Núcleo de Tradução das Sociais

O presente texto foi modificado para otimizar arquivos PDF, o processo envolve a separação de uma página em duas, e o reconhecimento de texto em imagens, de maneira que o arquivo se torne grifável por meio de programas OCR (Optical Character Recognition).

Vale lembrar que, a disponibilização de arquivos digitais de qualidade na faculdade também é uma pauta de permanência estudantil, uma vez que a experiência de leitura – tão crucial num curso de ciências sociais – é extremamente influente no processo de entendimento do material.

Caso tenha interesse em participar do nosso projeto, entre em contato no instagram: @nts.usp

(b) Muitos artigos de jornais que sugerem que um alimento, medicamento ou agente ambiental em particular é nocivo ou benéfico devem ser vistos com ceticismo, a não ser que saibamos mais sobre o delineamento e a análise estatística utilizados no estudo.

(c) Esse resultado sugere que você deve ser cético quanto aos resultados de estudos médicos publicados que não sejam estudos aleatórios e controlados.

(d) Controlar a tendenciosidade tanto suspeita quanto insuspeita é necessário na pesquisa médica, mas não na pesquisa social, porque as ciências sociais lidam com verdades subjetivas em vez de objetivas.

2.37 Um PSG recente perguntou às pessoas se elas apoiavam a legalização do aborto em cada uma de sete diferentes circunstâncias. O percentual que apoiava a legalização do aborto variou entre 45% (se a mulher o deseja por qualquer motivo) a 92% (se a saúde da mulher está seriamente em perigo devido à gravidez). Isso indica que:

(a) as respostas podem depender muito da maneira como a pergunta é formulada.

(b) os levantamentos amostrais lidam com uma pequena parte da população e nunca podemos confiar neles.

(c) a amostra não precisa ser selecionada aleatoriamente.

(d) a amostra deve ter problemas de tendenciosidade como resultado de as pessoas não dizerem a verdade.

2.38 Um pesquisador está na entrada de um *shopping* popular realizando entrevistas. Verdadeiro ou falso: como não podemos prever quem será entrevistado, a amostra obtida é um exemplo de amostra aleatória. Explique.

2.39 Em um concurso recente de Miss América, os telespectadores podiam votar se cancelavam o desfile de traje de banho ligando para um número fornecido pelo canal de televisão. Aproximadamente 1 milhão de telespectadores ligaram e registraram a sua opinião, dos quais 79% disseram que queriam ver as candidatas vestidas como beldades em traje de banho. Verdadeiro ou falso: visto que todos tiveram a chance de ligar, trata-se de uma amostra aleatória simples de todos os telespectadores desse programa. Explique.

***2.40** Uma escala intervalar para a qual as proporções são válidas é chamada de uma **escala de proporções**. Tais escalas têm um ponto 0 bem definido, assim, por exemplo, podemos considerar o valor 20 como o dobro da quantidade do valor 10. Explique por que a renda anual é mensurada em uma escala de proporções, mas a temperatura (em Fahrenheit ou Centígrado) não é. O QI, como uma medida de inteligência, é uma variável expressa em uma escala de proporções?

☑ NOTAS

[1] D. M. Wilbur, *Public Perspective,* disponível em http://roperweb.ropercenter.uconn.edu, Maio-Junho de 1993.

[2] Coluna de T. Friedman, *New York Times*, 2 de março de 2006.

[3] Veja Crossom (1994).

[4] *Washington Post*, 26 de junho de 1995.

[5] *Newsweek*, 25 de julho de 1994.

[6] *Fonte: A Mathematician Reads the Newspaper,* por J. A. Paulos, Basic Books, 1995, p. 15.

☑ 3 ESTATÍSTICA DESCRITIVA

Vimos que os métodos estatísticos são *descritivos* ou *inferenciais*. O propósito da estatística descritiva é resumir os dados, facilitar a assimilação da informação. Este capítulo apresenta os métodos básicos da estatística descritiva.

Apresentamos, em primeiro lugar, tabelas e gráficos que descrevem os dados mostrando o número de vezes em que vários resultados ocorrem. As variáveis quantitativas também apresentam duas características-chave para descrevê-las numericamente:

- O **centro** dos dados – uma observação típica.
- A **variabilidade** dos dados – a dispersão em torno do centro.

Aprenderemos a descrever dados quantitativos com estatísticas que resumem o centro e a variabilidade e, finalmente, com estatísticas que especificam certas posições nos conjuntos de dados que resumem tanto o centro quanto a variabilidade.

3.1 DESCREVENDO DADOS COM TABELAS E GRÁFICOS

As tabelas e gráficos são úteis para todos os tipos de dados. Começaremos com as variáveis categóricas.

Frequências relativas: dados categóricos

Para dados categóricos listamos as categorias e mostramos a frequência (o número de observações) em cada categoria. Para facilitar a comparação de diferentes categorias, relatamos, também, as proporções ou percentuais, também chamados de **frequências relativas.**

> ☑ **Frequência relativa**
>
> A **frequência relativa** para uma categoria é a **proporção** ou o **percentual** das observações que pertencem àquela categoria.

A *proporção* é igual ao número de observações em uma dada categoria dividida pelo número total de observações. É um número entre 0 e 1 que expressa o percentual de observações naquela categoria.

EXEMPLO 3.1 Estrutura domiciliar norte-americana

A Tabela 3.1 lista tipos diferentes de domicílios nos Estados Unidos em 2005. De 111,1 milhões de domicílios, por exemplo, 24,1 milhões eram de um casal com filhos. A proporção 24,1/111,1 = 0,22 era de um casal com filhos.

O percentual é a proporção expressa em relação a 100, isto é, a vírgula é movida duas posições para a direita. Por exemplo, 0,22 é a proporção de famílias casadas com filhos, então o percentual é 0,22 = 22%. A Tabela 3.1 mostra as proporções e os percentuais para todas as categorias. ■

A soma das proporções é igual a 1,00. A soma dos percentuais é igual a 100. (Na

☑ **Tabela 3.1** Estrutura domiciliar dos Estados Unidos, 2005

Tipo de Família	Número (milhões)	Proporção	Percentual
Casal com filhos	24,1	0,22	22
Casal sem filhos	31,1	0,28	28
Solteiro, sem parceiro	19,1	0,17	17
Morando sozinho	30,1	0,27	27
Outros domicílios	6,7	0,06	6
Total	111,1	1,00	100

Fonte: Agência do Censo dos Estados Unidos, *2005 American Community Survey*, Tabelas B1 1001, C11003.

prática, os valores podem ter a soma de um número levemente diferente, como 99,9 ou 100,1, por causa do arredondamento.)

É suficiente, em uma tabela como esta, relatar os percentuais (proporções) e o tamanho da amostra, visto que cada frequência é igual à proporção correspondente multiplicada pelo tamanho da amostra. Por exemplo, a frequência de casais com filhos é igual a 0,22(111,1) = 24 milhões. Quando representar os percentuais, mas não as frequências, sempre inclua o tamanho da amostra.

Distribuições de frequências e diagramas de colunas: dados categóricos

A Tabela 3.1 lista as categorias de domicílios e o número de domicílios de cada tipo. Tal lista é chamada de *distribuição de frequências.*

☑ **Distribuição de frequências**

Uma **distribuição de frequências** é uma lista de valores possíveis para uma variável, junto com o número de observações de cada valor. Uma **distribuição de frequências relativas** lista os valores possíveis juntamente com suas proporções ou percentuais.

Para construir uma distribuição de frequências para uma variável categórica, liste as categorias e conte o número de observações em cada uma.

Para mais facilmente ter uma ideia dos dados, é útil olhar um gráfico para a distribuição de frequências relativas. Um **diagrama de colunas** apresenta colunas retangulares desenhadas sobre cada categoria. A altura da coluna mostra a frequência relativa daquela categoria. A Figura 3.1 é um diagrama de colunas para os dados na Tabela 3.1. As colunas estão separadas para enfatizar que a variável é categórica e não quantitativa contínua. Visto que a estrutura domiciliar é uma variável nominal, não existe uma ordem natural particular para as colunas. A ordem de apresentação de uma variável nominal é a ordem natural das categorias.

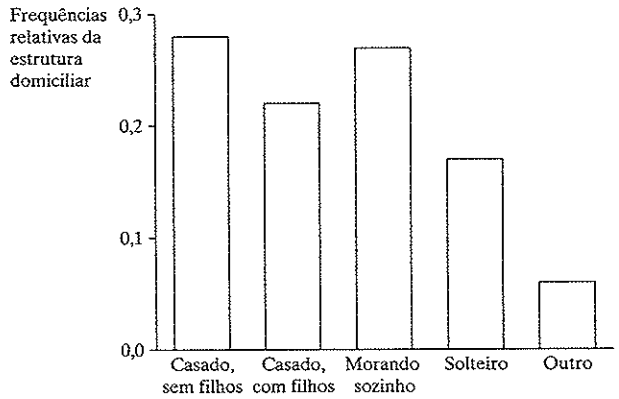
Outro tipo de gráfico, o *diagrama de pizza*, é um círculo tendo "uma fatia da pizza" para cada categoria. O tamanho da fatia representa o percentual das observações na categoria. O diagrama de colunas é mais preciso do que o diagrama de pizza para comparações visuais das categorias com frequências relativas similares.

Distribuições de frequências: dados quantitativos

As distribuições de frequências e gráficos também são úteis para variáveis quantitativas. O próximo exemplo ilustra um diagrama para uma variável quantitativa contínua.

EXEMPLO 3.2 Taxas de crimes violentos no estado

A Tabela 3.2 lista todos os 50 estados americanos e suas taxas de crimes violentos em 2005. Cada taxa mensura o número de crimes violentos naquele estado em 2005 por uma população de 10000. Por exemplo, se um estado teve 12000 crimes violentos e uma população de 2300000, sua taxa de crimes violentos seria (12000/2300000)×10000 = 52. É difícil saber mais, simplesmente lendo as taxas de crimes violentos. As tabelas, os gráficos e as medidas numéricas nos ajudam a absorver a informação desses dados em maior profundidade.



☑ **Figura 3.1** Frequências relativas da estrutura domiciliar dos Estados Unidos, 2005.

Primeiro, podemos resumir os dados com uma distribuição de frequências. Para fazer isso, dividimos a escala de mensuração para a taxa de crimes violentos em intervalos e contamos o número de observações em cada intervalo. Aqui usamos o conjunto intervalar {0-11, 12-23, 24-35, 36-47, 48-59, 60-71, 72-83}. Os valores que a Tabela 3.2

☑ **Tabela 3.2** Lista dos estados com as taxas de crimes violentos mensuradas como número de crimes violentos para uma população de 10000

Alabama	43	Louisiana	65	Ohio	33
Alaska	59	Maine	11	Oklahoma	51
Arizona	51	Maryland	70	Oregon	30
Arkansas	46	Massachussets	47	Pennsylvania	40
Califórnia	58	Michigan	51	Rhode Island	29
Colorado	34	Minnesota	26	Carolina do Sul	79
Connecticut	31	Mississippi	33	Dakota do Sul	17
Delaware	66	Missouri	47	Tennessee	69
Flórida	73	Montana	36	Texas	55
Geórgia	45	Nebraska	29	Utah	25
Havaí	27	Nevada	61	Vermont	11
Idaho	24	New Hampshire	15	Virginia	28
Illinois	56	Nova Jersey	37	Washington	35
Indiana	35	Novo México	66	Virgínia do Oeste	26
Iowa	27	Nova Iorque	46	Wisconsin	22
Kansas	40	Carolina do Norte	46	Wyoming	26
Kentucky	26	Dakota do Norte	8		

registra foram arredondados; assim, por exemplo, o intervalo 12-23 representa os valores entre 11,5 e 23,5. Contando o número de estados com taxas de crimes violentos em cada intervalo, conseguimos a distribuição de frequências mostrada na Tabela 3.3. Observamos que existe uma variabilidade considerável nas taxas de crimes violentos.

A Tabela 3.3 também mostra as frequências relativas usando proporções e percentuais. Por exemplo, 3/50 = 0,06 é a proporção para o intervalo 0-11 e 0,06 = 6% é o percentual. Assim como com qualquer método de resumo, perdemos alguma informação ao custo de obter objetividade. A distribuição de frequências não identifica quais estados têm as taxas mais altas ou mais baixas de crimes violentos nem as taxas exatas de crimes violentos são exibidas. ■

Os intervalos dos valores nas distribuições de frequências têm, geralmente, a mesma largura. A largura é igual a 12 na Tabela 3.3. Os intervalos devem incluir todos os valores possíveis da variável. Além disso, qualquer valor possível deve se ajustar somente em um intervalo; isto é, eles devem ser **mutuamente exclusivos**.

Histogramas

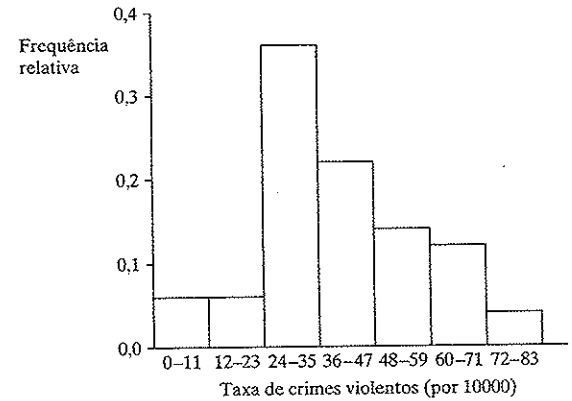
Um gráfico de uma distribuição de frequências relativas para uma variável quantitativa contínua é chamado de **histograma**. Cada intervalo tem uma coluna sobre ele, com a altura representando o número de observações naquele intervalo. A Figura 3.2 é um histograma para as taxas de crimes violentos.

Escolher intervalos para as distribuições de frequências e histogramas é, primeiramente, uma questão de bom senso. Se poucos intervalos são usados, muita informação é perdida. Por exemplo, a Figura 3.3 é um histograma da taxa de crimes violentos usando os intervalos 0-29, 30-59, 60-89. Isto é muito pouco para ser informativo. Se muitos intervalos são usados, eles são tão estreitos que a informação apresentada fica difícil de ser entendida e o histograma pode ser irregular e o padrão geral dos resultados pode ficar obscuro. Em condições ideais, duas observações no mesmo intervalo devem ser similares em um sentido prático. Para resumir a renda anual, por exemplo, se a diferença de $5000 na renda não é considerada praticamente importante, porém a diferença de $15000 é considerável, devemos escolher intervalos de largura menor do que $15000, como $0-$9999, $10000-$19999, $20000-$29999 e assim por diante. Um *software* estatístico pode, automaticamente, escolher os intervalos para nós e construir distribuições de frequências e histogramas.

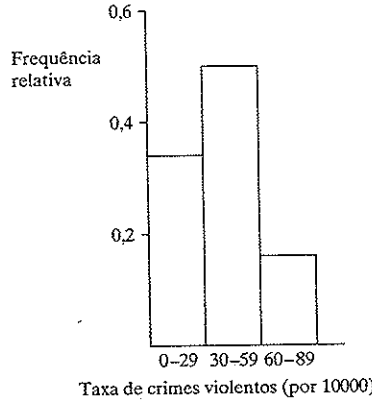
Para uma variável discreta com relativamente poucos valores, um diagrama adequado é de colunas com uma coluna representando cada valor possível. Para uma variável contínua ou uma variável discreta com muitos valores possíveis, você precisa dividir os valores possíveis em intervalos, como fizemos com as taxas de crimes violentos.

☑ **Tabela 3.3** Distribuição de frequências absolutas e relativas para as taxas de crimes violentos

Taxa de crimes violentos	Frequência	Frequência relativa	Percentual
0-11	3	0,06	6
12-23	3	0,06	6
24-35	18	0,36	36
36-37	11	0,22	22
48-59	7	0,14	14
60-71	6	0,12	12
72-83	2	0,04	4
Total	50	1,00	100,0



☑ **Figura 3.2** Histograma de frequências relativas para as taxas de crimes violentos dos estados norte-americanos.



☑ **Figura 3.3** Histograma de frequências relativas para a taxa de crimes violentos, usando poucos intervalos.

Diagrama de caule e folhas

A Figura 3.4 mostra uma representação gráfica alternativa dos dados da taxa de crimes violentos. Esta figura, chamada de **diagrama de caule e folhas**, representa cada observação por seu(s) primeiro(s) dígito(s) (o *caule*) e por seu dígito final (a *folha*). Cada caule é um número à esquerda da barra vertical e uma folha é um número à direita dela. Por exemplo, na segunda linha, o caule 1 e as folhas 1, 1, 5 e 7 representam a taxa de crimes violentos 11, 11, 15, 17. O diagrama ordena as folhas em cada linha, do menor para a maior valor.

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

☑ **Figura 3.4** Diagrama de caule e folhas para os dados da taxa de crimes violentos da Tabela 3.2.

Um diagrama de caule e folhas contém a mesma informação de um histograma. Virando de lado, ele tem a mesma forma do histograma. Na verdade, visto que o diagrama de caule e folhas mostra cada observação, ele exibe informações que são perdidas pelo histograma. Da Figura 3.4 a maior taxa de crimes violentos era 79 e a menor era de 8 (exibido como 08 com um caule de 0 e uma folha de 8). Não é possível determinar estes valores exatos a partir do histograma na Figura 3.2.

Os diagramas de caule e folhas são úteis para representações rápidas de pequenos conjuntos de dados. À medida que o tamanho da amostra aumenta, você pode acomodar o aumento nas folhas separando os caules. Por exemplo, você pode listar cada caule duas vezes colocando as folhas de 0 e 4 em uma linha e as folhas de 5 e 9 em outra. Quando um número tem vários dígitos é mais simples para uma representação gráfica suprimir o último digito ou dois. Por exemplo, para um diagrama de caule e folhas da renda anual em milhares de dólares, um valor de $27,1 mil tem um caule de 2 e uma folha de 7 e um valor de $106,4 mil tem um caule de 10 e uma folha de 6.

Comparando grupos

Muitos estudos comparam diferentes grupos considerando alguma variável. As distribuições de frequências relativas, histogramas e diagramas de caule e folhas são úteis para fazer comparações.

EXEMPLO 3.3 Comparando as taxas de assassinatos norte-americanas e canadenses

Os diagramas de caule e folhas podem fornecer comparações visuais de duas amostras pequenas de uma variável quantitativa. Para facilitar a comparação, os resultados são representados graficamente "um contra o outro". Cada diagrama usa o mesmo caule com folhas para uma amostra à sua esquerda e folhas para a outra amostra à direita. Para ilustrar, a Figura 3.5 mostra diagramas de caule e folhas "um contra o outro" das taxas de assassinatos recentes (mensuradas como o número de assassinatos de uma população de 10000) para os 50 estados dos Estados Unidos e para as províncias do Canadá. Desta figura, fica claro que as taxas de assassinato tendem a ser mais baixas no Canadá, variando entre 0,7 (Prince Edward Island) a 2,9 (Manitoba) enquanto nos Estados Unidos elas variaram entre 1,6 (Maine) a 20,3 (Louisiana).

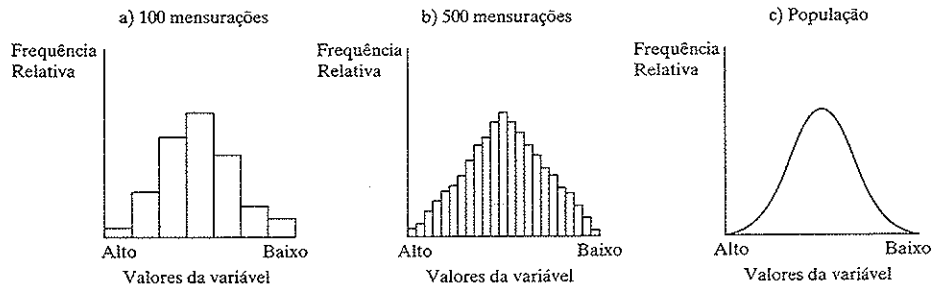
Distribuição da população e distribuição dos dados amostrais

As distribuições de frequência e histogramas se aplicam tanto à população e às amostras daquela população. O primeiro tipo é chamado de **distribuição da população** e o segundo tipo é chamado de **distribuição dos dados amostrais**. De certa forma, a distribuição dos dados amostrais é uma foto indistinta da população. À medida que o tamanho da amostra aumenta, a proporção da amostra em qualquer intervalo chega próximo da proporção populacional verdadeira. Dessa forma, a distribuição dos dados amostrais se parece mais com a distribuição da população.

Canadá	Caule	Estados Unidos
7	0	
3 2 1	1	6 7
9 7 6 3 2 0	2	0 3 9
	3	0 1 4 4 4 6 8 9 9 9
	4	4 6
	5	0 2 3 8
	6	0 3 4 6 8 9
	7	5
	8	0 3 4 6 9
	9	0 8
	10	2 2 3 4
	11	3 3 4 4 6 9
	12	7
	13	1 3 5
	14	
	15	
	16	
	17	
	18	
	19	
	20	3

☑ **Figura 3.5** Diagramas de caule e folhas "um contra o outro" das taxas de assassinatos norte-americanas e canadenses. Ambos dividem o mesmo caule, com as folhas do Canadá à esquerda e as folhas dos Estados Unidos à direita.

Para uma variável contínua, imagine o tamanho da amostra crescendo indefinidamente, com o número de intervalos aumentando simultaneamente, assim suas larguras ficam mais estreitas. Então, a forma do histograma da amostra gradualmente se aproxima de uma curva suave. Este livro usa tais curvas para representar as distribuições da população. A Figura 3.6 mostra dois histogramas com base em amostras, uma de tamanho 100 e outra de tamanho 500 e também uma curva suave representando a distribuição da população. Mesmo se uma variável for discreta, uma curva suave geralmente se aproxima bem da distribuição da



☑ **Figura 3.6** Histogramas para uma variável contínua. Usamos curvas suaves para representar as distribuições populacionais de variáveis contínuas.

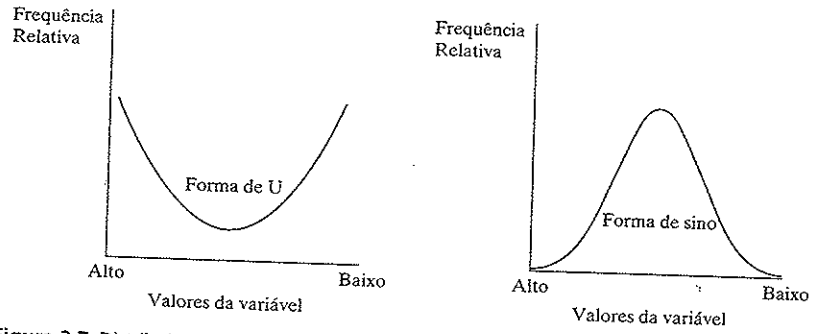
população, especialmente quando o número de valores possíveis da variável é grande.

A forma da distribuição

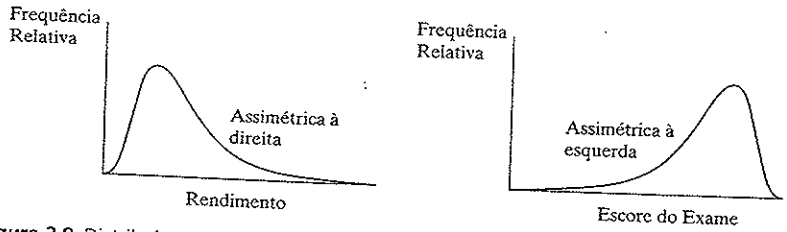
Uma forma de resumir uma amostra ou a distribuição de uma população é descrever sua forma. Um grupo para o qual a distribuição tem a forma de sino é fundamentalmente diferente de um grupo para o qual a distribuição tem a forma de U, por exemplo. Veja a Figura 3.7. Na distribuição com a forma de U, os pontos mais altos (representando as frequências mais altas) estão nos escores mais baixos e mais altos, enquanto, na distribuição com a forma de sino, o ponto mais alto está próximo do valor do meio. Uma distribuição com a forma de U indica a polarização na variável entre dois conjuntos de sujeitos. Uma distribuição com forma de sino indica que a maioria dos sujeitos tendem a estar próximos a um valor central.

As distribuições na Figura 3.7 são **simétricas**. O lado abaixo do valor central da distribuição é uma imagem espelhada do lado acima daquele valor central. Muitas das distribuições encontradas nas ciências sociais não são simétricas. A Figura 3.8 ilustra essa relação. As partes da curva para os valores mais baixos e para os valores mais altos são chamadas de **caudas** da distribuição. Geralmente, como na Figura 3.8, uma cauda é mais longa do que a outra. Uma distribuição é dita **assimétrica à direita** ou **assimétrica à esquerda**, de acordo com a cauda mais longa.

Para comparar as distribuições de frequências ou histogramas para dois grupos, você pode dar uma descrição verbal usando características como a inclinação. É também útil fazer comparações numéricas como: "Na média, a taxa de assassinatos dos estados americanos é 5,4 mais alta do



☑ **Figura 3.7** Distribuições de frequências em forma de U e de sino.



☑ **Figura 3.8** Distribuições de frequências assimétricas. A cauda mais longa indica a direção da inclinação.

que a das províncias canadenses". Agora, voltamos nossa atenção para as estatísticas descritivas numéricas.

3.2 DESCREVENDO O CENTRO DOS DADOS

Esta seção apresenta a estatística que descreve o centro de uma distribuição de frequências para uma variável quantitativa. A estatística mostra como é uma observação *típica*.

A média

A medida do centro mais conhecida e mais comumente usada é a **média**.

☑ **Média**

A **média** é a soma de todas as observações dividida pelo número de observações.

EXEMPLO 3.4 Atividade econômica feminina na Europa

A Tabela 3.4 mostra um indicador da atividade econômica feminina para os países da América do Sul e do Leste Europeu em 2003. O número especifica as mulheres empregadas como um percentual dos homens empregados. Na Argentina, por exemplo, o número de mulheres na força de trabalho era de 48% do número de homens na força de trabalho. (O valor era de 83 nos Estados Unidos e no Canadá.)

Para as oito observações do Leste Europeu, a soma é igual a

$$83 + 82 + 72 + 80 + 80 + 81 + 84 + 81 = 643$$

A atividade econômica feminina média é igual a 643/8 = 80,4. Por comparação, você pode verificar que a média para os 11 países da América do Sul é igual a 573/11 = 52,1. A atividade econômica feminina tende a ser consideravelmente mais baixa na América do Sul do que no Leste Europeu. ■

Usamos a seguinte notação para a média nas fórmulas para ela e para estatísticas que usem a média.

☑ **Notação para observações e média amostral**

O tamanho da amostra é representado por n. Para uma variável representada por y, suas observações são representadas por $y_1, y_2, \ldots, y_n$. A média amostral é representada por $\bar{y}$.

☑ **Tabela 3.4** Atividade econômica feminina na América do Sul e no Leste Europeu; mulheres empregadas como um percentual dos homens empregados

América do Sul		Leste Europeu	
País	**Atividade**	**País**	**Atividade**
Argentina	48	República Tcheca	83
Bolívia	58	Estônia	82
Brasil	52	Hungria	72
Chile	50	Letônia	80
Colômbia	62	Lituânia	80
Equador	40	Polônia	81
Guiana	51	Eslováquia	84
Paraguai	44	Eslovênia	81
Peru	45		
Uruguai	68		
Venezuela	55		

Fonte: Human Development Report 2005, United Nations Development Programme.

O símbolo $\bar{y}$ para a média amostral é lido como "y barra". Por todo o livro, as letras próximas ao final do alfabeto representam as variáveis. As n observações da amostra em uma variável y são representadas por y_1 para a primeira observação, y_2 na segunda e assim por diante. Por exemplo, para a atividade econômica feminina no Leste Europeu, $n = 8$ e as observações são $y_1 = 83$, $y_2 = 82$,..., $y_8 = 81$. Uma barra sobre a letra representa a média amostral para aquela variável. Por exemplo, $\bar{x}$ representa a média amostral para a variável representada por x.

A definição para a média amostral diz que

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

O símbolo Σ (a letra maiúscula grega sigma) representa a soma. Por exemplo, Σy_i representa a soma de $y_1 + y_2 + \ldots + y_n$. Este símbolo representa a soma dos valores, y, onde o índice i representa um valor típico no intervalo de 1 a n. Para ilustrar, para os dados do Leste Europeu, tem-se:

$$\sum y_i = y_1 + y_2 + \cdots + y_8 =$$
$$83 + 82 + \cdots + 81 = 643.$$

O símbolo Σ é, algumas vezes, até mesmo abreviado como Σy. Usando este símbolo de somatório, temos a expressão abreviada para a média amostral das n observações.

$$\bar{y} = \frac{\Sigma y_i}{n}.$$

Propriedades da média

Aqui estão algumas propriedades da média:

- A fórmula para a média usa valores numéricos para as observações. Assim a média é apropriada somente para variáveis quantitativas. Não é sensato calcular a média para as observações em uma escala nominal. Por exemplo, para a religião mensurada em categorias como (protestante, católica, judaica, outra), a religião média não faz sentido, embora esses níveis, algumas vezes, possam estar representados por números por conveniência. De forma similar, não podemos encontrar a média das observações em uma avaliação ordinal como excelente, bom, regular e ruim, a não ser que designemos números como 4, 3, 2, 1 aos níveis ordenados, tratando-os como quantitativos.
- A média pode ser grandemente influenciada por uma observação que esteja bem acima ou bem abaixo da grande maioria dos dados, chamada de **valor atípico**.

EXEMPLO 3.5 Efeito de um valor atípico na renda média

O proprietário da Pizzaria Leonardo relata que a renda anual dos empregados no negócio é de \$40900. Na verdade, os rendimentos anuais dos sete empregados são de \$11200, \$11400, \$11700, \$12200, \$12300, \$12500 e \$215000. O rendimento de \$215000 é o salário do filho do proprietário, que também é um empregado. A média calculada para as outras seis observações somente é igual a \$11883, bem diferente da média de \$40900 incluindo o valor atípico. ■

Este exemplo mostra que a média não é sempre típica das observações da amostra. Isto comumente acontece com amostras pequenas quando pelo menos uma observação é muito maior ou muito menor do que as outras, como em distribuições altamente assimétricas.

- A média é atraída na direção da cauda mais longa de uma distribuição assimétrica, relativa à maioria dos dados. No Exemplo 3.5, a observação maior de \$215000 resulta em uma assimetria extrema à direita da distribuição das rendas. Essa assimetria atrai a média acima de seis das sete observações. Em geral, quanto mais altamente assimétrica é a distribuição, menos típica é a média dos dados.
- A média é o ponto de equilíbrio da linha dos valores quando um peso igual está em cada ponto de observação.

Por exemplo, a Figura 3.9 mostra que, se um peso igual for colocado em cada observação do Leste Europeu na atividade econômica feminina do Exemplo 3.4, então a linha se equilibra colocando um fulcro no ponto 80,4. A média é o *centro de gravidade* (ponto de equilíbrio) das observações. Isso significa que a soma das distâncias para média das observações *acima* da média é igual a soma das distâncias para a média das observações *abaixo* da média.

- Represente as médias amostrais para dois conjuntos de dados com tamanhos amostrai n_1 e n_2 por $\bar{y}_1$ e $\bar{y}_2$. A média amostral total para o conjunto combinado de observações $(n_1 + n_2)$ é a **média ponderada.**

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

O numerador $n_1\bar{y}_1 + n_2\bar{y}_2$ é a soma de todas as observações, desde que $n\bar{y} = \Sigma y$ para cada conjunto de observações. O denominador é o tamanho da amostra.

Para ilustrar, considere os dados da atividade econômica feminina da Tabela 3.4, as observações da América do Sul têm $n_1 = 11$ e $\bar{y}_1 = 52{,}1$ e as observações do Leste Europeu têm $n_2 = 8$ e $\bar{y}_2 = 80{,}4$. A atividade econômica geral para as 19 nações é igual a:

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{11(52{,}1) + 8(80{,}4)}{11 + 8} = \frac{(573 + 643)}{19} = \frac{1216}{19} = 64.$$

A média ponderada de 64 está mais próxima de 52,1, o valor para a América do Sul, do que 80,4, o valor para o Leste Europeu. Isto acontece porque mais observações vêm da América do Sul do que do Leste Europeu.

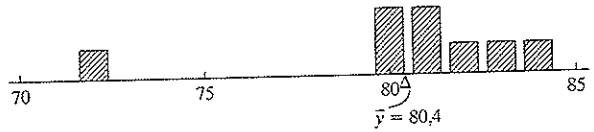
A mediana

A média é uma medida simples de centro. Mas outras medidas são também informativas e, algumas vezes, mais apropriadas. A mais importante é a *mediana*. Ela divide a amostra em duas partes com um número igual de observações, quando elas estão ordenadas da mais baixa para a mais alta.

> ☑ **Mediana**
> A mediana é a observação que está no centro da amostra ordenada. Quando o tamanho da amostra n é ímpar, existe uma única observação central. Quando o tamanho da amostra é par, existem duas observações centrais e a mediana é a média entre as duas.

Para ilustrar, as observações ordenadas da renda para os sete empregados no Exemplo 3.5 são

\$11200, \$11400, \$11700, \$12200, \$12300, \$12500, \$215000.



☑ **Figura 3.9** A Média como o centro de gravidade, para os dados do Leste Europeu do exemplo 3.4. A linha se equilibra sobre um fulcro em 80,4.

A mediana é a observação do centro, \$12200. Este é um valor mais típico para esta amostra do que a média amostral de \$40900. Quando uma distribuição é altamente assimétrica, a mediana é um valor típico melhor do que a média.

Na Tabela 3.4, os valores ordenados da atividade econômica para as nações do Leste Europeu são:

72, 80, 80, 81, 81, 82, 83, 84.

Visto que $n = 8$ é par, a mediana é a média entre os dois valores centrais, 81 e 81, que é $(81 + 81)/2 = 81$. Este resultado está próximo da média amostral de 80,4, porque este conjunto de dados não apresenta valores atípicos.

A posição da observação central é $(n + 1)/2$. Isto é, a mediana é o valor da observação $(n + 1)/2$ na amostra ordenada. Quando $n = 7$, $(n + 1)/2 = (7 + 1)/2 = 4$, assim a mediana é a quarta observação menor, ou equivalentemente a quarta observação maior. Quando n for par, $(n + 1)/2$ está entre os valores centrais e a mediana é o centro das observações com aqueles índices. Por exemplo, quando $n = 8$, $(n + 1)/2 = 4{,}5$, assim a mediana é o ponto central entre a quarta e quinta menores observações.

EXEMPLO 3.6 Mediana para dados agrupados ou ordinais

A Tabela 3.5 resume a distribuição do mais alto grau educacional da população norte-americana de 25 anos ou mais, como estimado pelo Levantamento de Dados da Comunidade Americana de 2005 feito pela Agência do Censo dos Estados Unidos. As respostas possíveis formam uma escala ordinal. O tamanho da população era de $n = 189$ (em milhões). O escore mediano é o $(n + 1)/2 = (189 + 1)/2 = 95^{o}$ mais baixo. Temos 30 respostas na primeira categoria, $(30 + 56) = 86$ nas duas primeiras, $(30 + 56 + 38) = 124$ nas três primeiras e assim por diante. Os 87^{o} ao 124^{o} escores mais baixos estão na categoria 3, que, portanto, contém o 95^{o} mais baixo, que é a mediana. A resposta mediana é "Superior Incompleto". De forma equivalente, dos percentuais da última coluna da tabela $(15{,}9\% + 29{,}6\%) = 45{,}5\%$ está nas primeiras categorias e $(15{,}9\% + 29{,}6\% + 20{,}1\%) = 65{,}6\%$ está nas primeiras três, assim o ponto central de 50% está na terceira categoria. ■

Propriedades da mediana

- A mediana, assim como a média, é apropriada para variáveis quantitativas. Visto que requer somente observações ordenadas para calculá-la, ela também é válida para dados de escala ordinal, como mostrou o exemplo anterior. Não é apropriada para dados de escala nominal, já que as observações não podem ser ordenadas.
- Para distribuições simétricas, como a da Figura 3.7, a mediana e a média são idênticas. Para ilustrar, a amostra das observações 4, 5, 7, 9, 10 é simétrica sobre 7; 5 e 9 estão igualmente distantes dela em direções opostas, como o 4 e 10. Portanto, 7 é tanto a média quanto a mediana.
- Para distribuições assimétricas, a média está situada na direção da assimetria da curva (a cauda mais longa) em relação à mediana. Veja Figura 3.10.

☑ **Tabela 3.5** Grau educacional completo mais alto, para uma amostra de norte-americanos

Grau educacional mais alto	Frequência (milhões)	Percentual
Ensino médio incompleto	30	15,9
Somente ensino médio	56	29,6
Superior incompleto	38	20,1
Grau universitário de dois anos	14	7,4
Graduado	32	16,9
Mestrado	13	6,9
Doutorado ou profissional	6	3,2

Por exemplo, considere a taxa de crimes violentos da Tabela 3.2. A mediana é 36,5. A média é $\bar{y} = 40{,}2$, um pouco maior do que a mediana. A Figura 3.2 mostrou que os valores da taxa de crimes violentos são assimétricos à direita. A média é maior do que a mediana para distribuições que são assimétricas à direita. As distribuições do rendimento tendem a ser assimétricas à direita. Por exemplo, a renda por domicílio nos Estados Unidos em 2005 tinha uma média de aproximadamente \$61000 e uma mediana de aproximadamente \$44000 (Agência do Censo dos Estados Unidos).

A distribuição das notas em um exame tende a ser assimétrica à esquerda quando alguns estudantes têm um desempenho consideravelmente pior do que outros. Neste caso, a média é menor do que a mediana. Por exemplo, suponha que as notas de um exame que variaram de 0 a 100 têm uma mediana de 88 e uma média de 76. A maioria dos estudantes teve um bom desempenho (a metade acima de 88), mas aparentemente alguns escores foram muito mais baixos para baixar a média para 76.

- A mediana é insensível às distâncias das observações do centro, visto que usa somente características ordinais dos dados. Por exemplo, todos os quatro conjuntos seguintes de observações têm medianas de 10:

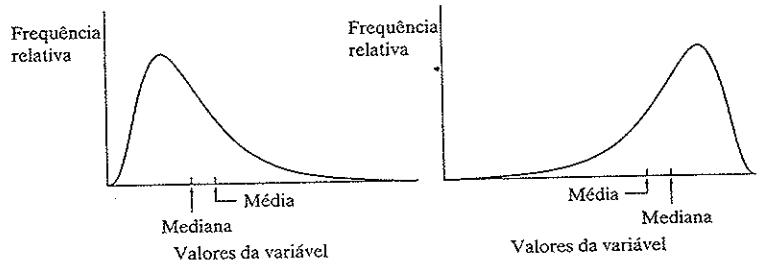
  Conjunto 1: 8, 9, 10, 11, 12
  Conjunto 2: 8, 9, 10, 11, 100
  Conjunto 3: 0, 9, 10, 10, 10
  Conjunto 4: 8, 9, 10, 100, 100

- A mediana não é afetada pelos valores atípicos. Por exemplo, as rendas dos sete empregados no Exemplo 3.5 têm uma mediana de \$12200 quer a observação maior seja \$20000, \$215000 ou \$2000000.

A mediana comparada à média

A mediana é geralmente mais apropriada do que a média quando a distribuição for altamente assimétrica, como observamos com os rendimentos dos empregados da Pizzaria Leonardo. A média pode ser fortemente afetada pelos valores atípicos, enquanto a mediana não.

Para a média, precisamos de dados quantitativos (escala intervalar). A mediana também pode ser calculada em uma escala ordinal (veja Exemplo 3.6). Para usar a média para dados ordinais, precisamos



☑ **Figura 3.10** A Média e a Mediana para distribuições assimétricas. A média é puxada na direção da cauda mais longa.

atribuir escores às categorias. Na Tabela 3.5, se atribuirmos os escores 10, 12, 13, 14, 16, 18, 20 às categorias com os graus educacionais mais altos, representando o número aproximado de anos de educação, obtemos uma média amostral de 13,4.

A mediana tem algumas desvantagens. Para dados discretos que assumem relativamente poucos valores, padrões bem diferentes de dados podem ter a mesma mediana. Por exemplo, a Tabela 3.6, de um PSG, resume as 365 respostas de mulheres à pergunta: "Quantos parceiros sexuais você teve nos últimos 12 meses?". Somente seis respostas distintas ocorreram e 63,8% delas são 1. A resposta mediana é 1. Para encontrar a média amostral, multiplicamos cada possível valor das 365 observações por sua frequência de ocorrência e, então, somamos. Isto é,

$$\sum y_i = 102(0) + 233(1) + 18(2) + 9(3) + 2(4) + 1(5) = 309.$$

A média amostral das respostas é

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{309}{365} = 0{,}85.$$

Se a distribuição das 365 observações entre estas categorias eram (0, 233, 18, 9, 2, 103) (isto é, nós trocamos as 102 respostas de 0 a 5), então a mediana seria 1, mas a média mudaria para 2,2. A média usa os valores das observações, não apenas suas posições (postos).

A forma mais extrema deste problema ocorre para **dados binários**, que podem assumir somente dois valores, como 0 e 1. A média se iguala ao resultado mais comum, mas não fornece informação sobre o número relativo de observações nos dois níveis. Por exemplo, considere amostras de tamanho 5 para a variável número de casamentos. As amostras (1, 1, 1, 1, 1) e (0, 0, 1, 1, 1) têm a mesma mediana de 1. Já a média é 1 para a amostra (1, 1, 1, 1, 1) e 3/5 para a amostra (0, 0, 1, 1, 1). *Quando as observações assumem apenas os valores 0 e 1, a média é igual a proporção das observações que é igual a 1.* Geralmente, para dados discretos a média é mais informativa do que a mediana.

Em resumo,

- Se uma distribuição for altamente assimétrica, a mediana é geralmente preferida porque ela representa melhor o que é típico.
- Se a distribuição for quase simétrica ou levemente assimétrica ou se ela for discreta com poucos valores distintos, a média é mais recomendada porque ela usa os valores de todas as observações.

A moda

Outra medida, a *moda*, indica o resultado que é mais comum.

☑ Moda

A **moda** é o valor que ocorre com maior frequência.

A moda é mais comumente usada com variáveis discretas como com dados categóricos. Por exemplo, na Tabela 3.5, no grau educacional completo mais alto a moda é "Somente Ensino Médio", visto que a frequência para aquela categoria é maior do que a frequência para qualquer outra categoria. Na Tabela 3.6, que considera o número de parceiros sexuais do ano anterior, a moda é 1.

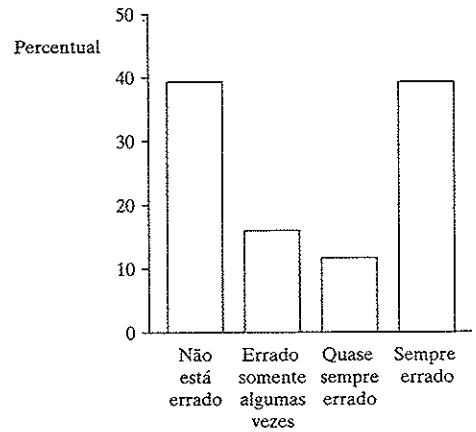
☑ **Tabela 3.6** Número de parceiros sexuais ano passado, para respondentes femininos na PSG

Respostas	Frequência	Percentual
0	102	27,9
1	233	63,8
2	18	4,9
3	9	2,5
4	2	0,5
5	1	0,3

Propriedades da moda

- A moda é apropriada para todos os tipos de dados. Por exemplo, podemos mensurar a moda para a religião na Austrália (escala nominal), para a avaliação de um professor (escala ordinal) ou o número de anos completos de educação de norte-americanos de origem hispânica (escala intervalar).
- Uma distribuição de frequências é denominada **bimodal** se ela tiver dois picos distintos de igual tamanho. As distribuições bimodais geralmente ocorrem com variáveis atitudinais quando as populações são polarizadas, com respostas sendo fortemente orientadas em uma direção ou outra. Por exemplo, a Figura 3.11 mostra a distribuição da frequência relativa das respostas da Pesquisa Social Geral à pergunta: "Você particularmente acha que é errado ou não é errado uma mulher fazer um aborto se a família tem um rendimento muito baixo e não pode sustentar mais uma criança?". As frequências relativas nas duas categorias extremas são mais altas do que as categorias do centro.
- A média, a mediana e a moda são idênticas para uma distribuição unimodal e simétrica como a distribuição em forma de sino.

A média, a mediana e a moda são medidas complementares. Elas descrevem aspectos diferentes dos dados. Em qualquer exemplo, alguns ou todos os valores podem ser úteis. Tome cuidado com análises estatísticas equivocadas, como usar uma estatística quando outra seria mais informativa. As pessoas que apresentam conclusões estatísticas geralmente escolhem a estatística dando a impressão de que elas querem transmitir uma ideia. Lembre-se do Exemplo 3.5 (p. 58) dos empregados da Pizzaria Leonardo com uma observação da renda muito extrema. Tenha cuidado com a média quando a distribuição é altamente assimétrica.



☑ **Figura 3.11** Distribuição bimodal para a opinião sobre se o aborto é errado.

3.3 DESCREVENDO A VARIABILIDADE DOS DADOS

Uma medida do centro isolada não é adequada para descrever dados numericamente para uma variável quantitativa. Ela descreve um valor típico, mas não a dispersão dos dados em torno do valor típico. As duas distribuições na Figura 3.12 ilustram a situação. Os cidadãos da nação A e os da nação B têm a mesma renda média anual ($25000). Contudo, as distribuições dessas rendas diferem fundamentalmente, com a renda da nação B sendo muito menos variável. Uma renda de $30000 é extremamente alta para a nação B, mas não especialmente alta para a nação A. Esta seção introduz estatísticas que descrevem a variabilidade de um conjunto de dados.

O intervalo ou amplitude

A diferença entre as observações maiores e menores é a forma mais simples de descrever a variabilidade.

☑ **Intervalo ou amplitude**

O **intervalo ou amplitude** é a diferença entre o maior e o menor valor de todas as observações.

Para a nação A, da Figura 3.12, o intervalo dos valores da renda é aproximadamente $50000 – 0 = $50000. Para a nação B, o intervalo é aproximadamente $30000 – $20000 = $10000. A nação A tem a maior variabilidade de rendas.

O intervalo não é, contudo, sensível a outras características da variabilidade dos dados. As três distribuições da Figura 3.13 têm a mesma média ($25000) e amplitude ($50000), mas elas diferem na forma da variabilidade dos valores em torno do centro. Em termos de distâncias das observações em relação à média, a nação A tem a maior variabilidade e a nação B a menor. As rendas na nação A tendem a estar mais longe da média e as rendas na nação B tendem a estar mais perto.

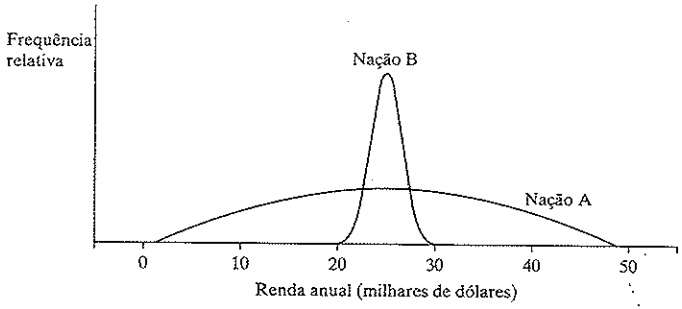
O desvio padrão

Outras medidas da variabilidade são baseadas nos desvios dos dados em relação a uma medida de centro, como a própria média.

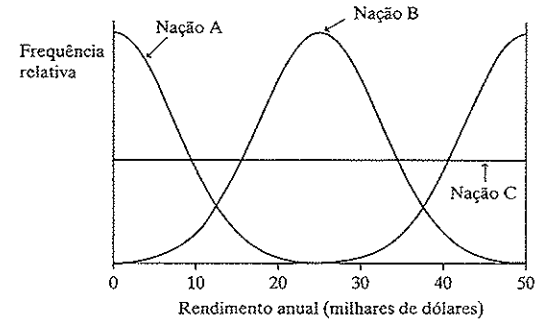
☑ **Desvio**

O **desvio** de uma observação y_i em relação à média da amostra $\bar{y}$ é $(y_i - \bar{y})$, a diferença entre os dois valores.

Cada observação tem um desvio. O desvio é *positivo* quando a observação está *acima* da média. O desvio é *negativo* quando a observação está *abaixo* da média. A interpretação de $\bar{y}$ como o centro da gravidade



☑ **Figura 3.12** Distribuições com a mesma média, mas com diferentes variabilidades.



☑ **Figura 3.13** Distribuições com a mesma média e amplitude, mas com diferentes variabilidades em relação à média.

dos dados implica que a soma dos desvios positivos seja igual à soma dos desvios negativos. Portanto, a soma de todos os desvios em relação à média, $\Sigma(y_i - \bar{y})$, é igual a 0. Em virtude disto, as medidas de variabilidade usam tanto os valores absolutos quanto os quadrados dos desvios. A mensuração mais popular usa os quadrados.

☑ **Desvio padrão**

O **desvio padrão** s de n observações é dado por:

$$s = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{\text{soma dos desvios ao quadrado}}{\text{tamanho da amostra -1}}}.$$

Isto é a raiz quadrada da **variância** s^2, que é dada por:

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n - 1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}$$

A *variância* é aproximadamente uma média dos desvios ao quadrado. As unidades de mensuração são os quadrados dos desvios em relação à média dos dados. Isto torna a variância difícil de ser interpretada. Por isso, usamos a sua raiz quadrada, o *desvio padrão.*

A expressão $\Sigma(y_i - \bar{y})^2$ nestas fórmulas é chamada de **soma dos quadrados.** Ela representa elevar ao quadrado cada desvio e, então, somar esses quadrados. É incorreto primeiro somar os desvios e, depois, elevar a soma ao quadrado, mesmo porque isto dará um valor igual a 0. Quanto maior forem os desvios, maior será a soma dos quadrados e, portanto, maior o valor de s.

Embora sua fórmula pareça ser complicada, a interpretação básica do desvio padrão s é muito simples: o s é um tipo de *distância típica* de uma observação em relação à média. Assim, quanto maior for o desvio padrão s, maior a dispersão dos dados.

EXEMPLO 3.7 Comparando a variabilidade de resultados de testes

Cada um dos seguintes conjuntos de resultados de testes de duas pequenas amostras de estudantes apresentam uma média de 5 e uma amplitude de 10:

Amostra 1: 0, 4, 4, 5, 7, 10
Amostra 2: 0, 0, 1, 9, 10, 10.

Por inspeção, a amostra 1 mostra menos variação em relação à média do que a amostra 2. A maioria dos escores da amostra 1 está próxima da média que é 5, enquanto todos os escores da amostra 2 estão distantes da média que, também, é 5.

Para a amostra 1,

$$\Sigma(y_i - \bar{y})^2 = (0 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (7 - 5)^2 + (10 - 5)^2 = 56,$$

assim, a variância é igual a

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n - 1} = \frac{56}{6 - 1} = \frac{56}{5} = 11{,}2.$$

O desvio padrão para a amostra 1 é igual $s = \sqrt{11{,}2} = 3{,}3$. Para a amostra 2, você pode verificar que $s^2 = 26{,}4$ e $s = \sqrt{26{,}4} = 5{,}1$. Visto que $3{,}3 < 5{,}1$, os desvios padrão nos dizem que a amostra 1 é menos variável do que a amostra 2. ■

Softwares estatísticos e calculadoras podem encontrar o desvio padrão. Você deve fazer os cálculos para alguns conjuntos de dados pequenos para perceber o que esta medida representa. A resposta que você achar pode diferir levemente do valor obtido pelo *software*, dependendo dos arredondamentos que tenha feito na realização dos cálculos.

Propriedades do desvio padrão

- $s \geq 0$.
- $s = 0$ somente quando todas as observações têm o mesmo valor. Por exemplo, se as idades para uma amostra de cinco estudantes são 19, 19, 19, 19, 19, então a média da amostra é igual a 19, cada um dos cinco desvios valem 0 e assim $s = 0$. Esta é a variabilidade mínima possível.
- Quanto maior a variabilidade em relação à média, maior é o valor de s. Por exemplo, a Figura 3.5 mostra que as taxas de assassinato são muito mais variáveis entre os estados dos Estados Unidos do que entre as províncias canadenses. Na verdade, os desvios padrão são $s = 4{,}0$ para os Estados Unidos e $s = 0{,}8$ para o Canadá.
- A razão para usar $(n - 1)$, em vez de n, no denominador de s (e s^2) é técnica e diz respeito à inferência sobre os parâmetros da população discutidos no Capítulo 5. Quando temos dados para toda uma população substituímos $(n - 1)$ pelo tamanho real da população; a variância da população é, então, precisamente a média dos desvios ao quadrado. Neste caso, o desvio padrão não pode ser maior do que metade da amplitude.
- Se os dados são redimensionados, o desvio padrão é também redimensionado. Por exemplo, se mudarmos os rendimentos anuais de dólares (como 34000) para milhares de dólares (como 34,0), o desvio padrão também muda pelo mesmo fator de 1000 (assim ele passa de 11800 para 11,8).

Interpretando a magnitude de *s*

Uma distribuição com $s = 5{,}1$ tem uma variabilidade maior do que uma com $s = 3{,}3$, mas como interpretamos *quão grande é* $s = 5{,}1$? Vimos que uma resposta aproximada é que s é uma distância típica de uma observação da média. Para ilustrar, suponha que a primeira prova no seu curso, classificada em uma escala de 0 a 100, tem uma média amostral de 77. Um valor de $s = 0$ é improvável, visto que todos os estudantes devem, então, ter um escore de 77. Um valor com $s = 50$ parece improvavelmente grande para uma distância típica da média. Valores de s como 8 ou 12 parecem ser muito mais realistas.

Formas mais precisas de interpretar s requerem conhecimento adicional da forma da distribuição da frequência. A regra seguinte fornece uma interpretação para muitos conjuntos de dados.

☑ Regra Empírica

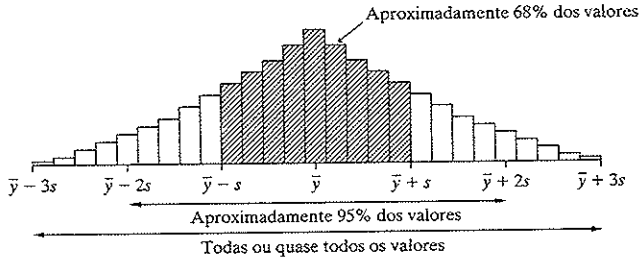
Se o histograma dos dados tem uma forma aproximada de sino, então:

1. Aproximadamente 68% das observações estão entre $\bar{y} - s$ e $\bar{y} + s$.
2. Aproximadamente 95% das observações estão entre $\bar{y} - 2s$ e $\bar{y} + 2s$.
3. Todas ou quase todas as observações estão entre $\bar{y} - 3s$ e $\bar{y} + 3s$.

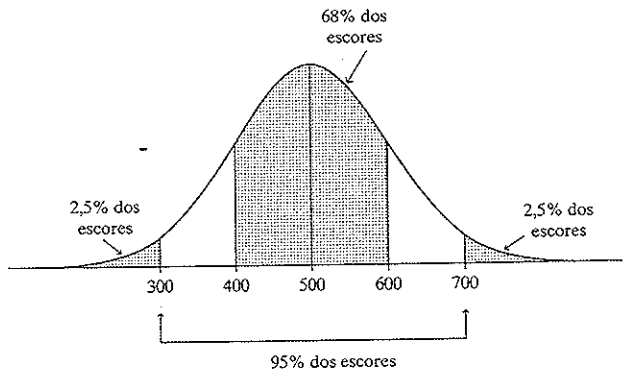
A regra é chamada de Regra Empírica porque muitas distribuições vistas na prática (isto é, *empiricamente*) têm uma forma aproximada de sino. A Figura 3.14 é uma descrição gráfica da regra.

EXEMPLO 3.8 Descrevendo a distribuição dos escores do SAT

O SAT (Scholastic Aptitude Test – Teste de Aptidão Acadêmica, veja www.collegeboard.com) é composto de três partes: leitura crítica, matemática e redação. Para cada parte, a distribuição dos escores tem aproximadamente a forma de sino. Cada parte tem uma média de aproximadamente 500 e um desvio padrão de aproximadamente 100. A Figura 3.15 representa isto. Pela Regra Empírica, para cada parte, aproximadamente 68% dos escores está entre 400 e 600,



☑ **Figura 3.14** Regra Empírica: interpretação do desvio padrão para uma distribuição com forma de sino.



☑ **Figura 3.15** Uma distribuição com forma de sino dos escores para uma parte do SAT, com média de 500 e desvio padrão de 100.

porque 400 e 600 são os números que estão *um* desvio padrão abaixo e acima da média 500. Aproximadamente 95% dos escores estão entre 300 e 700, valores que estão a *dois* desvios padrão da média. Os 5% restantes estão ou abaixo de 300 ou acima de 700. A distribuição é aproximadamente simétrica em torno de 500, assim aproximadamente 2,5% dos escores estão acima de 700 e aproximadamente 2,5% estão abaixo de 300. ■

A Regra Empírica se aplica somente a distribuições que têm aproximadamente a forma de sino. Para outras formas, o percentual que está entre dois desvios padrão da média não precisa estar próximo de 95%. Poderia estar tão baixo quanto 75% ou tão alto quanto 100%. A Regra Empírica pode não funcionar bem se a distribuição for altamente assimétrica ou se ela for muito discreta, com a variável assumindo poucos valores. Os percentuais exatos dependem da forma da distribuição, como mostra o próximo exemplo.

EXEMPLO 3.9 Convivência com vítimas da AIDS

Uma PSG perguntou: "Quantas pessoas você conhece pessoalmente, vivas ou mortas, que contraíram AIDS?". A Tabela 3.7 mostra parte de uma saída de computador resumindo as 1598 respostas desta variável. Ela indica que 76% das respostas foram 0.

A média e o desvio padrão são $\bar{y} = 0{,}47$ e $s = 1{,}09$. Os valores 0 e 1 estão a menos de um desvio padrão da média. Agora, 88,8% da distribuição está representada por estes dois pontos ou entre $\bar{y} \pm s$. Isto é consideravelmente maior do que os 68% que a Regra Empírica declara. A Regra Empírica não se aplica a esta distribuição porque ela não tem, nem mesmo aproximadamente, a forma de sino. Ao contrário, é altamente assimétrica à direita, como você pode verificar traçando um histograma para a Tabela 3.7. O valor menor na distribuição (0) é menor do que um desvio padrão abaixo da média; o valor maior na distribuição (8) está aproximadamente a sete desvios padrão acima da média. ■

Sempre que a observação menor ou maior for menor do que um desvio padrão da média, isto é uma evidência de severa assimetria. Por exemplo, um exame recente de estatística tendo uma escala de 0 a 100 teve $\bar{y} = 86$ e $s = 15$. O limite superior de 100 é menor do que um desvio padrão acima da média. A distribuição é altamente assimétrica à esquerda.

O desvio padrão, como a média, pode ser muito afetado por um valor atípico, especialmente para conjunto de pequenos de dados. Por exemplo, as taxas de assassinatos mostradas na Figura 3.5 para os 50 estados dos Estados Unidos têm $\bar{y} = 7{,}3$ e $s = 4{,}0$. A distribuição é um pouco irregular, mas 68% dos estados têm taxas de assassinatos dentro de um desvio padrão da média e 98% dentro de dois desvios padrão. Suponha, agora, que incluímos a taxa de assassinatos para o Distrito de Columbia que é igual a 78,5 no conjunto de dados. Então, $\bar{y} = 8{,}7$ e $s = 10{,}7$. O desvio padrão mais do que duplica. Agora, 96,1% das taxas de assassinato (todas exceto D.C. [Distrito de Columbia, Washington capital] e Louisiana) estão dentro de um desvio padrão da média.

3.4 MEDIDAS DE POSIÇÃO

Outra forma de descrever uma distribuição é com uma medida de **posição**. Ela nos diz o ponto no qual um certo percentual dos dados está abaixo (ou acima) daquele ponto. Como nos casos especiais, algumas medidas de posição descrevem o centro e algumas descrevem a variabilidade.

Quartis e outros percentis

O intervalo ou amplitude usa duas medidas de posição, o valor máximo e o valor mínimo. A mediana é uma medida de posição, com metade dos dados estando abaixo dela e outra metade acima. A mediana é um caso especial de um conjunto de medidas de posição chamado de *percentis*.

☑ **Tabela 3.7** Uma distribuição de frequências do número de pessoas com AIDS conhecidas pessoalmente*

AIDS	Frequência	Percentual
0	1214	76,0
1	204	12,8
2	85	5,3
3	49	3,1
4	19	1,2
5	13	0,8
6	5	0,3
7	8	0,5
8	1	0,1

N	1598
Média	0,47
Desvio Padrão	1,09

☑ **Percentil**

O **enésimo percentil** é o ponto tal que p% das observações estão abaixo ou naquele ponto e $(100 - p)$% estão acima dele.

Substituindo $p = 50$ nesta definição temos o 50º percentil. Isto é a mediana. A mediana é maior do que 50% das observações e menor do que os outros $(100 - 50) = 50$%. Os dois outros percentis mais usados são o *quartil inferior* e o *quartil superior*.

☑ **Quartis inferior e superior**

O 25º percentil é chamado de **quartil inferior.** O 75º percentil é chamado de **quartil superior.** Um quarto dos dados está abaixo do quartil inferior. Um quarto está acima do quartil superior.

Os quartis resultam de $p = 25$ e $p = 75$ na definição do percentil. O quartil inferior é a mediana para as observações que estão abaixo da mediana, isto é, para a metade inferior dos dados. O quartil superior é a mediana para as observações que estão acima da mediana, isto é, para a metade superior dos dados. Os quartis junto com a mediana dividem a distribuição em quatro partes, cada uma contendo um quarto das observações. Veja a Figura 3.16.

Para as taxas de crimes violentos da Tabela 3.2, o tamanho da amostra é $n = 50$ e a mediana é igual a 36,5. Assim como com a mediana, os quartis podem ser encontrados facilmente do diagrama de caule e folhas dos dados (Figura 3.4), que era:

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

* N. de T. T.: Para facilitar a leitura, as tabelas com as saídas do SPSS e SAS foram traduzidas em sua maioria, contudo ainda não existem versões desses *softwares* que fornecem saídas em português.

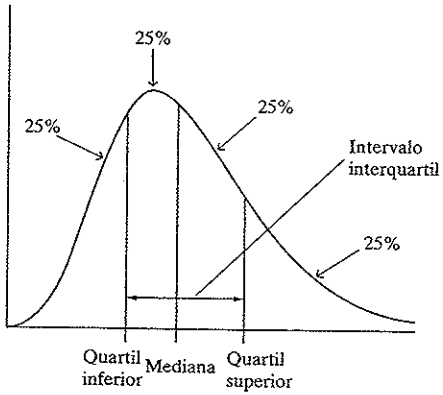


Figura 3.16 Os quartis e o intervalo interquartil.

O quartil inferior é a mediana para as 25 observações abaixo da mediana. É a 13ª observação menor ou 27. O quartil superior é a mediana para as 25 observações acima da mediana. É a 13ª observação maior ou 51.

Em resumo, desde que

quartil inferior = 27, mediana = 36,5,
quartil superior = 51,

aproximadamente um quarto dos estados tinha taxas de crimes violentos (i) abaixo de 27, (ii) entre 27 e 36,5, (iii) entre 36,5 e 51 e (iv) acima de 51. A distância entre o quartil superior e a mediana, 51 – 36,5 = 14,5, excede a distância de 36,5 – 27 = 9,5 entre o quartil inferior e a mediana. Isso geralmente acontece quando a distribuição é assimétrica à direita.

Um *software* pode facilmente encontrar quartis assim como outros percentis. Na prática, os percentis, com exceção da mediana, não são geralmente relatados para conjunto de pequenos dados.

Mensurando a variabilidade: o intervalo interquartil (IIQ)

A diferença entre os quartis superiores e inferiores é chamada de **intervalo interquartil**, denominado de IIQ. Essa medida descreve a dispersão da metade das observações. Para as taxas de crimes violentos dos Estados Unidos da Tabela 3.2, o intervalo interquartil é IIQ = 51 – 27 = 24. A metade das taxas de assassinatos está entre um intervalo de 24. Como o intervalo e o desvio padrão, o IIQ aumenta à medida que a variabilidade aumenta e ele é útil para comparar a variabilidade de grupos diferentes. Por exemplo, 12 anos atrás, em 1993, os quartis das taxas de crimes violentos dos estados dos Estados Unidos eram 33 e 77, dado um IIQ de 77 – 33 = 44 e mostrando muito mais variabilidade.

Uma vantagem do IIQ sobre o intervalo comum ou desvio padrão é que ele não é suscetível aos valores atípicos. As taxas de crimes violentos dos Estados Unidos têm um intervalo de 8 a 79, assim o intervalo é 71. Quando incluímos as observações para D.C., que era de 161, o IIQ muda somente de 24 a 28. Em contraposição, o intervalo muda de 71 a 153.

Para as distribuições com forma de sino, a distância da média para qualquer quartil é aproximadamente 2/3 de um desvio padrão. Então o IIQ é aproximadamente (4/3)s. A indiferença do IIQ em relação aos valores atípicos aumentou recentemente a sua popularidade, embora na prática o desvio padrão ainda seja mais comum.

Diagramas de caixa e bigodes: fazendo um gráfico das posições do resumo dos cinco números

A mediana, os quartis e o máximo e mínimo são as cinco posições geralmente usadas como um conjunto para descrever o centro e a dispersão. Por exemplo, um *software* relata o seguinte resumo dos cinco números para as taxas de crimes violentos (onde Q1 = quartil inferior, Q3 = quartil superior, considerando a mediana como o segundo quartil):

```
100% Max      79,0
 75% Q3       51,0
 50% Med      36,5
 25% Q1       27,0
  0% Min       8,0
```

O resumo de cinco números fornece uma descrição simples dos dados. Ele é a base de uma representação gráfica denominada de **diagrama de caixa e bigodes** que resume o centro e a variabilidade. A *caixa* do diagrama contém os 50% centrais da distribuição, do quartil inferior ao quartil superior. A mediana é marcada por uma linha dentro da caixa. As linhas que se estendem da caixa são chamadas de *bigodes*. Elas se estendem ao máximo e mínimo, exceto para os valores atípicos, que são marcados separadamente.

A Figura 3.17 mostra o diagrama de caixa e bigode para as taxas de crimes violentos, no formato fornecido com o *software* SPSS. O bigode superior e a metade superior da caixa central são mais longos do que os inferiores. Isto indica que a cauda da direita da distribuição, que corresponde aos valores relativamente maiores, é mais longa do que a cauda da esquerda. O diagrama reflete a assimetria à esquerda das taxas de crimes violentos. (Alguns *softwares* também um traçam a média no diagrama de caixa e bigodes, representando-a por um sinal de +.)

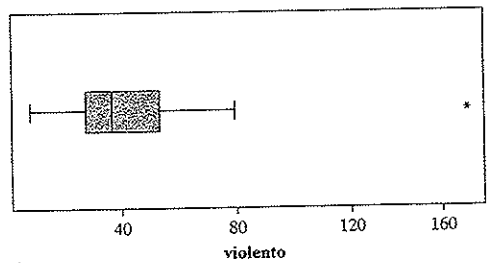
Diagramas de caixa e bigodes lado a lado são úteis para comparar duas distribuições. A Figura 3.5 mostrou diagramas de caule e folhas lado a lado das taxas de assassinato dos Estados Unidos e Canadá. A Figura 3.18 mostra os diagramas de caixa e bigodes lado a lado. Estes diagramas revelam que as taxas de assassinato nos Estados Unidos tendem a ser mais altas e têm maior variabilidade.

Valores atípicos

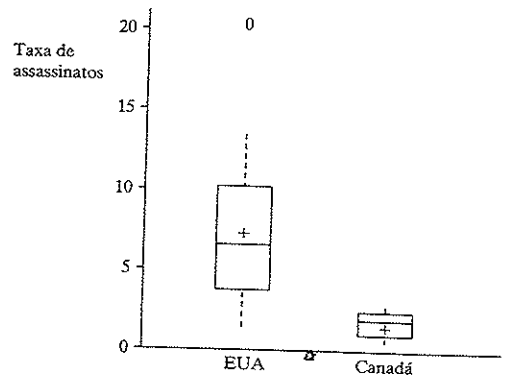
Os diagramas de caixa e bigodes identificam os valores atípicos separadamente. Para explicar isto, apresentamos agora uma definição formal de um valor atípico.

☑ **Valor atípico**

Uma observação é um **valor atípico** se estiver a mais do que 1,5 (IIQ) acima do quartil superior ou a mais do que 1,5 (IIQ) abaixo do quartil inferior.



☑ **Figura 3.17** Diagrama de caixa e bigodes para as taxas de crimes violentos dos Estados Unidos e D.C.



☑ **Figura 3.18** Diagramas de caixa e bigodes para as taxas de assassinatos dos Estados Unidos e Canadá.

Nos diagramas de caixa e bigodes, os bigodes se estendem às observações menores e maiores somente se estes valores não forem valores atípicos; isto é, se eles não estiverem a mais do que 1,5 (IIQ) além dos quartis. Por outro lado, os bigodes se estendem às observações mais extremas dentro de 1,5 (IIQ) e os valores atípicos são marcados separadamente. Por exemplo, o *software* estatístico SAS assinala com um O (O para *outlier* – valor atípico) um valor entre 1,5 e 3,0 (IIQ) do diagrama e com um asterisco (*) um valor mais distante.

A Figura 3.18 mostra um valor atípico para os Estados Unidos com uma taxa de assassinato muito alta. Esta é a taxa de assassinato de 20,3 (da Louisiana). Para estes dados, o quartil inferior é 3,9 e o quartil superior é 10,3, assim o $IIQ = 10,3 - 3,9 = 6,4$. Portanto,

$$\text{Quartil superior} = 1,5(\text{IIQ}) = 10,3 + 1,5(6,4) = 19,9.$$

Visto que $20,3 > 19,9$, o diagrama de caixa e bigodes destaca a taxa de 20,3 como um valor atípico.

Por que destacar valores atípicos? Pode ser informativo investigá-los. A observação foi, talvez, registrada incorretamente? O sujeito era fundamentalmente, de alguma forma, diferente dos outros? Geralmente faz sentido repetir a análise estatística sem um valor atípico para ter certeza de que as conclusões não estão muito suscetíveis a uma única observação. Outra razão para mostrar os valores atípicos separadamente em um diagrama de caixa e bigodes é que eles não fornecem muita informação sobre a forma da distribuição, especialmente para conjuntos de dados grandes.

Na prática, o critério de 1,5(IIQ) para um valor atípico é, de alguma forma, arbitrário. É melhor considerar uma observação satisfazendo esse critério como um valor atípico *potencial* em vez de um valor atípico definido. Quando uma distribuição tem uma cauda direita longa, algumas observações podem estar a mais do que 1,5 IIQ acima do quartil superior mesmo se elas não estão separadas do total dos dados.

Quantos desvios padrão da média? O escore-*z*

Outra forma de avaliar a posição é pelo número de desvios padrão que um ponto está da média. Por exemplo, as taxas de assassinatos dos Estados Unidos exibidas no diagrama de caixa e bigodes na Figura 3.18 têm uma média de 7,3 e um desvio padrão de 4,0. O valor de 20,3 da Louisiana está $20,3 - 7,3 = 13,0$ acima da média. Agora, 13 é $13/4 = 3,25$ desvios padrão. Assim a taxa de assassinatos da Louisiana está 3,25 desvios padrão acima da média.

O número de desvios padrão que uma observação está da média é chamado de **escore-*z***. Para as taxas de assassinatos da Figura 3.18, a Louisiana tem um escore-*z* de

$$z = \frac{20,3 - 7,3}{4,0} = \frac{\text{Observação} - \text{Média}}{\text{Desvio Padrão}} = 3,25.$$

Pela Regra Empírica, para uma distribuição com forma de sino é muito incomum que uma observação esteja a mais do que três desvios padrão da média. Um critério alternativo considera uma observação um valor atípico se ela tiver um escore-*z* maior do que 3 em valor absoluto. Por esse critério, a taxa de assassinatos da Louisiana é um valor atípico.

Estudaremos os escores-*z* em mais detalhes no próximo capítulo. Veremos que eles são especialmente úteis para as distribuições com forma de sino.

3.5 ESTATÍSTICA DESCRITIVA BIVARIADA

Neste capítulo nós aprendemos como resumir graficamente e numericamente variáveis categóricas e quantitativas. Nos próximos três capítulos iremos aprender ideias básicas de inferência estatística para uma variável categórica ou quantitativa. A maioria dos estudos tem mais do que uma variável, entretanto, os Capítulos 7 a 16 apresentam métodos que podem tratar de duas ou mais variáveis ao mesmo tempo.

Associação entre a variável resposta e a explicativa

Em uma análise multivariada, o foco principal é no estudo das *associações* entre as variáveis. É dito haver uma **associação** entre duas variáveis se certos valores de uma variável tendem a ir com certos valores da outra variável.

Por exemplo, considere a "afiliação religiosa" com as categorias (protestante, católica, outra) e "grupo étnico" com as categorias (anglo-americano, afro-americano, hispânico). Nos Estados Unidos, os anglo-americanos são mais prováveis de serem protestantes do que os hispânicos, que são, na sua maioria, católicos. Os afro-americanos são ainda mais prováveis de serem protestantes. Existe uma associação entre a afiliação religiosa e grupo étnico porque a proporção de pessoas que tem uma afiliação religiosa em particular muda à medida que o grupo étnico muda.

Uma análise da associação entre duas variáveis é chamada de análise **bivariada** porque existem duas variáveis. Geralmente uma é a variável de saída na qual as comparações são feitas nos níveis da outra variável. A variável de saída é chamada de **variável resposta**. A variável que define os grupos é chamada de **variável explicativa**. A análise estuda como um resultado da variável resposta *depende* ou *é explicado* pelo valor da variável explicativa. Por exemplo, quando descrevemos como a afiliação religiosa depende do grupo étnico, a afiliação religiosa é a variável resposta. Na comparação da renda de homens e mulheres, a renda é a variável resposta e o gênero é a variável explicativa. A renda pode depender do gênero, não o gênero da renda.

Normalmente, a variável resposta é chamada de **variável dependente** e a variável explicativa é chamada de **variável independente**. A terminologia *variável dependente* se refere ao objetivo de investigar o grau no qual a resposta naquela variável

depende do valor da outra variável. Preferimos não usar estes termos, visto que *dependente* e *independente* são usados para muitas outras coisas nos métodos estatísticos.

Comparar dois grupos é uma análise bivariada

O Capítulo 7 irá apresentar os métodos descritivos e inferenciais para a comparação de dois grupos. Por exemplo, suponha que gostaríamos de saber se são os homens ou as mulheres que têm mais melhores amigos, em média. Uma PSG relata (para a variável "NUMFREND") que o número médio de bons amigos é 7,0 para os homens (s = 8,4) e 5,9 para as mulheres (s = 6,0). As duas distribuições têm uma aparência similar, ambas são assimétricas à direita e com mediana de 4.

Aqui está uma análise das duas variáveis – número de melhores amigos e gênero. A variável resposta, número de melhores amigos, é quantitativa. A variável explicativa, gênero, é categórica. Nesse caso, é comum comparar as médias na variável resposta considerando as categorias da variável qualitativa. Os gráficos são também úteis, como os diagramas de caixa e bigodes lado a lado.

Dados categóricos bivariados

O Capítulo 8 irá apresentar métodos para analisar a associação entre duas variáveis categóricas. A Tabela 3.8 é um exemplo desses dados. Essa tabela é o resultado das respostas a duas perguntas da Pesquisa Social Geral de 2006. Uma pergunta era se as relações homossexuais eram erradas. A outra perguntava sobre o fundamentalismo/liberalismo da religião do respondente. Uma tabela deste tipo, chamada de **tabela de contingência**, exibe o número de sujeitos observados em combinações de possíveis resultados para as duas variáveis. Ela exibe como os resultados de uma variável resposta são *contingentes* na categoria da variável explicativa.

A Tabela 3.8 tem oito combinações possíveis de respostas. (O outro resultado possível, "moderado" para a variável religião, não é exibido aqui.) Poderíamos listar as categorias em uma distribuição de frequência ou construir um diagrama de colunas. Geralmente, é mais informativo fazer isto para as categorias da variável resposta, separadamente para cada categoria da variável explicativa. Por exemplo, se tratarmos a opinião sobre relações homossexuais como a variável resposta, poderíamos relatar os percentuais nas quatro categorias para relações homossexuais, separadamente para cada categoria da religião.

Considere aqueles que se declararam fundamentalistas. Visto que 416/547 = 0,76, 76% deles acreditam que as relações homossexuais estão sempre erradas. Da mesma forma, você pode verificar que 5% acreditam que elas estão quase sempre erradas, 4% acreditam que elas estão algumas vezes erradas e 15% acreditam que elas não estão erradas. Para aqueles que se declararam liberais, visto que 213/586 = 0,36, 36% acreditam que as relações homossexuais estão sempre erradas. Da mesma forma, você pode verificar que 5%

☑ **Tabela 3.8** Classificação cruzada da religião e opinião sobre relações homossexuais

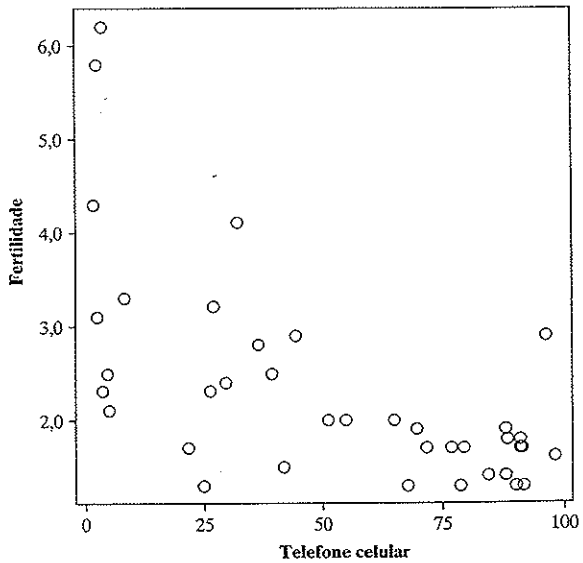
	Opinião sobre relações homossexuais				
Religião	Sempre erradas	Quase sempre erradas	Algumas vezes erradas	Não estão erradas	Total
Fundamentalista	416	26	22	83	547
Liberal	213	29	52	292	586

acreditam que elas estão quase sempre erradas, 9% acreditam que elas estão algumas vezes erradas e 50% acreditam que elas não estão erradas. Parece haver uma associação definitiva entre a opinião sobre a homossexualidade e a crença religiosa, com os fundamentalistas religiosos sendo mais negativos sobre a homossexualidade. O Capítulo 8 irá mostrar outras formas de analisar este tipo de dados.

Dados quantitativos bivariais

Quando ambas as variáveis são quantitativas, um diagrama que ainda não discutimos é útil. A Figura 3.19 mostra um exemplo usando o SPSS para fazer um gráfico dos dados de 38 países entre a fertilidade (o número médio de filhos por mulher adulta) e o percentual da população adulta usando telefones celulares. (Os dados são exibidos mais tarde no livro na Tabela 9.13.) Aqui, os valores do uso de telefones celulares estão representados graficamente no eixo horizontal, chamado de **eixo x**, e os valores da fertilidade estão representados graficamente no eixo vertical, chamado de **eixo y**. Os valores das duas variáveis para qualquer observação em particular formam um ponto relativo a estes eixos. Para descrever graficamente os dados amostrais, representamos graficamente as 38 observações como 38 pontos. Por exemplo, o ponto no lado superior esquerdo do gráfico representa o Paquistão, que tem uma fertilidade de 6,2 filhos por mulher, mas o uso do telefone celular é somente de 3,5%. Esta representação gráfica é chamada de **diagrama de dispersão**.

O diagrama de dispersão mostra a tendência de que os países com taxas mais altas de uso de telefones celulares apresentam níveis mais baixos de fertilidade. No Capítulo 9 iremos aprender sobre as duas formas de descrever isto como uma tendência. Uma forma, chamada de **correlação**, descreve quão forte é a associação, em termos de



☑ **Figura 3.19** Diagrama de dispersão entre as taxas de fertilidade e de uso de telefones celulares para 38 países. Os dados estão na Tabela 9.13 no Capítulo 9.

quão próximo os dados seguem *uma tendência linear*. Para a Figura 3.19, a correlação é -0,63. O valor negativo significa que a fertilidade tende a *diminuir* à medida que o uso do telefone celular *aumenta*. Ao contrário, o uso do telefone celular e o PIB (produto interno bruto, *per capita*) têm uma correlação positiva de 0,83. À medida que um aumenta, o outro também tende a aumentar.

A correlação assume valores entre -1 e +1. Quanto maior ela é, em valor absoluto, isto é, quanto mais longe de 0, mais forte a associação. O uso do telefone celular tem uma associação um pouco mais forte com o PIB do que com a fertilidade porque a correlação de 0,87 é maior em valor absoluto do que a correlação de -0,63.

A segunda ferramenta útil para descrever a tendência é a **análise de regressão**. Ela fornece uma fórmula direta para prever o valor da variável resposta de um valor dado da variável explicativa. Para a Figura 3.19, esta equação é

$$\text{Fertilidade prevista} = 3{,}4 - 0{,}02\,(\text{uso do telefone celular}).$$

Para um país sem o uso do telefone celular, a fertilidade prevista é de $3{,}4 - 0{,}02(0) = 3{,}4$ filhos por mãe. Para um país com 100% de adultos usando telefones celulares, a fertilidade prevista é de somente $3{,}4 - 0{,}02(100) = 1{,}4$ filhos por mãe.

O Capítulo 9 mostra como encontrar a correlação e a linha de regressão. Capítulos subsequentes mostram como estender a análise para lidar com variáveis categóricas assim como com variáveis quantitativas.

Analisando mais do que duas variáveis

Esta seção deu uma rápida olhada na análise de associações entre duas variáveis. Uma lição importante para mais tarde no livro é: *só porque duas variáveis têm uma associação não significa que exista uma conexão casual*. Por exemplo, ter mais pessoas em uma nação usando o telefone celular não significa que esta é a razão de que a taxa de fertilidade seja mais baixa (por exemplo, porque as pessoas estão falando nos telefones celulares em vez de fazer aquilo que gera bebês). Talvez valores altos no uso do telefone celular e valores baixos na fertilidade sejam ambos um subproduto de um país mais avançado economicamente.

A maioria dos estudos tem *várias* variáveis. A segunda metade deste livro (Capítulos 10 a 16) mostra como conduzir uma análise *multivariada*. Por exemplo, para estudar o que afeta o número de melhores amigos, poderemos simultaneamente considerar o gênero, idade, estado civil, nível de educação, se participa de eventos religiosos regularmente e se vive em um centro urbano ou rural.

3.6 ESTATÍSTICAS AMOSTRAIS E PARÂMETROS POPULACIONAIS

Das mensurações introduzidas neste capítulo, a média $\bar{y}$ é a medida de centro mais comumente relatada e o desvio padrão s é a medida mais comum de dispersão. Iremos usá-las frequentemente no restante do livro.

Visto que os valores $\bar{y}$ e s dependem da amostra selecionada, eles variam de amostra para amostra. Neste sentido, eles são variáveis. Os seus valores são desconhecidos antes de a amostra ser escolhida. Uma vez que a amostra é selecionada e eles são calculados, eles se tornam estatísticas amostrais conhecidas.

Na estatística inferencial, devemos fazer distinção entre um valor amostral (estatística) e as medidas correspondentes na população. A Seção 1.2 introduziu o termo *parâmetro* para uma medida resumindo a população. Uma estatística descreve uma amostra, enquanto um parâmetro descreve a população da qual a amostra foi coletada. Neste livro, letras gregas minúsculas geralmente representam os parâmetros e as letras romanas representam as estatísticas.

> ☑ **Notação para os parâmetros**
>
> A letra grega μ (mi) e a σ (sigma) representam a média e o desvio padrão de uma população.

Representamos por μ a **média da população** e por σ o **desvio padrão da população**. A média da população é um representante de todas as observações da população. O desvio padrão descreve a variabilidade destas observações em torno da média da população.

Enquanto as estatísticas $\bar{y}$ e s são variáveis, com valores dependendo da amostra escolhida, os parâmetros μ e σ são constantes. Isto ocorre porque μ e σ se referem apenas a um grupo em particular de observações, a saber, as observações de toda população. Os valores dos parâmetros são geralmente desconhecidos, e esta é uma razão para coletar amostras e calcular estatísticas amostrais para estimar os seus valores. Boa parte do restante deste livro trata das formas de se fazer inferências sobre parâmetros desconhecidos (como μ) usando uma estatística amostral (como $\bar{y}$). Antes de estudar os métodos de inferência, você precisa aprender algumas ideias básicas de *probabilidade*, que servem como fundamento a eles. A probabilidade é o assunto do próximo capítulo.

3.7 RESUMO DO CAPÍTULO

Este capítulo introduziu a **estatística descritiva** – formas de *descrever* os dados para resumir suas características básicas.

3.7.1 Visão geral das tabelas e gráficos

- Uma **distribuição de frequências** resume as contagens dos valores possíveis ou intervalos de valores. Uma distribuição de **frequências relativas** relata essa informação usando percentuais ou proporções.
- Um **diagrama de colunas** usa colunas sobre os valores possíveis para representar uma distribuição de frequências para uma variável categórica. Para uma variável quantitativa, um gráfico semelhante é chamado de **histograma***. Ele mostra se a distribuição tem aproximadamente uma forma de sino, forma de U, se é assimétrica à direita (a cauda mais longa apontando para a direita) ou tem qualquer outra forma.
- O **diagrama de caule e folhas** é uma exibição alternativa dos dados para uma variável quantitativa. Agrupa as observações que têm o mesmo dígito dominante (caule) e mostra também o seu dígito final (folha). Para amostras pequenas ele exibe as observações individuais.
- O **diagrama de caixa e bigodes** descreve os quartis, os valores extremos e os valores atípicos. O diagrama de caixa e bigodes e o diagrama de caule e folhas também podem fornecer comparações um contra o outro entre dois grupos.

3.7.2 Visão geral das medidas de centro

As **medidas do centro** descrevem o centro dos dados, em termos de uma observação típica.

- A **média** é a soma de todas as observações divida pelo tamanho da amostra. Ela é o centro de gravidade dos dados.
- A **mediana** divide o conjunto ordenado de dados em duas partes com o mesmo número de observações, metade delas abaixo da mediana e a outra metade acima.
- A quarta parte inferior das observações está abaixo do **quartil inferior** e

* N. de T. T.: De fato, em um diagrama de colunas elas estão separadas enquanto que em um histograma elas são justapostas.

a quarta parte superior está acima do **quartil superior**. Estes são os 25º e o 75º **percentis**. A mediana é o 50º percentil. Os quartis e a mediana dividem os dados em quatro partes iguais. Eles são menos afetados do que a média por valores atípicos ou assimetria extrema.

- A **moda** é o valor que ocorre mais frequentemente. Ela é válida para qualquer tipo de dados, embora seja usada com mais frequência com dados categóricos ou variáveis discretas aceitando relativamente poucos valores.

3.7.3 Visão geral das medidas de variabilidade

As **medidas de variabilidade** descrevem a dispersão dos dados.

- O **intervalo** é a diferença entre a maior e a menor observação. O **intervalo interquartil** é o intervalo que compreende a metade dos dados entre os quartis superior e inferior. Ele é menos afetado pelos valores atípicos.
- A **variância** calcula a média dos quadrados dos desvios em torno da média. Sua raiz quadrada, o **desvio padrão**, é fácil de interpretar, descrevendo a distância típica da média.
- A **Regra Empírica** declara que, em uma distribuição com a forma de sino, aproximadamente 68% das observações estão dentro de um desvio padrão da média, aproximadamente 95% estão dentro de dois desvios padrão, e aproximadamente todas, senão todas, estão dentro de três desvios padrão.

A Tabela 3.9 resume as medidas do centro e de variabilidade. Uma **estatística** resume uma amostra. Um **parâmetro** resume uma população. A **inferência estatística** usa estatísticas para fazer previsões sobre os parâmetros.

3.7.4 Visão geral das estatísticas descritivas bivariadas

As **estatísticas bivariadas** são usadas para analisar dados de duas variáveis em conjunto.

☑ **Tabela 3.9** Resumo das medidas de centro e de variabilidade

Medida	Definição	Interpretação
Centro		
Média	$\bar{y} = \Sigma y_i/n$	Centro da gravidade
Mediana	Observação do meio da amostra ordenada	50º percentil, divide a amostra em duas partes iguais.
Moda	Valor mais frequente	Resultado mais provável, válido para todos os tipos de dados
Variabilidade		
Desvio padrão	$s = \sqrt{\Sigma(y_i - \bar{y})^2/(n-1)}$	Regra Empírica: se tiver forma de sino, 68%, 95% dentro de s, $2s$ de $\bar{y}$
Intervalo	A diferença entre a maior e a menor observação	Maior com mais variabilidade
Intervalo interquartil	A diferença entre o quartil superior (75º percentil) e o quartil inferior (25º percentil)	Abrange a metade dos dados

- Muitos estudos analisam como o resultado de uma **variável resposta** depende do valor de uma variável explicativa.
- Para variáveis categóricas, uma **tabela de contingência** mostra o número de observações em todas as combinações possíveis de resultados para as duas variáveis.
- Para variáveis quantitativas, um **diagrama de dispersão** faz uma representação gráfica das observações, exibindo um ponto para cada observação. A variável resposta é representada graficamente no eixo y e a variável explicativa no eixo x.
- Para variáveis quantitativas, a **correlação** descreve a força da associação linear. Ela varia entre -1 e +1 e indica se a variável resposta tende a aumentar (correlação positiva) ou diminui (correlação negativa) à medida que a variável explicativa aumenta ou diminui.
- Uma **análise de regressão** fornece uma equação linear para prever o valor da variável resposta usando a variável explicativa. Estudaremos a correlação e a regressão em detalhes no Capítulo 9.

EXERCÍCIOS

Praticando o básico

3.1 A Tabela 3.10 mostra o número (em milhões) de habitantes, em 2004, dos Estados Unidos nascidos no exterior classificados por local de nascimento.

(a) Construa uma distribuição de frequências relativas.

(b) Represente os dados em um diagrama de colunas.

(c) O "local de nascimento" é uma variável quantitativa ou categórica?

(d) Use uma das seguintes medidas que seja adequada para estes dados: média, mediana, moda.

☑ **Tabela 3.10**

Local de nascimento	Habitantes
Europa	4,7
Ásia	8,7
Caribe	3,3
América Central	12,9
América do Sul	2,1
Outro	2,6
Total	34,3

Fonte: Statistical Abstract of the United States, 2006.

3.2 De acordo com www.adherents.com, em 2006, o número dos seguidores das cinco maiores religiões do mundo era de 2,1 bilhões para o cristianismo, 1,3 bilhões para o islamismo, 0,9 bilhões para o hinduísmo, 0,4 bilhões para o confucionismo e 0,4 bilhões para o budismo.

(a) Construa uma distribuição de frequências relativas.

(b) Construa um diagrama de colunas.

(c) Você pode determinar a média, a mediana ou a moda para estes dados? Se sim, determine e interprete.

3.3 Um professor mostra para a sua turma os escores da prova do meio do semestre em um diagrama de caule e folhas:

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

(a) Identifique o número de estudantes e os escores mínimos e máximos.

(b) Construa um histograma com quatro intervalos.

3.4 De acordo com o *2005 American Community Survey*, naquele ano, os Estados Unidos tinham 30,1 milhões de domicílios com uma pessoa, 37,0 milhões com duas pessoas, 17,8 milhões com três pessoas, 15,3 milhões com quatro pessoas e 10,9 milhões com cinco ou mais pessoas.

(a) Construa uma distribuição de frequências relativas.

(b) Construa um histograma. Qual é a forma?

(c) Relate e interprete (i) a mediana, (ii) a moda do tamanho do domicílio.

3.5 Copie o arquivo de dados *2005 statewide crime* do *site* do livro em www.grupoa.com.br. Use a variável taxa de assassinatos (por 100000 habitantes). Neste exercício, não use as observações para D.C. Use um *software*.

(a) Construa uma distribuição de frequências relativas.

(b) Construa um histograma. Como você descreveria a forma do histograma?

(c) Construa um diagrama de caule e folhas. Como podemos comparar este diagrama ao histograma da parte (b)?

3.6 A OECD (Organization for Economic Cooperation and Development – Organização para a Cooperação e Desenvolvimento Econômico) consiste em países desenvolvidos e industrializados que aceitam os princípios de democracia representativa e uma economia de mercado livre. A Tabela 3.11 mostra os dados das Nações Unidas para os países da OECD em várias variáveis: produto interno bruto (PIB, *per capita* em dólares americanos), percentual de desempregados, uma medida de desigualdade baseada na comparação da riqueza dos 10% mais ricos com os 10% mais pobres, gasto público em saúde (como um percentual do PIB), o número de médicos por 100000 pessoas, emissão de dióxido de carbono (CO_2, *per capita* em toneladas métricas), o percentual de mulheres no parlamento e atividade econômica feminina como um percentual da taxa masculina. Estes dados são do arquivo *OECD data* do *site* do livro.

☑ **Tabela 3.11** Dados para os países do OECD, disponíveis no arquivo *OECD data* no *site* do livro

País	PIB	Desempr.	Desig.	Saúde	Médicos	CO_2	Mulheres no parlamento	Econ. feminina
Austrália	30,331	5,1	12,5	6,4	247	18	28,3	79
Áustria	32,276	5,8	6,9	5,1	338	8,6	32,2	75
Bélgica	31,096	8,4	8,2	6,3	449	8,3	35,7	72
Canadá	31,263	6,8	9,4	6,9	214	17,9	24,3	83
Dinamarca	31,914	4,9	8,1	7,5	293	10,1	36,9	84
Finlândia	29,951	8,6	5,6	5,7	316	13	37,5	86
França	29,300	10,0	9,1	7,7	337	6,2	13,9	79
Alemanha	28,303	9,3	6,9	8,7	337	9,8	30,5	76
Grécia	22,205	10,6	10,2	5,1	438	8,7	13	66
Islândia	33,051	2,5	..	8,8	362	7,6	33,3	87
Irlanda	38,827	4,3	9,4	5,8	279	10,3	14,2	72
Itália	28,180	7,7	11,6	6,3	420	7,7	16,1	61
Japão	29,251	4,4	4,5	6,4	198	9,7	10,7	65
Luxemburgo	69,961	4,6	..	6,2	266	22	23,3	68
Holanda	31,789	6,2	9,2	6,1	315	8,7	34,2	76
Nova Zelândia	23,413	3,6	12,5	6,3	237	8,8	32,2	81
Noruega	38,454	4,6	6,1	8,6	313	9,9	37,9	87
Portugal	19,629	7,5	15	6,7	342	5,6	21,3	79
Espanha	25,047	9,1	10,3	5,5	330	7,3	30,5	65
Suécia	29,541	5,6	6,2	8	328	5,9	45,3	87
Suíça	33,040	4,1	9	6,7	361	5,6	24,8	79
Reino Unido	30,821	4,8	13,8	6,9	230	9,4	18,5	79
Estados Unidos	39,676	5,1	15,9	6,8	256	19,8	15	81

Fonte: hdr.undp.org/statistics/data
Desempr. = % desempregados, Desig. = medida de desigualdade, Mulheres no parlamento = % de mulheres no parlamento, Econ. feminina = atividade econômica feminina (% da taxa masculina).

(a) Construa em diagrama de caule e folhas dos valores do PIB, arredondando e apresentando os valores em milhares de dólares (por exemplo, substituindo $19629 por 20).

(b) Construa um histograma correspondente ao diagrama de caule e folhas em (a).

(c) Identifique o valor atípico em cada diagrama.

3.7 Recentemente, o número de abortos em todos os estados, por 1000 mulheres entre 15 a 41 anos, para os estados da região do Pacífico nos Estados Unidos eram: Washington, 26; Oregon, 17; Califórnia, 236; Alasca, 2; e Havaí, 6 (*Statistical Abstract of the United States, 2006*).

(a) Calcule a média.

(b) Determine a mediana. Por que ela é tão diferente da média?

3.8 O aquecimento global parece ser principalmente o resultado da atividade humana que produz emissões de dióxido de carbono e outros gases do efeito estufa. O *Human Development Report 2005* (Relatório do Desenvolvimento Humano), publicado pelo Programa do Desenvolvimento Humano das Nações Unidas, determinou as emissões *per capita* em 2002 para os oito maiores países considerando o número de habitantes, em toneladas métricas (1000 kg) por pessoa: Bangladesh 0,3; Brasil 1,8; China 2,3; Índia 1,2; Indonésia 1,4; Paquistão 0,7; Rússia 9,9; Estados Unidos 20,1.

(a) Para estes oito valores, encontre a média e a mediana.

(b) Alguma observação parece ser um valor atípico? Discuta o seu impacto na comparação entre a média e a mediana.

3.9 Um levantamento de dados da Organização Roper perguntou: "Até onde foram as leis e regulamentos de proteção do meio ambiente?". Para as respostas possíveis *não muito longe, o suficiente*, e *muito longe*, o percentual das respostas foi 51%, 33% e 16%.

(a) Qual resposta é a moda?

(b) É possível calcular a média ou a mediana para estes dados? Se sim, calcule; se não, explique.

3.10 Um pesquisador de um centro de tratamento do alcoolismo, para estudar o tempo de permanência no centro de pacientes internados pela primeira vez, selecionou aleatoriamente dez relatórios de indivíduos internados nos últimos dois anos. O tempo de permanência, em dias, foi: 11, 6, 20, 9, 13, 4, 39, 13, 44 e 7.

(a) Construa um diagrama de caule e folhas.

(b) Encontre a média e o desvio padrão e interprete.

(c) Para um estudo similar 25 anos atrás, o tempo de permanência para dez indivíduos amostrados foi de 32, 18, 55, 17, 24, 31, 20, 40, 24, 15. Compare os resultados com aqueles do estudo recente usando (i) um diagrama de caule e folhas, (ii) a média, (iii) o desvio padrão. Interprete todas as diferenças que encontrar.

(d) Na verdade, o estudo recente também selecionou outro relatório. Aquele paciente ainda está internado após 40 dias. Assim, o tempo de permanência daquele paciente é de pelo menos 40 dias, mas o valor real é desconhecido. Você pode calcular a média ou a mediana para a amostra completa de 11 incluindo esta observação parcial? Explique. (Esta observação é dita ser *censurada*, significando que o valor observado é truncado do seu valor verdadeiro desconhecido.)

3.11 Acesse o PSG (Pesquisa Social Geral) em http://sda.berkeley.edu/GSS. Entre com a variável "TVHOURS" e com o ano de "2006" no filtro de seleção para obter os dados em horas por dia de assistir à televisão nos Estados Unidos em 2006.

(a) Construa a distribuição de frequências relativas para os valores 0, 1, 2, 3,4, 5, 6,7 ou mais.

(b) Como você descreveria a forma da distribuição?

(c) Explique porque a mediana é dois.

(d) A média é maior do que 2. Por que você acha que ela é assim?

☑ **Tabela 3.12**

País	Atividade econ. feminina	País	Atividade econ. feminina	País	Atividade econ. feminina
Áustria	66	Alemanha	71	Noruega	86
Bélgica	67	Grécia	60	Portugal	72
Chipre	63	Irlanda	54	Espanha	58
Dinamarca	85	Itália	60	Suécia	90
Finlândia	87	Luxemburgo	58	Reino Unido	76
França	78	Holanda	68		

Fonte: Human Development Report, 2005, United Nations Development Programme.

3.12 A Tabela 3.12 mostra a atividade econômica feminina em 2003 (número de mulheres na força de trabalho por 100 homens na força de trabalho), para países da Europa Ocidental. Construa um diagrama de caule e folhas "um contra o outro" destes valores em contraposição com os valores da América do Sul da Tabela 3.4. Qual é a sua interpretação?

3.13 De acordo com a Statistics Canada, em 2000 a renda domiciliar no Canadá tinha uma mediana de $46752 e uma média de $71600. O que você preveria sobre a forma da distribuição? Por quê?

3.14 A Tabela 3.13 resume as respostas de 2333 sujeitos da Pesquisa Social Geral de 2006 à pergunta: "Aproximadamente quantas vezes você fez sexo nos últimos 12 meses?".

(a) Determine a mediana e a moda. Interprete.

(b) Trate esta escala de uma forma quantitativa atribuindo os escores 0; 0,1; 1,0; 2,5; 4,3; 10,8 e 17 às categorias para a frequência mensal aproximada. Encontre a média amostral e interprete.

☑ **Tabela 3.13**

Número de vezes que fez sexo	Frequência
Nenhuma vez	595
Uma ou duas vezes	205
Aproximadamente uma vez por mês	265
2 a 3 vezes por mês	361
Aproximadamente uma vez por semana	343
2 a 3 vezes por semana	430
Mais do que 3 vezes por semana	134

3.15 A PSG de 2004 perguntou aos respondentes: "Com que frequência você lê o jornal?". As respostas possíveis eram (todos os dias, algumas vezes por semana, uma vez por semana, menos do que uma vez por semana, nunca) e as contagens nestas categorias foram (358, 222, 134, 121, 71).

(a) Identifique a resposta moda e a mediana.

(b) Seja y = número de vezes que você lê o jornal numa semana, medida como descrito acima. Para os escores (7; 3; 1; 0,5; 0) para as categorias, encontre $\bar{y}$. Como ela se compara à média de 4,4 da PSG de 1994?

3.16 De acordo com a Agência Governamental do Censo Americano, *2005 American Community Survey*, a mediana dos rendimentos nos últimos 12 meses foi de $32168 para mulheres e $41965 para homens, enquanto a média foi de $39890 para mulheres e $56724 para homens.

(a) Isto sugere que a distribuição da renda para cada gênero é simétrica ou assimétrica à direita ou à esquerda? Explique.

(b) Os resultados se referem aos 73,8 milhões de mulheres e 83,4 milhões de homens. Encontre a média geral do rendimento.

3.17 Em 2003, nos Estados Unidos, a renda média familiar era de $55800 para famílias brancas, $34400 para famílias negras e $34300 para famílias hispânicas (*Statistical Abstract of the United States, 2006* – Resumo Estatístico dos Estados Unidos, 2006).

(a) Identifique a variável resposta e a explicativa para esta análise.

(b) Existe informação suficiente para encontrar a mediana quando todos os dados são combinados dos três grupos? Por que ou por que não?

(c) Se os valores relatados fossem médias, de que mais você precisaria para encontrar a média geral?

3.18 A PSG perguntou: "Durante os últimos 12 meses quantas pessoas você conheceu pessoalmente e que foram vítimas de homicídio?". A Tabela 3.14 mostra uma listagem das respostas analisadas.

(a) A distribuição tem uma forma de sino, é assimétrica à direita ou assimétrica à esquerda?

(b) A Regra Empírica se aplica a esta distribuição? Por que ou por que não?

(c) Determine a mediana. Se as 500 observações variam de 0 a 6, como a mediana muda? Que propriedade isto ilustra para a mediana?

3.19 Em outubro de 2006, um artigo na wikipedia.org sobre o "salário mínimo" divulgou o salário mínimo por hora (em dólares americanos) de cinco países: $10,00 na Austrália, $10,25 na Nova Zelândia, $10,46 na França, $10,01 no Reino Unido e $5,15 nos Estados Unidos. Encontre a mediana, a média, o intervalo (amplitude) e o desvio padrão **(a)** excluindo os Estados Unidos, **(b)** utilizando os cinco países. Use os dados para explicar o efeito dos valores atípicos nestas medidas.

3.20 A revista *National Geographic Traveler* recentemente apresentou dados sobre o número anual médio de dias de férias de residentes de oito países diferentes. Eles relataram 42 dias para a Itália, 37 dias para a França, 35 para a Alemanha, 34 para o Brasil, 28 para a Grã-Bretanha, 26 para o Canadá, 25 para o Japão e 13 para os Estados Unidos.

(a) Encontre a média e o desvio padrão. Interprete.

(b) Determine o resumo de cinco números. (*Dica:* Você pode encontrar o quartil inferior encontrando a mediana dos quatro valores abaixo dela.)

3.21 O Índice de Desenvolvimento Humano (IDH) é um índice que as Nações Unidas usam para atribuir uma classificação a cada país, baseado na expectativa de vida ao nascer, grau de instrução e renda. Em 2006, os dez países (em ordem) com a taxa de IDH mais alta, seguidos em parênteses pelo percentual de cadeiras no parlamento ocupadas por mulheres (que é uma medida de poder do gênero), foram Noruega (38), Islândia (33), Austrália (28), Irlanda (14), Suécia (45), Canadá (24), Japão (11), Estados Unidos (15), Suíça (25), Holanda (34). Encontre a média e o desvio padrão do número de mulheres no parlamento e interprete.

3.22 O *Human Development Report 2006* (Relatório do Desenvolvimento Humano) publicado pelas Nações Unidas mostrou a expectativa de vida por país. Para a Europa Ocidental, os valores foram:

Dinamarca 77, Portugal 77, Holanda 78, Finlândia 78, Grécia 78, Irlanda 78, Reino Unido 78, Bélgica 79, França 79, Alemanha 79,

☑ **Tabela 3.14**

VÍTIMAS	Frequência	Percentual
0	1244	90,8
1	81	5,9
2	27	2,0
3	11	0,8
4	4	0,3
5	2	0,1
6	1	0,1

N	Média	Desvio padrão	Máximo	Q3	Mediana	Q1	Mínimo
1370	0,146	0,546	6	0	0	0	0

Noruega 79, Itália 80, Espanha 80, Suécia 80, Suíça 80.

Para a África, os valores relatados (muitos dos quais eram substancialmente mais baixos do que cinco anos antes por causa da prevalência da Aids) foram:

Botsuana 37, Zâmbia 37, Zimbábue 37, Maláui 40, Angola 41, Nigéria 43, Ruanda 44, Uganda 47, Quênia 47, Mali 48, África do Sul 49, Congo 52, Madagascar 55, Senegal 56, Sudão 56, Gana 57.

(a) Que grupo de expectativa de vida você acha que tem o maior desvio padrão? Por quê?

(b) Encontre o desvio padrão para cada grupo. Compare-os para ilustrar que s é maior para o grupo que se mostra mais disperso.

3.23 Um relatório indica que o salário anual dos professores em Ontário tem uma média de \$50000 e um desvio padrão de \$10000 (dólares canadenses). Suponha que a distribuição tenha aproximadamente a forma de sino.

(a) Dê um intervalo de valores que contenha aproximadamente (i) 68%, (ii) 95%, (iii) todos ou aproximadamente todos os salários.

(b) Um salário de \$100000 seria incomum? Por quê?

3.24 Excluindo os Estados Unidos, o número médio de feriados e dias de férias em um ano para os países da OECD (veja Exercício 3.6) tem aproximadamente a forma de sino com a média de 35 dias e desvio padrão de 3 dias.[1]

(a) Use a Regra Empírica para descrever a variabilidade.

(b) A observação para os Estados Unidos é 19. Se ela for incluída com as outras observações, (i) a média irá aumentar ou diminuir, (ii) o desvio padrão irá aumentar ou diminuir?

(c) Usando a média e o desvio padrão dos outros países, em quantos desvios padrão os Estados Unidos diferem da média?

3.25 Para os dados da PSG em "número de pessoas que você conhece que cometeram suicídio", 88,8% das respostas foram 1, e as outras respostas tiveram valores mais altos. A média foi igual a 0,145 e o desvio padrão foi igual a 0,457.

(a) Qual o percentual das observações que está dentro de um desvio padrão da média?

(b) A Regra Empírica é apropriada para a distribuição desta variável? Por que ou por que não?

3.26 O primeiro exame do seu curso de Estatística é classificado em uma escala de 0 a 100 e a média é 76. Que valor é mais plausível para o desvio padrão: -20, 0, 10 ou 50? Por quê?

3.27 A média geral das notas dos formandos na Universidade de Rochester deve estar entre 2,0 e 4,0. Considere os possíveis valores do desvio padrão: -10,0; 0,0; 0,4; 1,5; 6,0.

(a) Que valor é o mais realista? Por quê?

(b) Que valor é *impossível*? Por quê?

3.28 De acordo com a Agência do Censo Americano, o preço médio de venda das casas vendidas em todo o país, em 2005, foi de \$184100. Qual dos seguintes é o valor mais plausível para o desvio padrão: **(a)** -15000. **(b)** 1000, **(c)** 10000, **(d)** 60000, **(e)** 1000000? Por quê?

3.29 Para todas as casas em Gainesville, Flórida, o consumo de energia elétrica[2] para o ano de 2006 teve uma média de 10449 e um desvio padrão de 7489 quilowatts-hora (kWh). O uso máximo foi de 336240 kWh.

(a) Que forma você espera que essa distribuição tenha?

(b) Você espera que essa distribuição tenha valores atípicos? Explique.

3.30 O consumo de água em residências (em milhares de galões) em Gainesville, Flórida, em 2006 teve uma média de 78 e um desvio padrão de 119. Que forma você espera que esta distribuição tenha? Por quê?

3.31 De acordo com *Statistical Abstract of the United States 2006*, a média salarial (em dólares) de professores do ensino médio, em 2004, dos Estados Unidos variava entre os estados com um resumo dos cinco números de

```
100% Max  61,800 (Illinois)
 75% Q3   48,850
 50% Med  42,700
 25% Q1   39,250
  0% Mín  33,100 (Dakota do Sul)
```

(a) Encontre e interprete o intervalo.

(b) Encontre e interprete o intervalo interquartil.

3.32 Considere o exercício anterior.

(a) Faça um diagrama de caixa e bigodes.

(b) Baseado em (a), faça uma previsão da direção da assimetria para esta distribuição. Explique.

(c) Se a distribuição, embora assimétrica, tem aproximadamente a forma de sino, que valor é mais plausível para o desvio padrão: (i) 100, (ii) 1000, (iii) 7000, (iv) 25000? Explique.

3.33 A Tabela 3.15 exibe parte de uma saída computacional para analisar as taxas de assassinatos (por 100000) no arquivo de dados *2005 statewide crime* no *site* do livro. A primeira coluna se refere a todo o conjunto de dados, e a segunda coluna exclui a observação para D.C. Para cada estatística relatada, avalie o efeito de incluir a observação atípica para D.C.

3.34 Durante um semestre recente na Universidade da Flórida, o uso do computador[3] por estudantes que têm conta num computador principal foi resumido por uma média de 1921 e um desvio padrão de 11495 kilobytes de utilização do *drive*.

(a) A Regra Empírica se aplica a esta distribuição? Por quê?

(b) O resumo dos cinco números foi: mínimo = 4, $Q1$ = 256, mediana = 530, $Q3$ = 1105, e máximo = 320000. O que isto sugere sobre a forma da distribuição? Por quê?

(c) Use o critério 1,5(IIQ) para determinar se existem valores atípicos presentes.

☑ **Tabela 3.15**

Variável = ASSASSINATOS			
N	51	N	50
Média	5,6	Média	4,8
Desvio Padrão	6,05	Desvio Padrão	2,57
Quartis		Quartis	
100% Max	44	100% Max	13
75% Q3	6	75% Q3	6
50% Med	5	50% Med	5
25% Q1	3	25% Q1	3
0% Mín	1	0% Mín	1
Intervalo	43	Intervalo	12
Q3-Q1	3	Q3-Q1	3
Moda	3	Moda	3

3.35 Para cada uma das alternativas seguintes, esboce um histograma e explique qual medida será maior: a média ou a mediana.

(a) O preço de casas novas em 2008.

(b) O número de filhos nascidos de mulheres com 40 anos ou mais.

(c) O escore de um exame fácil (média = 88, desvio padrão = 10, máximo possível = 100).

(d) O número de carros pertencentes a uma família.

(e) O número de meses no qual o sujeito dirigiu um carro no ano passado.

3.36 Para cada uma das seguintes variáveis indique se você esperaria que o histograma de frequências relativas tivesse: a forma de sino, a forma de U, assimetria à direita ou assimetria à esquerda.

(a) Escores de um exame fácil, com média = 88, desvio padrão = 10, mínimo = 65, quartil inferior = 77, mediana = 85, quartil superior = 91, máximo = 100.

(b) QI de uma população.

(c) Número de vezes que foi preso no ano passado.

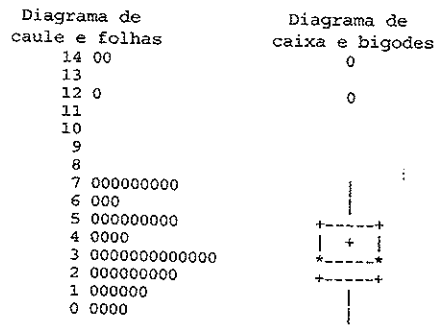
(d) O tempo necessário para completar um exame difícil (o tempo máximo é de 1 hora).
(e) Idade ao morrer.
(f) Contribuição semanal para a igreja (a mediana é de $10 e a média é de $17).
(g) A atitude em relação à legalização do aborto.

3.37 Para as partes (a), (b) e (f) do exercício anterior, esboce diagramas de caixa e bigodes que seriam plausíveis para a variável.

3.38 As taxas de desemprego, em janeiro de 2007, dos 27 países da União Europeia vão de 3,2 (Dinamarca) a 12,6 (Polônia), com quartil inferior = 5,0, mediana = 6,7, quartil superior = 7,9, média = 6,7 e desvio padrão = 2,2. Esboce um diagrama de caixa e bigodes identificando quais destes valores são usados no diagrama.

3.39 Para os dados do levantamento sobre o número de vezes por semana que estudantes leem jornal, mencionado no Exercício 1.11, a Figura 3.20 mostra uma saída computacional de um diagrama de caule e folhas e de um diagrama de caixa e bigodes.

(a) Do diagrama de caixa e bigodes, identifique o mínimo, o quartil inferior, a mediana, o quartil superior e o máximo.
(b) Identifique os mesmos cinco números usando o diagrama de caule e folhas.
(c) Os dados parecem conter valores atípicos? Se sim, identifique-os.
(d) O desvio padrão é um dos seguintes valores: - 0,3; 3; 13; 23. Qual você acha que é e por quê?



☑ **Figura 3.20** Saída computacional de diagrama de caule e folhas e diagrama de caixa e bigodes.

3.40 As taxas de mortalidade infantil (número de mortes infantis por 1000 nascidos com vida) são registradas pelas Nações Unidas. No seu relatório de 2006, os valores para a África apresentaram um resumo dos cinco números de:

mín = 54, Q1 = 76, mediana = 81,
Q3 = 101, máx = 154.

Os valores para a Europa Ocidental apresentaram um resumo dos cinco números de:

mín. = 3, Q1 = 4, mediana = 4,
Q3 = 4, máx. = 5.

Faça diagramas de caixa e bigodes lado a lado e use-os para descrever as diferenças entre as distribuições. (O diagrama para a Europa mostra que os quartis, como a mediana, são menos úteis quando os dados são altamente discretos.)

3.41 Em 2004, o resumo dos cinco números para o percentual considerando todos os estados para pessoas sem seguro de saúde tinha um mínimo de 8,9% (Minnesota), *Q1* = 11,6, Med = 14,2, *Q3* = 17,0 e máximo de 25,0% (Texas) (*Statiscal Abstract of the United States, 2006*).

(a) Faça um diagrama de caixa e bigodes.
(b) Você acha que a distribuição é simétrica, assimétrica à direita ou assimétrica à esquerda? Por quê?

3.42 As taxas de graduação do ensino médio nos Estados Unidos em 2004 tinham um mínimo de 78,3 (Texas), quartil inferior de 83,6, mediana de 87,2, quartil superior de 88,8 e máximo de 92,3 (Minnesota) (*Statiscal Abstract of the United States, 2006*).

(a) Determine e interprete o intervalo e o intervalo interquartil (IIQ).
(b) Existem valores atípicos de acordo com o critério 1,5(IIQ)?

3.43 Usando um *software*, analise as taxas de assassinatos do arquivo de dados *2005 statewide crime* do *site* do livro.

(a) Usando o conjunto de dados sem o D.C., encontre o resumo dos cinco números.
(b) Construa um diagrama de caixa e bigodes e interprete-o.

(c) Repita a análise, incluindo o valor do D.C. e compare os resultados.

3.44 Um relatório da OECD[4] indicou que o consumo anual de água para as nações na OECD (veja Exercício 3.6) era assimétrico à direita, com os valores (em metros cúbicos *per capita*) tendo uma mediana de aproximadamente 500 e um intervalo de aproximadamente 200 na Dinamarca a 1700 nos Estados Unidos. Considere os valores possíveis para o IIQ: -10, 0, 10, 350, 1500. Que valor é o mais realista? Por quê?

3.45 De acordo com os valores do *Human Development Report* (Relatório do Desenvolvimento Humano) publicado pelas Nações Unidas (hdr.undp.org), as emissões de dióxido de carbono em 2005, para os 25 países da União Europeia (EU) em 2005, tinham uma média de 8,3 e um desvio padrão de 3,3 em toneladas métricas *per capita*. Todos os valores estavam abaixo de 12, exceto Luxemburgo, que tinha um valor de 21,1.

(a) Quantos desvios padrão acima da média estava o valor de Luxemburgo?
(b) O valor da Suécia era de 5,8. Quantos desvios padrão abaixo da média ele estava?
(c) As emissões de dióxido de carbono eram de 16,5 para o Canadá e 20,1 para os Estados Unidos. Relativo à distribuição para a União Europeia, encontre e interprete o escore-*z* para (i) Canadá, (ii) Estados Unidos.

3.46 A publicação das Nações Unidas *Energy Statistics Yearbook* (unstats.un.org/unsd/energy) lista o consumo de energia. Para os 25 países que constituíam a União Europeia em 2006, os valores da energia (em quilogramas *per capita*) tinham uma média de 4998 e um desvio padrão de 1786.

(a) A Itália tinha um valor de 4222. Quantos desvios padrão da média estava este valor?
(b) O valor para os Estados Unidos era de 11067. Em relação à distribuição para a União Europeia, encontre o escore-*z*. Interprete-o.
(c) Se a distribuição dos valores de energia da União Europeia tinha a forma de sino, um valor de 11067 seria excepcionalmente alto? Por quê?

3.47 Um estudo compara os Democratas e os Republicanos em sua opinião sobre o seguro nacional de saúde (a favor ou contra).

(a) Identifique a variável resposta e a variável explicativa.
(b) Explique como os dados poderiam ser resumidos em uma tabela de contingência.

3.48 A Tabela 3.16 mostra a felicidade relatada para aqueles sujeitos da PSG de 2004 que disseram frequentar serviços religiosos raramente ou frequentemente (variáveis "ATTEND".e "HAPPY").

(a) Identifique a variável resposta e a variável explicativa.
(b) Em cada nível de comparecimento aos serviços religiosos encontre o percentual dos que declararam ser muito felizes.
(c) Parece haver uma associação entre estas variáveis? Por quê?

3.49 Dados recentes das Nações Unidas sobre vários países forneceram uma equação de previsão relacionando a fertilidade (o número médio de filhos por

☑ **Tabela 3.16**

	Felicidade			
Comparecimento a cultos religiosos	**Muito feliz**	**Moderadamente feliz**	**Não muito feliz**	**Total**
Quase toda semana ou mais	200	220	29	449
Nunca ou menos do que uma vez ao ano	72	185	53	310

mulher adulta) com o percentual de pessoas que usam a internet:

Fertilidade prevista =
3,2 – 0,04 (uso da internet)

(a) Compare a fertilidade prevista de um país com 50% de uso da internet (os Estados Unidos) com um país com 0% de uso (Iêmen).

(b) A correlação é -0,55. Explique o que o valor negativo representa.

3.50 Considere o exercício anterior. Uma equação de previsão relacionando a fertilidade com o percentual de pessoas que usam métodos contraceptivos é:

Fertilidade prevista =
6,6 – 0,065(uso de contraceptivo)

e a correlação é -0,89.

(a) Que tipo de padrão você esperaria ver em um diagrama de dispersão destes dados?

(b) Que variável parece ser mais fortemente associada à fertilidade – uso da internet ou o uso do contraceptivo? Por quê?

3.51 Considerando os dados dos países da OECD, apresentados na Tabela 3.11 do Exercício 3.6, utilize um *software* para construir um diagrama de dispersão relacionando y = emissões de dióxido de carbono e x = PIB.

(a) Baseado neste diagrama, você esperaria que a correlação entre estas variáveis fosse positiva ou negativa? Por quê?

(b) Você vê uma observação que está longe das demais? Identifique o país.

3.52 Considere o exercício anterior. A correlação com as emissões de dióxido de carbono é de 0,03 para as atividades econômicas femininas e -0,52 para o número de médicos. Qual das variáveis está mais fortemente associada às emissões de dióxido de carbono? Por quê?

3.53 Qual é a diferença entre as medidas descritivas simbolizadas por

(a) $\bar{y}$ e μ?

(b) s e σ?

Conceitos e aplicações

3.54 Utilize o arquivo de dados *Student survey* no *site* do livro (veja Exercício 1.11 na página 25) e, com a ajuda de um *software*, determine resumos numéricos e representações gráficas para:

(a) a distância da cidade natal.

(b) as horas semanais assistindo à televisão. Descreva as formas das distribuições e resuma suas descobertas.

3.55 Considere o arquivo de dados que sua turma criou para o Exercício 1.12 (página 26). Para as variáveis escolhidas pelo seu professor, realize uma análise estatística descritiva. No seu relatório, dê um exemplo de uma questão de pesquisa que poderia ser feita usando sua análise, identificando as variáveis resposta e explicativa. Resuma e interprete suas descobertas.

3.56 A Tabela 3.17 mostra as taxas anuais de morte por arma de fogo (incluindo homicídio, suicídio e mortes acidentais) por 100000 habitantes em países industrialmente desenvolvidos. Prepare um relatório no qual você resuma os dados usando métodos gráficos e numéricos apresentados neste capítulo.

3.57 Para o conjunto de dados *2005 statewide crime* do *site* do livro, considere a taxa de crimes violentos e o percentual de pessoas com renda abaixo do nível de pobreza. Proponha uma questão de pesquisa para estas variáveis relacionada à direção das suas associações, identificando a variável resposta e a explicativa. Usando um *software*, construa um diagrama de dispersão e encontre a correlação. Interprete e indique o que o diagrama de dispersão e a correlação sugerem sobre a pergunta de pesquisa.

3.58 Considere o Exercício 3.6. Proponha uma questão de pesquisa relacionando a correlação entre o gasto público com saúde e o número de médicos por 100000 habitantes. Usando um *software*, analise os dados da Tabela 3.11 para tratar desta pergunta e resuma sua análise fornecendo conclusões.

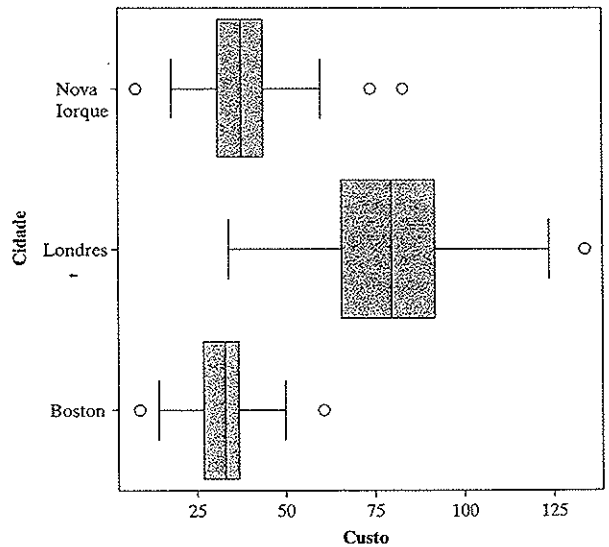
3.59 O guia de restaurantes Zagat publica avaliações de restaurantes para muitas cidades grandes ao redor do mundo (veja www.zagat.com). A análise crítica para cada restaurante fornece um resumo qualitativo assim como um escore de 0 a 30 pontos sobre a qualidade da comida, apresentação, serviço e o custo do jantar com uma bebida e gorjeta. A Figura 3.21 mostra um diagrama de caixa e bigodes lado a lado para restaurantes italianos de Boston, Londres e Nova Iorque (para os bairros Little Italy e Greenwich Village). Resuma o que você descobriu a partir desses diagramas.

☑ **Tabela 3.17**

País	Mortes por arma de fogo	País	Mortes por arma de fogo	País	Mortes por arma de fogo
Austrália	1,7	Grécia	1,8	Noruega	2,6
Áustria	3,6	Islândia	2,7	Portugal	2,1
Bélgica	3,7	Irlanda	1,5	Espanha	0,7
Canadá	3,1	Itália	2,0	Suécia	2,1
Dinamarca	1,8	Japão	0,1	Suíça	6,2
Finlândia	4,4	Luxemburgo	1,9	Reino Unido	0,3
França	4,9	Holanda	0,8	EstadosUnidos	9,4
Alemanha	1,5	Nova Zelândia	2,3		

Fonte: Small Armas Survey, Genebra, 2007.

3.60 Considere o exercício anterior. Os dados estão disponíveis no arquivo *Zagat restaurant rating of Italian restaurants* no *site* do livro. Para os 83 restaurantes listados em Londres, a avaliação da qualidade da comida tem uma correlação de 0,61 com a avaliação da apresentação, 0,81 com a avaliação do serviço e 0,53 com a avaliação do custo. Resuma o que você descobriu destas correlações.



☑ **Figura 3.21** Diagrama de caixa e bigodes lado a lado para restaurantes italianos de Boston.

3.61 O Exercício 3.21 introduziu o Índice de Desenvolvimento Humano (IDH). Vá a hdr.undp.org/statistics e obtenha as últimas avaliações do IDH para os países da África Subsaariana e separadamente para os países ocidentais listados no *site* como *High-income OECD*. (Uma forma de fazer isto é clicar em [*Access a wide range of data tools*] e então [*Build a table*] e fazer as escolhas apropriadas.) Usando métodos gráficos e numéricos deste capítulo, resuma os dados.

3.62 As rendas dos jogadores do time de beisebol do New York Yankees em 2006 podem ser resumidas pelos números[5] $2925000 e $7095078. Um destes números é a mediana e o outro a média. Qual valor você acha que é a média? Por quê?

3.63 Em 2001, o Banco Central dos Estados Unidos amostrou aproximadamente 4000 domicílios para estimar o patrimônio líquido geral de uma família. O Banco Central relatou as estatísticas $86100 e $395500. Um destes valores é a média e o outro a mediana. Qual deles você acha que é a mediana? Por quê?

3.64 Um estudo do Banco Central dos Estados Unidos em 2000 indicou que, para as famílias com renda anual acima de $100000, o patrimônio líquido mediano era de aproximadamente de $500000 tanto em 1995 quanto em 1998, mas seu patrimônio líquido médio aumentou de $1,4 milhões em 1995 para $1,7 milhões em 1998. Um artigo no jornal sobre isto diz que a média usa "um cálculo que capturou enormes ganhos conseguidos pelos norte-americanos mais ricos". Por que a mediana não iria necessariamente capturar estes ganhos?

3.65 A taxa de fertilidade (número médio de filhos por mulher adulta) varia nos países da Europa Ocidental entre um 1,3 baixo (Itália e Espanha) e um 1,9 alto (Irlanda). Para cada mulher, o número de filhos é um número inteiro, como 0, 1 ou 2. Explique por que faz sentido mensurar um número médio de filhos por mulher adulta (que não é um número inteiro); por exemplo, para comparar estas taxas entre países europeus ou com o Canadá (1,5), os Estados Unidos (2,0) e México (2,4).

3.66 De acordo com um relatório do U.S. National Center for Health Statistics (Centro Nacional dos Estados Unidos para Estatísticas da Saúde), para homens com idade entre 25-34 anos, 2% das suas alturas são 64 polegadas ou menos, 8% têm 66 polegadas ou menos, 27% têm 68 polegadas ou menos, 39% têm 69 polegadas ou menos, 54% têm 70 polegadas ou menos, 68% têm 71 polegadas ou menos, 80% têm 72 polegadas ou menos 93% têm 74 polegadas ou menos, e 98% têm 76 polegadas ou menos. Estes são chamados de *percentuais cumulativos*.

(a) Encontre a altura mediana masculina.

(b) Aproximadamente todas as alturas estão entre 60 e 80 polegadas, com menos do que 1% ficando fora desse intervalo. Se as alturas têm aproximadamente uma forma de sino, dê uma estimativa para o desvio padrão. Explique o seu raciocínio.

3.67 Dê um exemplo de uma variável para a qual a moda se aplica, mas não a média ou a mediana.

3.68 Dê um exemplo de uma variável que tenha uma distribuição que você espera ser:

(a) aproximadamente simétrica.

(b) assimétrica à direita.

(c) assimétrica à esquerda.

(d) bimodal.

(e) assimétrica à direita, com a moda e a mediana iguais a zero, mas com uma média positiva.

3.69 Para mensurar o centro de um conjunto de dados, por que

(a) a mediana é, algumas vezes, preferida à média?

(b) a média é, algumas vezes, preferida à mediana?

Em cada caso, dê um exemplo para ilustrar a sua resposta.

3.70 Para mensurar a variabilidade, por que

(a) o desvio padrão s é geralmente preferido à amplitude (intervalo)?

(b) o IIQ é, algumas vezes, preferido ao s?

3.71 Responda verdadeiro ou falso ao que segue:

(a) A média, a mediana e a moda nunca podem ser iguais.

(b) A média é sempre um dos valores do conjunto de dados.

(c) A mediana é igual ao $2^{\underline{o}}$ quartil e ao $50^{\underline{o}}$ percentil.

(d) Para 67 sentenças por assassinato recentemente impostas pelas diretrizes da Comissão de Sentenças dos Estados Unidos, a duração mediana foi de 160 meses e a média de 251 meses. Esta distribuição é provavelmente assimétrica à direita.

Para os problemas de escolha múltipla 3.72 a 3.74, selecione a melhor resposta.

3.72 No Canadá, baseado no censo de 2001, as categorias para a afiliação religiosa (católico, protestante, outra cristã, muçulmano, judeu, nenhuma, outra) apresentaram as seguintes frequências relativas (42%, 28%, 4%, 2%, 1%, 16%, 7%) (*Statistics Canada*).

(a) A religião mediana é a protestante.

(b) Somente 2,7% dos sujeitos estão entre um desvio padrão da média.

(c) A moda é a categoria "católica".

(d) A categoria "judeu" é um valor atípico.

3.73 A PSG de 2004 perguntou se fazer sexo antes do casamento é (sempre errado, quase sempre errado, somente algumas vezes errado, não é errado). As contagens da resposta nestas quatro categorias foram (238, 79, 157, 409). A distribuição é

(a) assimétrica à direita.

(b) aproximadamente em forma de sino.

(c) bimodal.

(d) a forma não faz sentido visto que a variável é nominal.

3.74 Em um estudo de alunos formandos que fizeram o *Graduate Record Exam* (GRE), o Educational Testing Service (Serviço de Avaliação Educacional) relatou que, para o exame quantitativo, os cidadãos norte-americanos tiveram uma média de 529 e um desvio padrão de 127, enquanto os cidadãos não norte-americanos tiveram uma média de 649 e um desvio padrão de 129.

(a) Ambos os grupos tiveram a mesma variabilidade nos seus escores, mas os cidadãos não norte-americanos tiveram uma melhor performance, em média, do que os cidadãos norte-americanos.

(b) Se a distribuição dos escores tivesse uma forma aproximada de sino, então significaria que quase nenhum cidadão não norte-americano teria um escore abaixo de 400.

(c) Se os escores tiveram um intervalo entre 200 e 800, então, provavelmente, os escores para os cidadãos não norte-americanos foram simétricos e com a forma de sino.

(d) Um cidadão não norte-americano que teve um escore de três desvios padrão abaixo da média teve um escore de 200.

3.75 Uma professora resume as notas de um exame de meio do semestre por:

Mín. = 26, Q1 = 67, Mediana = 80,
Q3 = 87, Máx. = 100,
Média = 76, Moda = 100,
Desvio padrão = 76, IIQ = 20.

Ela registrou incorretamente uma delas. Qual você acha que foi? Por quê?

3.76 Dez pessoas são selecionadas aleatoriamente na Flórida e outras dez pessoas são selecionadas aleatoriamente no Alabama. A Tabela 3.18 fornece um resumo informativo sobre o rendimento médio. A média é alta no Alabama tanto nas áreas rurais quanto nas áreas urbanas. Qual o estado que tem o rendimento médio geral mais alto? (A razão para esse paradoxo aparente é que a média das rendas urbanas é maior do que a média das rendas rurais para ambos os estados e a amostra da Flórida tem uma proporção maior de residentes urbanos.)

☑ **Tabela 3.18**

Estado	Rural	Urbano
Flórida	\$26000 ($n = 3$)	\$39000 ($n = 7$)
Alabama	\$27000 ($n = 8$)	\$40000 ($n = 2$)

3.77 Considere a Tabela 3.2 (página 51). Explique por que a média destas 50 observações não é necessariamente a mesma da taxa de crimes violentos para toda a população dos Estados Unidos.

3.78 Para uma amostra com média $\bar{y}$, adicionar uma constante c a cada observação altera a média para $\bar{y} + c$ e o desvio padrão s não muda. Multiplicar cada observação por c muda a média para $c\bar{y}$ e o desvio padrão para $|c|s$.

(a) Os escores de um exame difícil tiveram uma média de 57 e um desvio padrão de 20. O professor aumenta todos os escores em 20 pontos antes de dar as notas. Determine a média e o desvio padrão dos escores aumentados.

(b) Suponha que o rendimento anual dos advogados canadenses tem uma média de \$100000. Os valores são convertidos a libras britânicas para uma apresentação a uma audiência britânica. Se uma libra britânica é igual a \$2,00, determine a média e o desvio padrão da renda expressas na moeda britânica.

(c) As observações de um levantamento de dados que perguntou sobre o número de milhas viajadas todos os dias num transporte público devem ser convertidas para quilômetros (1 milha = 1,6 quilômetros). Explique como encontrar a média e o desvio padrão das observações convertidas.

***3.79** Mostre que $\Sigma = (y_i - \bar{y})$ deve ser igual a 0 para qualquer conjunto de observações $y_1, y_2, \ldots y_n$.

***3.80** O matemático russo Tchebysheff provou que para todo $k > 1$, a proporção das observações que estão mais do que k desvios padrão da média não pode ser maior do que $1/k^2$. Isto é válido para *qualquer* a distribuição e não apenas para aquelas com forma de sino.

(a) Encontre o limite superior para a proporção de observações que estão (i) mais do que dois desvios padrão da média, (ii) mais do que três desvios padrão da média, (iii) mais do que dez desvios padrão da média.

(b) Compare o limite superior para $k = 2$ com a proporção aproximada que está a mais do que dois desvios padrão da média em uma distribuição com forma de sino. Por que existe uma diferença?

***3.81** A propriedade dos **mínimos quadrados** para a média estabelece que os dados estão mais próximos de $\bar{y}$ do que de qualquer outro número c, no sentido de que a soma dos quadrados dos desvios dos dados em torno de sua média é menor do que a soma dos quadrados dos seus desvios em torno de c. Isto é,

$$\Sigma(y_i - \bar{y})^2 < \Sigma(y_i - c)^2.$$

Se você estudou cálculo, prove esta propriedade tratando $f(c) = \Sigma(y_i - c)^2$ como uma função de c que fornece um mínimo. (Dica: faça a derivada de f(c) em relação a c e igual a zero.)

☑ NOTAS

[1] *Fonte:* Tabela 8.9 em www.stateofworkingamerica.org, do The Economic Policy Institute (Instituto de Política Econômica).
[2] Dados fornecidos por Todd Kamhoot, Gainesville Regional Utilities.
[3] Dados fornecidos pelo Dr. Michael Conlon, Universidade da Flórida.
[4] *OECD Key Environmental Indicators 2005.*
[5] http://usatoday.com/sports/baseball/mlbsalaries/team

☑ 4 DISTRIBUIÇÕES DE PROBABILIDADE

Comparada à maioria das ciências matemáticas, a estatística é recente. A maioria dos métodos discutidos neste livro foi desenvolvida no século passado. Ao contrário, a probabilidade, o assunto deste capítulo, tem uma longa história. Por exemplo, os matemáticos usavam a probabilidade na França no século XVII para avaliar as várias estratégias de jogo. A probabilidade é um assunto altamente desenvolvido, mas este capítulo limita sua atenção ao básico de que iremos necessitar para a inferência estatística.

Após uma breve introdução à probabilidade na Seção 4.1, as Seções 4.2 e 4.3 apresentam as *distribuições de probabilidade,* as quais fornecem probabilidades para todos os resultados possíveis de uma variável. A *distribuição normal,* descrita por uma curva em forma de sino, é a distribuição de probabilidade mais importante para a análise estatística. As Seções 4.4 e 4.5 introduzem a *distribuição amostral,* um tipo de distribuição de probabilidade de fundamental importância para a inferência estatística. Ela nos permite prever quão próximo a média amostral está da média da população. Veremos que a razão principal para a importância da distribuição normal é o resultado notável de que as distribuições amostrais apresentam, em geral, a forma de sino, isto é, tendem a normal.

4.1 INTRODUÇÃO À PROBABILIDADE

No Capítulo 2, aprendemos que a aleatorização é a componente-chave de um bom método de coleta de dados. Considere uma amostra aleatória hipotética ou um experimento aleatório. Para cada situação os resultados possíveis são conhecidos, mas não sabemos ao certo qual deles irá ocorrer.

A probabilidade como uma frequência relativa de muitas repetições

Para um resultado possível em particular, de um fenômeno aleatório, a *probabilidade* é a proporção das vezes em que o resultado irá ocorrer em uma sequência bastante longa de observações ou repetições.

☑ **Probabilidade**

Com uma amostra ou um experimento aleatório, a **probabilidade** de ocorrência de um resultado, em particular, é a proporção de vezes em que o resultado é obtido em uma longa sequência de observações ou repetições.

Mais tarde neste capítulo, iremos analisar os dados para a eleição governamental da Califórnia em 2006, para a qual o vencedor foi o candidato Republicano Arnold