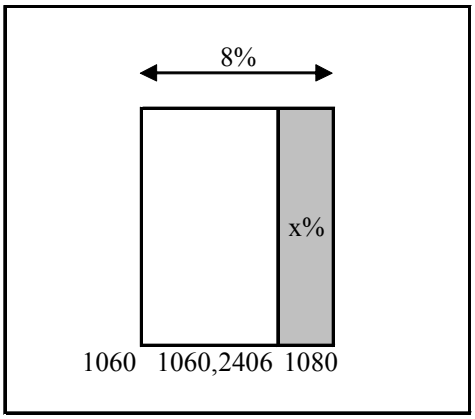


Figura 5.19: Percentual de reprodutores

2. O número de folhas de cartolina compradas em cada ano é:

$$98 : \frac{500}{0,35} \qquad 99 : \frac{550}{0,45} \qquad 00 : \frac{620}{0,50}$$

e o valor total gasto é 3×500. Logo, o preço médio por folha de cartolina é

$$\begin{aligned} p_m &= \frac{500 + 550 + 620}{\frac{500}{0,35} + \frac{550}{0,45} + \frac{620}{0,50}} = \frac{1670}{\frac{500}{0,35} + \frac{550}{0,45} + \frac{620}{0,50}} \\ &= \frac{1}{\frac{500}{1670} \times \frac{1}{0.35} + \frac{550}{1670} \times \frac{1}{0.45} + \frac{620}{1670} \times \frac{1}{0.50}} \\ &= 0,429 \end{aligned}$$

Note que o preço médio é a média harmônica *ponderada* dos preços pagos em cada ano. A ponderação é definida em termos do valor gasto em cada ano.

3. Veja o resumo na tabela a seguir:

	Homens	Mulheres
Média	5,2727	6,4000
Desvio padrão	1,8839	0,7642
CV	0,3573	0,1194

As mulheres, além da média mais alta, também tiveram menor variabilidade, o que pode ser visto pelo menor valor do CV.

4. Na tabela a seguir são dados os escores padronizados para cada UF:

UF	Escore z	UF	Escore z	UF	Escore z
RO	$-0,6125$	CE	$-0,0968$	RJ	$+3,0779$
AC	$-0,6354$	RN	$-0,0739$	SP	$+1,0263$
AM	$-0,6584$	PB	$+0,0178$	PR	$-0,1312$
RR	$-0,6584$	PE	$+0,2471$	SC	$-0,0280$
PA	$-0,6240$	AL	$+0,4877$	RS	$-0,2572$
AP	$-0,6354$	SE	$+0,2470$	MS	$-0,6125$
TO	$-0,6240$	BA	$-0,4062$	MT	$-0,6469$
MA	$-0,4865$	MG	$-0,3260$	GO	$-0,5094$
PI	$-0,5438$	ES	$+0,0980$	DF	$+3,3644$

Rio de Janeiro e Distrito Federal pdem ser considerados valores discrepantes, uma vez que seus escores z estão acima de $+3$. Os estados da região Norte têm densidade abaixo de média.

5. O crescimento global nos três dias foi de $\frac{9200}{2500} = 3,68$; logo, o percentual médio de crescimento foi de $100 \times \left(\sqrt[3]{3,68} - 1\right) = 100 \times (1,543889 - 1) = 54,39\%$. Aqui você tem que usar a média geométrica porque as novas bactérias também se reproduzem; é como se tivéssemos um regime de capitalização composta.

6. A inflação acumulada até novembro é:

$$1,007 \times 1,0105 \times \cdots \times 1,0095 = 1,08290$$

Como queremos a inflação anual no máximo de 9%, temos que ter

$$1,08290 \times i_{12} \leq 1,09 \Rightarrow i_{12} \leq \frac{1,09}{1,08290} = 1,006556$$

que equivale a uma taxa máxima de 0,66%.

7.

$$\begin{array}{rcll} \overline{x}_A & = & 55,7222 & \sigma_A = 7,4596 \qquad CV_A = 13,3871 \\ \overline{x}_B & = & 55,4286 & \sigma_B = 3,0949 \qquad CV_B = 5,5835 \end{array}$$

Rendimentos médios semelhantes mas dispersão da corretora A é muito maior. A corretora B parece ter um comportamento mais estável.

5.4 Capítulo 4

1. .

   (a) 40
   (b) $11 + 152 + 140 = 303$
   (c) $143 + 171 + 40 + 11 + 152 + 140 + 66 + 57 + 20 = 800$
   (d) $\dfrac{66}{66 + 57 + 20} = 0.4615$ ou 46,15%
   (e) Veja a Figura 5.20

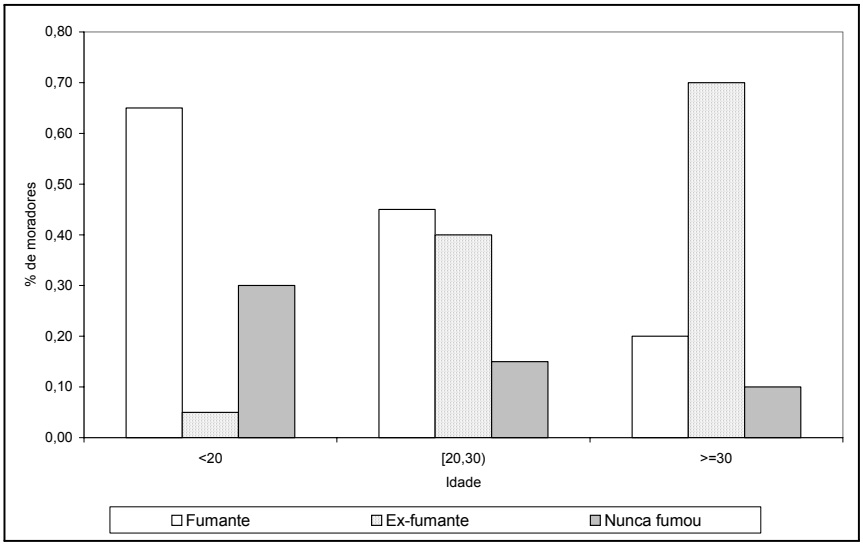


Figura 5.20: Distribuição dos moradores segundo a idade e o hábito de fumar

2. Seja X = "quantidade produzida" e Y = "preço". Note que atendfência é o preço diminuir à medida que a produção aumenta. Temos os seguintes resultados:

$$\begin{array}{rcll} \sum x_i & = & 67881,1 & \sum x_i^2 = 227383441,5 \\ \sum y_i & = & 3914,5 & \sum y_i^2 = 328733,49 \\ \sum x_i y_i & = & 4813211,18 & \end{array}$$

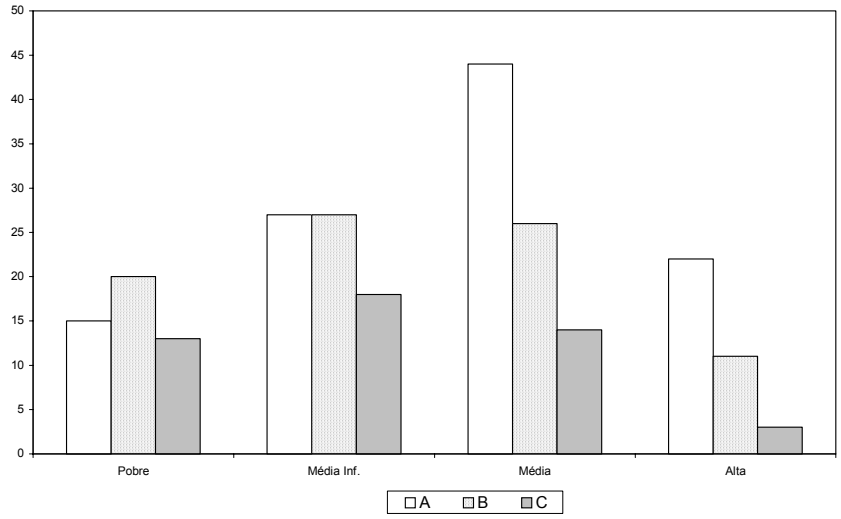


Figura 5.21:

Logo,

$$Cov(X,Y) = \frac{4813211.18}{50} - \frac{67881.1}{50} \times \frac{3914.5}{50} = -10024,00278$$

$$\sigma_x^2 = \frac{227383441.5}{50} - \left(\frac{67881.1}{50}\right)^2 = 2704531,334516$$

$$\sigma_y^2 = \frac{328733,49}{50} - \left(\frac{3914,5}{50}\right)^2 = 445,34570000$$

$$Corr(X,Y) = \frac{-10024.00278}{\sqrt{2704531.334516 \times 445.3457}} = -0,2888$$

Uma associação linear moderada entre preço e quantidade produzida, indicando que à medida que aumenta a quantidade produzida de ovos, o preço pago aos produtores tende a baixar.

3. Veja a Figura **??**.

   Na classe alta há $22 + 11 + 3 = 36$ pessoas; há $15 + 27 + 44 + 22 = 108$ leitores do jornal A; o percentual de leitores do jornal C entre os pobres é $\frac{13}{15+20+13} = 0.27083$ ou 27,80% dos pobres l˜eem o jornal C.

4. O diagrama de dispersão encontra-se na Figura **??**, onde podemos ver que à medida que a área da casa aumenta, opreço também aumenta.

   Fazendo $X =$ "área da casa" e $Y =$ "preço de venda", temos os seguintes resultados:

$$\begin{array}{rcll} \sum x_i & = & 14433 & \sum x_i^2 = 3736397 \\ \sum y_i & = & 10472 & \sum y_i^2 = 1976810 \\ \sum x_i y_i & = & 2667287 & \end{array}$$

   Logo

$$Cov(X,Y) = \frac{2667287}{59} - \frac{14433}{59} \times \frac{10472}{59} = 1789,013789$$

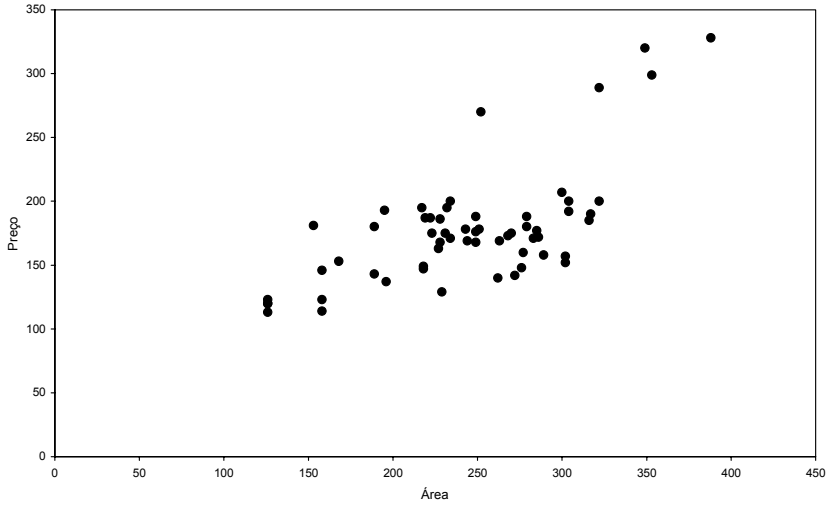


Figura 5.22:

$$\sigma_x^2 = \frac{3736397}{59} - \left(\frac{14433}{59}\right)^2 = 3486,335536$$

$$\sigma_y^2 = \frac{1976810}{59} - \left(\frac{10472}{59}\right)^2 = 2002,01264$$

$$Corr(X,Y) = \frac{1789.013789}{\sqrt{3486.335536 \times 2002.01264}} = 0,6772$$

Associação linear forte, indicando que à medida que a área da casa aumenta, o seu preço de venda também aumenta.

5. Para os 4 conjuntos, temos que as médias de X e Y são as mesmas, assim como o coeficiente de correlação.

$$\begin{array}{rcll} \overline{X} & = & 9 & \overline{Y} = 7,50091 \\ \sigma_X & = & 3,16228 & \sigma_Y = 1,93711 \\ \rho(X,Y) & = & 0,816 & \end{array}$$

No entanto, os conjuntos são completamente diferentes, conforme ilustrado pelos diagramas de dispersão da Figura **??**. Então, uma análise de dados não deve se basear em apenas uma medida descritiva; é importante que diferentes aspectos sejam analisados, inclusive através de representações gráficas adequadas.

6. A idéia é usar como *proxy* a variável mais fortemente associada com a variável de interesse, que é capacidade da produção instalada. Vamos, então, calcular os coeficientes de correlação entre essa variável e as duas "candidatas". Usando os valores dados, temos que:

$$\rho(X,Y) = \frac{361 - \frac{80 \times 38}{10}}{\sqrt{736 - \frac{80^2}{10}}\sqrt{182 - \frac{38^2}{10}}} = 0,9487$$

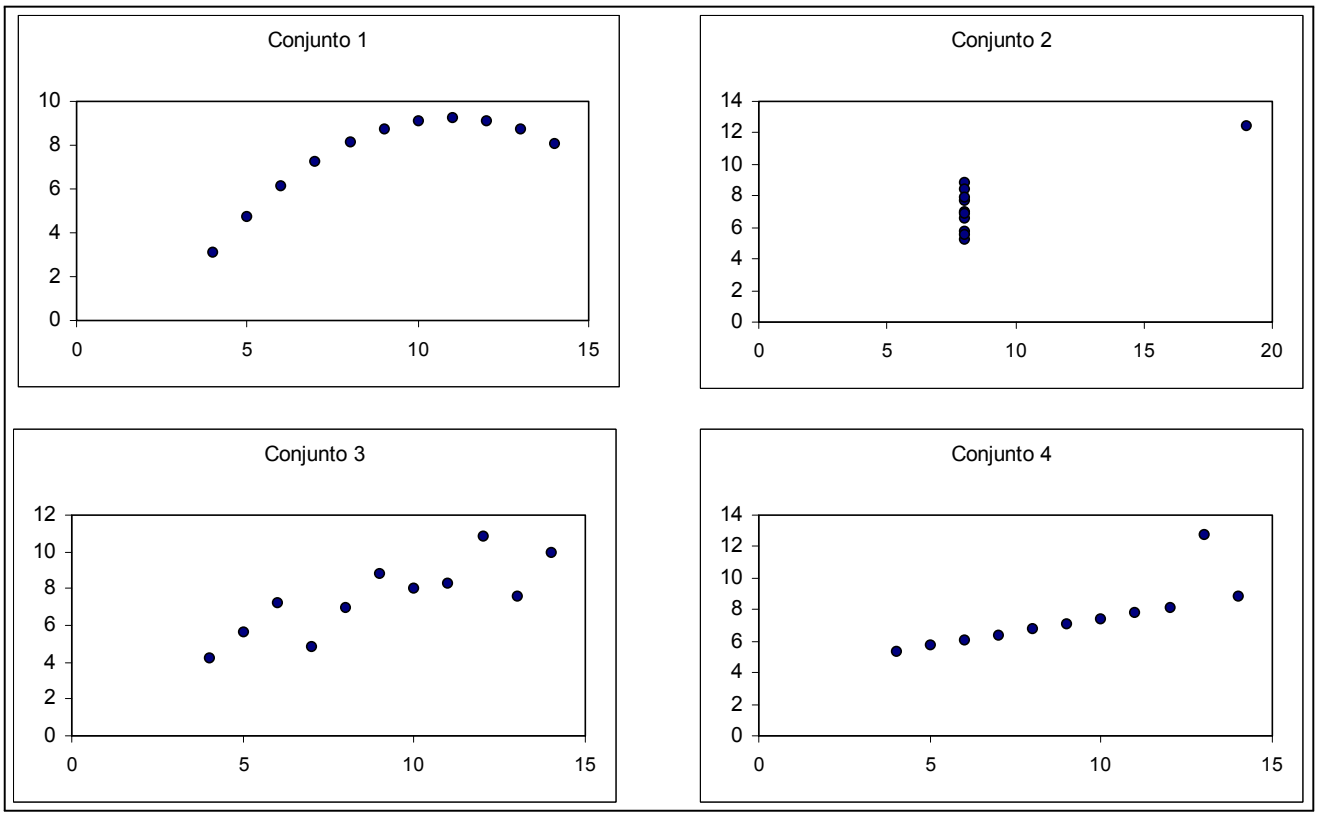


Figura 5.23: Diagramas de dispersão para os dados de Anscombe

$$\rho(X,Z) = \frac{848 - \frac{80\times100}{10}}{\sqrt{736 - \frac{80^2}{10}}\sqrt{1048 - \frac{100^2}{10}}} = 0,7071$$

Logo, a variável a ser utilizada como *proxy* deverá ser Potência Instalada, que apresenta maior correlação com a variável de interesse.

7. O sinal deve ser positivo, uma vez que, aumentando a renda, as despesas em alimentação tendem a aumentar. Para os dados em questão temos:

$$\begin{aligned} \sum X_i &= 5212,52 \qquad \sum Y_i = 27920 \qquad \sum X_iY_i = 3834936,4970 \\ \sum X_i^2 &= 758791,6732 \qquad \sum Y_i^2 = 21020623,02 \end{aligned}$$

Logo,

$$\rho(X,Y) = \frac{3834936,4970 - \frac{5212,52\times27920}{40}}{\sqrt{758791,6732 - \frac{(5212,52)^2}{40}}\sqrt{21020623,02 - \frac{(27920)^2}{40}}} = 0,5631$$

Bibliografia

[1] Anscombe, F.J. (1974), Graphs in statistical analysis, *The American Statistician,* 27(1973), pp. 17-21.

[2] Bussab, W.O. e Morettin, P.A. *Estatística Básica*, 5a. ed., São Paulo: Editora Saraiva, 2003 .

[3] Dunn, O.J. e Clark, V.A. *Applied Statistics: Analysis of Variance and Regression*, Nova York: John Wiley & Sons.

[4] Moore, D.S. e McCabe, G.P. *Introdução à Prática da Estatística*, 3ª ed., Rio de Janeiro: LTC Editora, 2002

[5] Moore, D.S., McCabe, G.P., Duckworth, W.M., Sclove, S.L. A *Prática da Estatística Empresarial*, Rio de Janeiro: LTC Editora, 2006

[6] Soares, J.F., Farias, A.A. e Cesar, C.C. *Introdução à Estatística*, Rio de Janeiro: LTC Editora, 2002