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

☑ **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

Schwarzenegger. Imagine o processo de entrevistar uma amostra aleatória de eleitores daquela eleição e perguntar em quem eles votaram. À medida que entrevistamos mais e mais pessoas, a proporção amostral dos que dizem ter votado em Schwarzenegger se aproxima mais e mais da proporção da população que votou nele. Elas serão as mesmas depois de entrevistarmos toda a população de eleitores. Suponha que a proporção de eleitores na população seja de 0,56. Então, a probabilidade de que uma pessoa aleatoriamente selecionada tenha votado em Schwarzenegger é de 0,56.

Por que a probabilidade se refere a muitas repetições? Porque precisamos de um grande número de observações para avaliar com precisão esse tipo de probabilidade. Se você amostrar apenas dez pessoas e elas são todas destras, você não pode concluir que a probabilidade de ser destro é igual a 1,0.

Este livro define a probabilidade como uma proporção, assim ela é um número entre 0 e 1. Na prática, as probabilidades são geralmente expressas também como percentuais, estando, então, entre 0 e 100. Por exemplo, se o meteorologista diz que a probabilidade de chuva, hoje, é de 70%, isto quer dizer que em uma longa série de dias com condições atmosféricas semelhante as de hoje, a chuva ocorreu em 70% dos dias.

A abordagem de muitas repetições para definir a probabilidade nem sempre é útil. Se você decidir iniciar um novo negócio, não terá muitas tentativas para estimar a probabilidade de que o seu negócio será um sucesso. Você deve, então, contar com informações *subjetivas* em vez de apenas dados *objetivos*. Na abordagem subjetiva, a probabilidade de um resultado é definida como sendo o seu grau de crença de que o resultado irá ocorrer baseado na informação disponível. Existe um ramo da estatística que utiliza a probabilidade subjetiva como base. Ela é denominada **estatística bayesiana,** em homenagem ao clérigo britânico Thomas Bayes que descobriu uma regra probabilística na qual essa estatística está baseada. Essa abordagem está além do propósito deste livro.

Regras básicas de probabilidade

Não é o propósito deste livro entrar em detalhes sobre as muitas regras para determinar probabilidades. Aqui, iremos mencionar brevemente quatro regras que são especialmente úteis. Não iremos tentar explicá-las com um raciocínio matemático preciso porque para os nossos propósitos é suficiente ter uma percepção intuitiva do que cada regra diz.

Considere $P(A)$ a representação da probabilidade de um resultado possível ou de um conjunto de resultados representados pela letra A. Então:

1. **P(não A) = 1 – $P(A)$.**
   Se você sabe a probabilidade de que um resultado ocorra, então a probabilidade de que ele *não* ocorra é 1 menos aquela probabilidade. Suponha que A represente o resultado de que um eleitor selecionado aleatoriamente vote em Schwarzenegger. Se $P(A) = 0{,}56$, então $1 - 0{,}56 = 0{,}44$ é a probabilidade de que ele *não* vote em Schwarzenegger, isto é, ao contrário, votar no candidato Democrata ou em outro candidato da cédula eleitoral.
2. **Se A e B são resultados distintos possíveis (sem sobreposição), então $P(A$ ou $B) = P(A) + P(B)$.**
   Suponha que você faça um levantamento de dados para estimar a proporção da população de pessoas que acredita que a pesquisa com células-tronco deva ser proibida pelo governo federal. Considere A a representação que você obteve em uma estimativa da proporção amostral que é muito baixa, ficando mais do que 0,10 *abaixo* da proporção populacional. Considere B a representação de sua estimativa da proporção amostral como sendo muito alta – pelo menos 0,10 *acima* da proporção populacional. Usando os métodos deste capítulo, talvez você encontre que $P(A) = P(B) = 0{,}03$. Então a probabilidade de que a proporção amostral esteja errada por mais do que 0,10 (sem especificar a direção do erro) é dada por:

$$P(A \text{ ou } B) = P(A) + P(B) = 0{,}03 + 0{,}03 = 0{,}06.$$

3. **Se A e B são resultados possíveis, então $P(A$ e $B) = P(A) \times P(B$ dado $A)$.**
   Pelos dados do censo norte-americano, a probabilidade de que um adulto aleatoriamente selecionado seja casado é igual a 0,56. Daqueles que são casados, a Pesquisa Social Geral indica que a probabilidade de que uma pessoa diga que está *muito feliz* quando solicitada a escolher entre (muito feliz, moderadamente feliz, não muito feliz) é de aproximadamente 0,40; isto é, dado que você seja casado, a probabilidade de que esteja muito feliz é de 0,40. Assim:

$$P(\text{casado e muito feliz}) = P(\text{casado}) \times P(\text{muito feliz dado que é casado}) = 0{,}56 \times 0{,}40 = 0{,}22.$$

   Aproximadamente 22% da população adulta é casada *e* muito feliz.
   Em alguns casos, A e B são "independentes", no sentido de que a ocorrência de um não depende da do outro. Isto é, $P(B \text{ dado } A) = P(B)$, assim a regra anterior simplifica para:
4. **Se A e B são independentes, então $P(A$ e $B) = P(A) \times P(B)$.**
   Por exemplo, um método de inferência apresentado no próximo capítulo geralmente é usado com a probabilidade de a inferência, para um dado conjunto, estar correta como sendo 0,95. Suponha que A represente uma inferência sobre homens na população de interesse (como uma previsão sobre a proporção de homens que votaram em Schwarzenegger) como sendo correta. Considere B a representação de uma inferência separada sobre mulheres como sendo correta. Então, visto que estas são amostras e inferências independentes, a probabilidade de que as duas inferências estejam corretas é:

$$P(A \text{ e } B) = P(A) \times P(B) = 0{,}95 \times 0{,}95 = 0{,}90.$$

4.2 DISTRIBUIÇÕES DE PROBABILIDADE PARA VARIÁVEIS DISCRETAS E CONTÍNUAS

Uma variável pode assumir, pelo menos, dois valores diferentes. Para uma amostra ou experimento aleatório, cada resultado possível tem uma probabilidade de ocorrer. A variável, propriamente, é, então, algumas vezes referida como uma **variável aleatória**. Essa terminologia enfatiza que o resultado varia de observação para observação de acordo com uma variação aleatória que pode ser resumida por probabilidades. Continuaremos a usar a terminologia mais simples "variável".

Relembre, da Seção 2.1, que uma variável é *discreta* se os resultados possíveis forem um conjunto de valores separados, por exemplo, uma variável expressada como "o número de..." com valores possíveis 0, 1, 2, ... Ela é *contínua* se os resultados possíveis forem um infinito contínuo. Uma **distribuição de probabilidade** lista os resultados possíveis e suas probabilidades. Veremos a seguir como isso é feito para as variáveis discretas e contínuas.

Distribuições de probabilidade para variáveis discretas

A distribuição de probabilidade de uma variável *discreta* atribui uma probabili-

dade a cada valor possível de uma variável. Cada probabilidade é um número entre 0 e 1. A soma das probabilidades de todos os valores possíveis é igual a 1.

Considere $P(y)$ a representação da probabilidade de um possível resultado para uma variável y. Então:

$$0 \leq P(y) \leq 1 \text{ e } \Sigma_{\text{todos } y} P(y) = 1,$$

onde a soma é sobre todos os valores possíveis da variável.

EXEMPLO 4.1 Número ideal de filhos de uma família

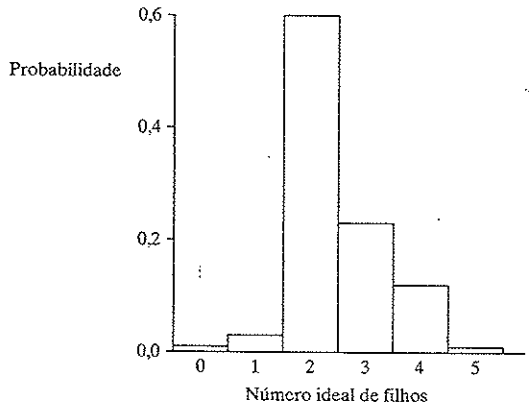
Considere y aos valores da resposta para a pergunta: "Qual é o número ideal de filhos que você acha que uma família deva ter?". Trata-se de uma variável discreta, assumindo os valores possíveis 0, 1, 2, 3 e assim por diante. De acordo com resultados de uma PSG, para uma pessoa escolhida aleatoriamente nos Estados Unidos, a distribuição de probabilidade de y é aproximadamente como a Tabela 4.1 mostra. A tabela mostra os valores y observados e suas probabilidades. Por exemplo, $P(4)$ é a probabilidade de que $y = 4$ filhos seja considerado ideal e é igual a 0,12. Cada probabilidade na Tabela 4.1 está entre 0 e 1, e a soma de todas é igual a 1. ■

Um *histograma** pode representar a distribuição de probabilidade. Uma barra retangular sobre um valor possível da variável tem a altura igual à probabilidade daquele valor. A Figura 4.1 é um histograma para a distribuição de probabilidade do número ideal de filhos da Tabela 4.1. A barra sobre o valor 4 tem a altura de 0,12, a probabilidade do resultado 4.

☑ **Tabela 4.1** Distribuição de probabilidade de y = número ideal de filhos de uma família

y	P(y)
0	0,01
1	0,03
2	0,60
3	0,23
4	0,12
5	0,01
Total	1,00

* N. de T. T.: De fato, aqui o diagrama ideal seria o de colunas simples, em vez de um histograma.



☑ **Figura 4.1** Histograma para a distribuição de probabilidade do número ideal de filhos de uma família.

Distribuições de probabilidade para variáveis contínuas

As variáveis *contínuas* têm um contínuo infinito de possíveis valores. As distribuições de probabilidade de variáveis contínuas designam probabilidades aos *intervalos* de números. A probabilidade de que uma variável caia em qualquer intervalo particular está entre 0 e 1, e a probabilidade de o intervalo conter todos os valores possíveis é igual a 1.

Um gráfico de uma distribuição da probabilidade de uma variável contínua é uma curva contínua suave. A *área* abaixo da curva para um intervalo de valores representa a probabilidade de que a variável assuma um valor naquele intervalo.

EXEMPLO 4.2 Tempo de viagem até o local de trabalho

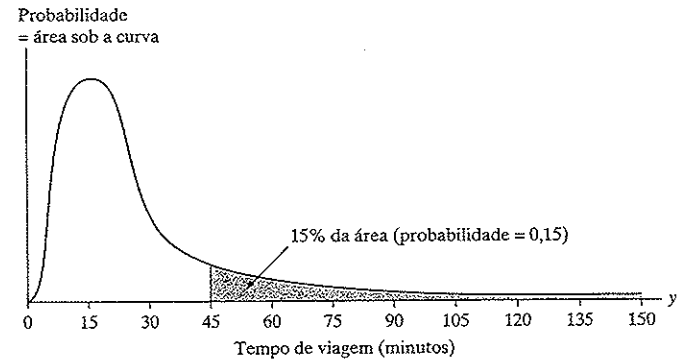
Um estudo recente sobre o tempo de viagem para chegar ao trabalho para trabalhadores norte-americanos que precisam viajar[1] mensurou y = tempo de viagem (em minutos). A distribuição de probabilidade de y fornece probabilidades como $P(y < 10)$, a probabilidade de que o tempo de viagem seja menor do que 10 minutos, ou $P(30 < y < 60)$, a probabilidade de que o tempo de viagem esteja entre 30 e 60 minutos.

A Figura 4.2 representa a distribuição de probabilidade aproximada de y. A área sombreada na figura se refere à região de tempos maiores do que 45 minutos. Essa área é igual a 15% da área total sob a curva, informando que existe uma probabilidade de 0,15 de que o tempo de viagem seja maior do que 45 minutos. As regiões nas quais a curva tem alturas relativamente altas representam maiores probabilidades de se observar valores. ■

Os parâmetros descrevem distribuições de probabilidade

A maioria das distribuições de probabilidade tem fórmulas para calcular as probabilidades. Para as outras, as tabelas e gráficos fornecem as probabilidades. A Seção 4.3 mostra como calcular as probabilidades para a distribuição de probabilidade mais importante.

A Seção 3.1 introduziu a *distribuição da população* de uma variável. Isto é, de modo equivalente, a distribuição de probabilidade de uma variável para um sujeito selecionado aleatoriamente da população. Por exemplo, se 0,12 é a proporção da população de adultos que acreditam que o número ideal de filhos é 4, então a pro-



☑ **Figura 4.2** Distribuição de probabilidade do tempo de viagem até o trabalho. A área sob a curva entre dois pontos representa a probabilidade daquele intervalo de valores.

babilidade de que um adulto selecionado aleatoriamente daquela população acredite nisso é, também, 0,12.

Como a distribuição da população, a distribuição de probabilidade tem *parâmetros* que descrevem o centro e a variabilidade. A *média* descreve o centro, e o *desvio padrão* descreve a variabilidade. Os valores do parâmetro são os valores que estas medidas iriam assumir em uma grande quantidade de repetições, se o experimento ou a amostra aleatórios fossem repetidamente observados ou extraídos da variável y tendo aquela distribuição de probabilidade.

Por exemplo, suponha que façamos observações da distribuição da Tabela 4.1. Em um grande número de repetições esperamos que $y = 0$ ocorra 1% das vezes, $y = 1$ ocorra 3% das vezes, e assim por diante. Em 100 observações, por exemplo, esperaríamos aproximadamente:

apenas um valor 0, três vezes o valor 1, sessenta vezes o valor 2, vinte e três vezes o valor 3, doze vezes o valor 4, e somente um valor 5.

Nesse caso, visto que a média é igual ao total de observações dividida pelo tamanho da amostra, a média seria igual a:

$$[(1)0 + (3)1 + (60)2 + (23)3 + 12(4) + 1(5)]/100 = \frac{245}{100} = 2{,}45.$$

Esse cálculo pode assumir o seguinte formato:

$$0(0{,}01) + 1(0{,}03) + 2(0{,}60) + 3(0{,}23) + 4(0{,}12) + 5(0{,}01),$$

a soma dos possíveis resultados vezes as suas probabilidades. Na verdade, para qualquer variável discreta y, a média da sua distribuição da probabilidade tem essa forma.

☑ A média de uma distribuição da probabilidade (valor esperado)

A **média da distribuição da probabilidade** para uma variável discreta y é

$$\mu = \sum yP(y).$$

A soma é realizada sobre todos os valores possíveis da variável. Esse parâmetro também é chamado de **valor esperado de y** e é representado por $E(y)$.

Para a Tabela 4.1, por exemplo,

$$\mu = \sum yP(y) = 0P(0) + 1P(1) + 2P(2) + 3P(3) + 4P(4) + 5P(5)$$
$$= 0(0{,}01) + 1(0{,}03) + 2(0{,}60) + 3(0{,}23) + 4(0{,}12) + 5(0{,}01)$$
$$= 2{,}45.$$

Esse é também o *valor esperado* de y, $E(y) = \mu = 2{,}45$. A terminologia reflete que $E(y)$ representa o que esperamos para o valor médio de y em uma longa série de repetições das observações ou extrações.

O **desvio padrão** de uma distribuição da probabilidade, representado por σ, avalia sua variabilidade. Quanto maior o valor de σ, maior a dispersão da distribuição. Em um sentido geral, σ descreve quão longe a variável y tende a estar da sua média. A Regra Empírica (Seção 3.3) nos ajuda a interpretar σ. Se a distribuição de probabilidade tem a forma aproximada de sino, aproximadamente 68% dos valores está entre $\mu - \sigma$ e $\mu + \sigma$, aproximadamente 95% está entre $\mu - 2\sigma$ e $\mu + 2\sigma$, e toda ou quase toda a distribuição está entre $\mu - 3\sigma$ e $\mu + 3\sigma$. Por exemplo, suponha que o tempo de viagem para o trabalho de um trabalhador aleatoriamente selecionado em Toronto tenha uma distribuição de probabilidade com uma forma de sino com $\mu = 24$ minutos e $\sigma = 8$ minutos. Assim,

existe uma probabilidade de aproximadamente 95% de que o tempo de viagem esteja entre $24 - 2(8) = 8$ minutos e $24 + 2(8) = 40$ minutos.

O desvio padrão é a raiz quadrada da **variância** da distribuição da probabilidade. A variância é a média dos quadrados dos desvios das observações em relação à média. Isto é, ela é o valor esperado de $(y - \mu)^2$. Não precisaremos calcular essa medida, assim não estudaremos esta fórmula aqui. (O Exercício 4.55 mostra a fórmula para σ para o caso de uma variável discreta.)

4.3 A DISTRIBUIÇÃO DE PROBABILIDADE NORMAL

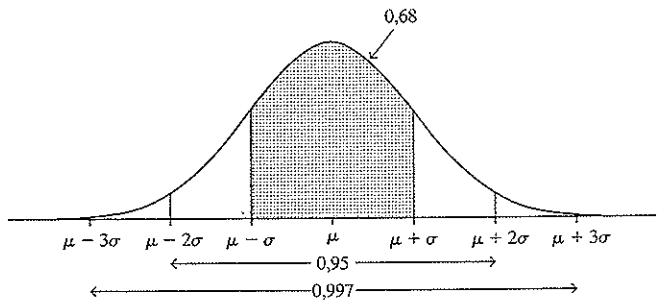
Algumas distribuições de probabilidade são importantes porque aproximam bem as distribuições das variáveis do mundo real. Algumas são importantes por causa do seu uso na inferência estatística. Esta seção introduz a **distribuição de probabilidade normal**, que é importante por duas razões. Sua curva em forma de sino descreve bem muitos histogramas de dados de muitas variáveis contínuas ou que assumem um grande número de possíveis valores. Ela é a distribuição mais importante para a inferência estatística, pois veremos que ela é útil mesmo quando os dados amostrais *não* têm a forma de sino.

☑ Distribuição normal

A **distribuição normal** é simétrica, com forma de sino e caracterizada por sua média μ e desvio padrão σ. A probabilidade dentro de qualquer número de desvios padrão em torno de μ é a mesma para todas as distribuições normais. Essa probabilidade é igual a 0,68 para um desvio padrão, 0,95 para dois desvios padrão e 0,997 para três desvios padrão.

Cada distribuição normal é especificada por dois parâmetros – a média μ e o desvio padrão σ. Para todo número real μ e todo número real não negativo σ, existe uma distribuição normal tendo essa média e desvio padrão. A Figura 4.3 ilustra. Essencialmente, toda a distribuição está entre $\mu - 3\sigma$ e $\mu + 3\sigma$.

Por exemplo, as alturas das mulheres adultas norte-americanas têm aproximadamente uma distribuição normal com média $\mu = 65{,}0$ polegadas e desvio padrão $\sigma = 3{,}5$ polegadas. A probabilidade é de aproximadamente 1,0 de que uma mulher aleatoriamente selecionada tenha uma altura entre $\mu - 3\sigma = 65{,}0 - 3(3{,}5) = 54{,}5$ polegadas e $\mu + 3\sigma = 65{,}0 + 3(3{,}5) = 75{,}5$ polegadas. A altura do homem norte-americano adulto tem uma distribuição normal com $\mu = 70{,}0$ e $\sigma = 4{,}0$ polegadas. Assim, a probabilidade é de aproximadamente 1,0



☑ **Figura 4.3** Para cada distribuição normal, a probabilidade é igual a (aproximadamente) 0,68 para um σ em torno de μ, 0,95 para 2σ em torno de μ, e 0,997 para 3σ em torno de μ.

de que um homem norte-americano aleatoriamente selecionado tenha uma altura entre $\mu - 3\sigma = 70{,}0 - 3(4{,}0) = 58$ polegadas e $\mu + 3\sigma = 70{,}0 + 3(4{,}0) = 82$ polegadas. Veja a Figura 4.4.

Probabilidades tabeladas da cauda direita da normal

Para a distribuição normal, para cada número fixo z, a probabilidade que está entre z desvios padrão a contar da média depende somente do valor de z. Essa é a área sob a curva normal em forma de sino $\mu - z\sigma$ e $\mu + z\sigma$. Para cada distribuição normal, esta probabilidade é de 0,68 para $z = 1$; 0,95 para $z = 2$; e aproximadamente 1,0 para $z = 3$.

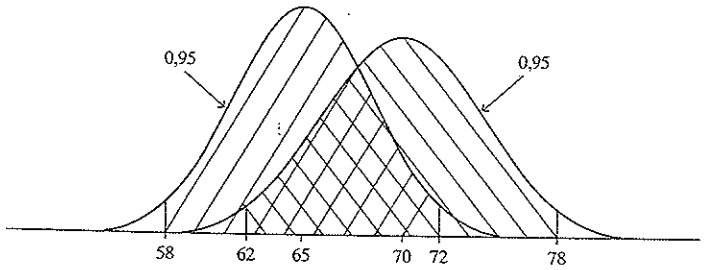
Para uma distribuição normal, a probabilidade concentrada dentro de $z\sigma$ de μ é a mesma para todas as curvas normais mesmo se z não for um número inteiro – por exemplo, $z = 1{,}43$ em vez de 1, 2 ou 3. A Tabela A (página 650) em Tabelas fornece a probabilidade da cauda à direita da curva para vários valores. Ela apresenta a probabilidade (área) para os valores de z que estão na cauda direita (iniciando em zero) com precisão centesimal. A coluna da margem esquerda da tabela lista os valores de z com uma casa decimal, com a segunda casa decimal (centésimo) sendo apresentada na linha acima das colunas.

A Tabela 4.2 exibe uma pequena parte da Tabela A. A probabilidade para $z = 1{,}43$ está na interseção da linha do 1,4 (unidade e décimo) com a coluna do 0,03 (centésimo) e é igual a 0,0764. Isso significa que para cada distribuição normal, a probabilidade da cauda direita acima de $\mu + 1{,}43\sigma$ (isto é, mais do que 1,43 desvios padrão acima da média) é igual a 0,0764.

Visto que os valores da Tabela A são probabilidades para a metade direita da distribuição normal acima de $\mu + z\sigma$, elas estão compreendidas no intervalo de 0 a 0,50. Pela simetria da curva normal, essas probabilidades da cauda direita também se aplicam à cauda esquerda abaixo de $\mu - z\sigma$. Por exemplo, a probabilidade abaixo de $\mu - 1{,}43\sigma$ também é igual a 0,0764. As probabilidades da cauda esquerda, chamadas de *probabilidades cumulativas*, podem ser obtidas por muitas calculadoras e *softwares*.

Probabilidades normais e a Regra Empírica

As probabilidades na Tabela A se aplicam à distribuição normal e também se aplicam aproximadamente a outras distribuições com forma de sino. Essa tabela produz as probabilidades para a Regra Empírica. Essa regra declara que, para histogramas com forma de sino, aproximadamente 68% dos dados estão a um desvio padrão em torno da média, 95% estão entre dois desvios padrão e todos ou quase todos estão entre três desvios padrão.



☑ **Figura 4.4** Distribuição normal para a altura das mulheres ($\mu = 65$, $\sigma = 3{,}5$) e dos homens ($\mu = 70$, $\sigma = 4{,}0$) norte-americanos.

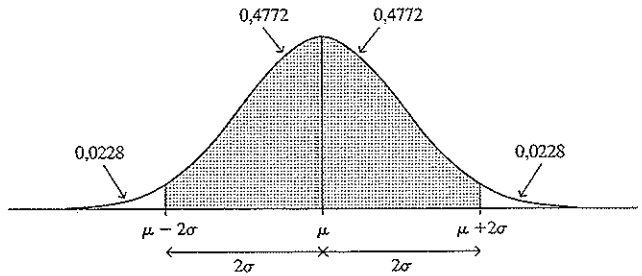
☑ **Tabela 4.2** Parte da Tabela A que exibe probabilidades da cauda direita da curva normal padrão

	Segunda casa decimal de z									
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
						...				
						...				
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
						...				
						...				

Por exemplo, o valor de dois desvios padrão acima da média tem um valor-z de 2,00. A probabilidade da curva normal listada na Tabela A oposta a $z = 2{,}00$ é 0,0228. A probabilidade da cauda direita, acima de $\mu + 2\sigma$, é igual a 0,0228 para qualquer distribuição normal. A probabilidade da caixa esquerda abaixo de $\mu - 2\sigma$ é também 0,0228, por simetria (veja Figura 4.5). A probabilidade total de mais do que dois desvios padrão da média é $2(0{,}0228) = 0{,}0456$. Uma vez que a probabilidade além de dois desvios padrão da média é 0,0456, a probabilidade de termos valores entre $\mu - 2\sigma$ e $\mu + 2\sigma$ (isto é, dentro de dois desvios padrão da média) é igual a $1 - 0{,}0456 = 0{,}9544$. (Aqui, usamos a regra (1) das regras da probabilidade do final da Seção 4.1, que é $P(\text{não } A) = 1 - P(A)$.) Quando a variável tem uma distribuição normal, aproximadamente 95% das observações estão dentro de dois desvios padrão a contar da média.

A probabilidade é igual a 0,50 acima da média, visto que a distribuição normal é simétrica em torno de μ. Assim, a probabilidade entre μ e $\mu + 2\sigma$ ou entre $\mu - 2\sigma$ e μ é igual a $0{,}50 - 0{,}0228 = 0{,}4772$, também mostrado na Figura 4.5. Novamente, vemos que a probabilidade total dentro de dois desvios padrão da média é igual a $2(0{,}4772) = 0{,}9544$ ou aproximadamente 95%.

Os percentuais aproximados da Regra Empírica são os percentuais reais para a distribuição normal, arredondados para duas casas decimais. Por exemplo, com a Tabela A você pode verificar que a probabilidade dentro de um desvio padrão da média de uma distribuição normal é igual



☑ **Figura 4.5** A probabilidade dentro de dois desvios padrão da média para uma distribuição normal é igual a $1 - 2(0{,}0228)$, o que é aproximadamente 0,95.

a 0,68. (*Dica:* considere $z = 1{,}00$). A Regra Empírica apresenta os percentuais como *aproximações* em vez de *exatos*. Por quê? Porque essa regra se refere a *todas as distribuições com aproximadamente a forma de sino*, não apenas a normal. Nem todas as distribuições com forma de sino são normais, somente aquelas descritas pela fórmula matemática mostrada no Exercício 4.56 no final do capítulo. Não precisamos daquela fórmula, mas usaremos as probabilidades tabuladas para ela na Tabela A ao longo de todo o livro.

Encontrando valores-z para certas probabilidades da cauda

Muitos métodos inferenciais usam os valores-z correspondentes a certas probabilidades da curva normal, o que requer o uso inverso da Tabela A. Iniciando com a probabilidade da cauda que está listada no corpo da Tabela A, encontramos o valor-z que fornece o número de desvios padrão que aquele número está da média.

Para ilustrar, vamos encontrar o valor-z que tem uma probabilidade unicaudal à direita de 0,025. Procuramos por 0,025 no corpo da Tabela A. Vemos que ela corresponde a $z = 1{,}96$. Isso significa que existe uma probabilidade de 0,025 acima de $\mu + 1{,}96\sigma$. Da mesma forma, uma probabilidade de 0,025 está abaixo de $\mu - 1{,}96\sigma$. Assim, a probabilidade total de $0{,}025 + 0{,}025 = 0{,}050$ está a mais do que $1{,}96\sigma$ de μ. Vimos na subseção anterior que 95% dos valores de uma distribuição normal estão dentro de dois desvios padrão da média. Mais precisamente, 0,9544 dos valores estão dentro de 2,00 desvios padrão, e aqui vimos que 0,950 estão dentro de 1,96 desvios padrão.

Para checar se você entendeu esse raciocínio, determine que o valor-z para uma probabilidade da cauda direita de 0,05 é $z = 1{,}64$, de 0,01 é $z = 2{,}33$ e de 0,005 é $z = 2{,}58$. Mostre, também, que 90% dos valores de uma distribuição normal estão entre $\mu - 1{,}64\sigma$ e $\mu + 1{,}64\sigma$.

EXEMPLO 4.3 Encontrando o 99º percentil dos escores do QI

Os escores do QI de Stanford-Binet têm aproximadamente uma distribuição normal com média = 100 e desvio padrão = 16. Quanto vale o 99º percentil dos escores do QI? Em outras palavras, qual é o escore do QI que está acima de 99% de todos os escores?

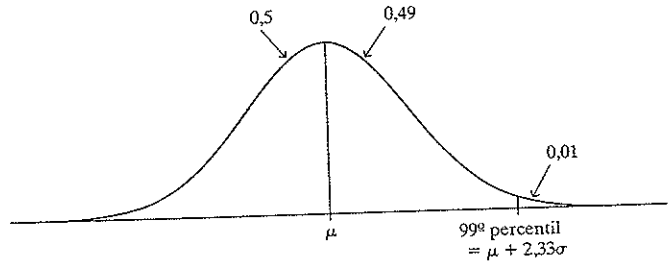
Para responder a isso, precisamos encontrar o valor de z tal que $\mu + z\sigma$ esteja acima de 99% de uma distribuição normal. Usamos a probabilidade da cauda direita da normal além do 99º percentil. Então, podemos usar a Tabela A para encontrar o valor-z correspondente àquela probabilidade. Agora, para $\mu + z\sigma$ representar o 99º percentil, a probabilidade abaixo de $\mu + z\sigma$ deve ser igual a 0,99, pela definição de percentil. Assim, 1% da distribuição está acima do 99º percentil. A probabilidade da cauda direita procurada é igual a 0,01, como mostra a Figura 4.6.

O corpo da Tabela A não contém uma probabilidade da cauda direita exatamente igual a 0,0100. A probabilidade mais próxima por excesso é igual a 0,0102 e corresponde a um valor-$z = 2{,}32$ e a probabilidade mais próxima por falta é igual a 0,0099 e corresponde a um valor-$z = 2{,}33$. Poderíamos interpolar, mas é suficiente usar o valor-z arredondado para duas casas decimais. Selecionamos aquele com a probabilidade próxima ao valor desejado de 0,0100. Assim, o 99º percentil está 2,33 desvios padrão acima da média. Em resumo, 99% de toda distribuição está localizada abaixo de $\mu + 2{,}33\sigma$, não mais do que 2,33 desvios padrão acima da média.

Para os escores do QI com média = 100 e desvio padrão = 16, o 99º percentil será igual a:

$$\mu + 2{,}33\sigma = 100 + 2{,}33(16) = 137.$$

Isto é, aproximadamente 99% dos escores do QI estão abaixo do QI 137. ■



☑ **Figura 4.6** O 99º percentil para uma distribuição tem 99% da distribuição abaixo dele e 1% acima.

Para verificar se você entende o raciocínio acima, mostre que o 95º percentil de uma distribuição normal é $\mu + 1{,}64\sigma$, e mostre, também, que o 95º percentil para a distribuição do QI é igual a 126.

O escore-z é o número de desvios padrão da média

O símbolo z em uma tabela normal se refere à distância entre um possível valor y de uma variável e a média μ da sua distribuição da probabilidade, em termos do *número de desvios padrão* que y está de μ.

EXEMPLO 4.4 Os escores-z para os testes de admissão para a universidade

Os escores em cada parte do Teste de Aptidão Escolar (Scholastic Aptitude Test – SAT), um exame exigido para admissão na universidade, tem tradicionalmente distribuição normal com uma média $\mu = 500$ e desvio padrão $\sigma = 100$. Um escore-y de 650 do teste tem um escore-z de $z = 1{,}50$, porque 650 está 1,50 desvios padrão acima da média. Em outras palavras, $y = 650 = \mu + z\sigma = 500 + z(100)$, onde $z = 1{,}50$. ■

Para dados amostrais, a Seção 3.4 introduziu o escore-z como uma medida de posição. Vamos revisar como encontrá-la. A distância entre y e a média μ é igual a $y - \mu$. O escore-z expressa essa diferença em unidades de desvios padrão.

Escore-z

O escore-z para um valor y de uma variável é o número de *desvios padrão* que y está de μ. Ele é igual a:

$$z = \frac{\text{Observação} - \text{Média}}{\text{Desvio Padrão}} = \frac{y - \mu}{\sigma}.$$

Para ilustrar, quando $\mu = 500$ e $\sigma = 100$, uma observação de $y = 650$ tem um escore-z de:

$$z = \frac{y - \mu}{\sigma} = \frac{650 - 500}{100} = 1{,}50.$$

Escores-z *positivos* ocorrem quando o valor de y está *acima* da média μ. Escores-z *negativos* ocorrem quando o valor de y está *abaixo* da média. Por exemplo, para os escores do SAT com $\mu = 500$ e $\sigma = 100$, um valor de $y = 350$ tem um escore-z de:

$$z = \frac{y - \mu}{\sigma} = \frac{350 - 500}{100} = -1{,}50.$$

O escore do teste de 350 está 1,50 desvios padrão abaixo da média. O valor $y = 350$ está abaixo da média, assim o escore-z é negativo.

A Tabela A contém somente valores-z positivos. Visto que a distribuição normal é simétrica em torno da média, a probabilidade da cauda esquerda abaixo de $-z$ é igual à probabilidade da cauda direita acima de $+z$. Procurando por $z = 1{,}50$ na Tabela A, vemos que a probabilidade de que

um escore do SAT esteja abaixo de 350 é de 0,0668, como mostra a Figura 4.7. Menos do que 7% dos escores estão abaixo de 350 e menos do que 7% estão acima de 650.

O próximo exemplo mostra que os escores-z fornecem uma forma útil para comparar posições para diferentes distribuições normais.

EXEMPLO 4.5 Comparando os escores dos testes SAT e ACT

Suponha que, quando você se inscreveu numa universidade, você fez o teste SAT, tendo um escore de 550. Seu amigo fez o teste ACT*, tendo um escore de 30. Qual o melhor escore?

Não podemos comparar os escores 550 e 30 dos dois testes diretamente, porque têm escalas diferentes. Nós os convertemos a escores-z, analisando quantos desvios padrão cada um está da média. Se o SAT tem $\mu = 500$ e $\sigma = 100$, um escore do SAT de $y = 550$ se converte em um escore-z de:

$$z = \frac{(y - \mu)}{\sigma} = \frac{(550 - 500)}{100} = 0{,}50.$$

O ACT tem aproximadamente $\mu = 18$ e $\sigma = 6$, assim o ACT = 30 se converte em um escore-z de $(30 - 18)/6 = 2{,}0$.

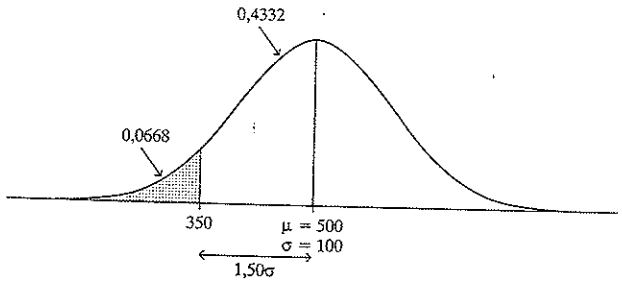
* N. de T.T.: O Teste para Universidades Norte-americanas (American College Testing – ACT) é um exame padronizado semelhante ao SAT (Scholastic Aptitude Test).

O escore do ACT de 30 é relativamente maior do que o escore do SAT de 550 porque 30 está 2,0 desvios padrão acima da média enquanto 550 está somente 0,5 desvios padrão acima da sua média. Os escores tanto do SAT quanto do ACT têm distribuições aproximadamente normais. Da Tabela A, $z = 2{,}0$ tem a probabilidade da cauda direita de 0,0228 e $z = 0{,}5$ tem uma probabilidade da cauda direita de 0,3085. De todos os estudantes que fizeram o ACT, apenas aproximadamente 2% tiveram um escore mais alto do que 30, enquanto que, de todos os estudantes que fizeram o SAT, aproximadamente 31% tiveram um escore maior do que 550. Em um sentido relativo, o escore do ACT é mais alto. ■

Aqui está um resumo de como usamos os escores-z:

☑ Usando os escores-z para encontrar probabilidades ou os valores-y.

- Se tivermos um valor y e precisarmos encontrar uma probabilidade, converta y em um escore-z usando $z = (y - \mu)/\sigma$ e use a tabela das probabilidades normais para obter a probabilidade de interesse.
- Se tivermos uma probabilidade e precisarmos encontrar um valor de y correspondente, converta a probabilidade a um valor da cauda direita e encontre o escore-z usando a tabela da normal padrão e, então, determine $y = \mu + z\sigma$.



☑ **Figura 4.7** Distribuição normal para os escores do SAT.

Para ilustrar, o Exemplo 4.5 usa a equação $z = (y - \mu)/\sigma$ para determinar quantos desvios padrão um teste do SAT está da média. O Exemplo 4.3 usou a equação $y = \mu + z\sigma$ para encontrar um percentil para uma distribuição normal dos escores de QI.

A distribuição normal padrão

Muitos métodos de estatística inferencial usam uma distribuição normal particular chamada de **distribuição normal padrão**.

☑ A distribuição normal padrão

A **distribuição normal padrão** é uma distribuição normal com média $\mu = 0$ e desvio padrão $\sigma = 1$.

Para a distribuição normal padrão, o número estando z desvios padrão acima da média é $\mu + z\sigma = 0 + z(1) = z$. Ele é simplesmente o próprio escore-z. Por exemplo, o valor de 2 está dois desvios padrão acima da média e o valor de $-1{,}3$ está 1,3 desvios padrão abaixo da média. Os valores originais são os mesmos dos escores-z. Veja a Figura 4.8.

Quando os valores para uma distribuição normal arbitrária são convertidos em escores-z, estes estão centrados em volta de 0 e têm um desvio padrão de 1. Os escores-z têm uma distribuição normal padrão.

☑ Os escores-z e a distribuição normal padrão

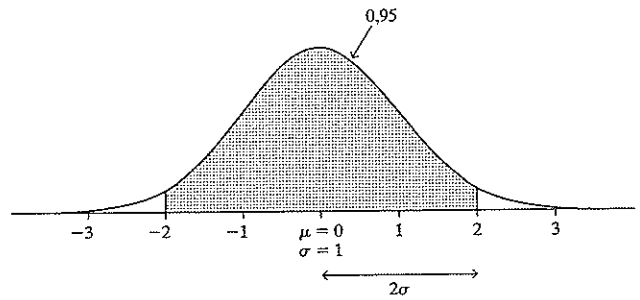
Se uma variável tem uma distribuição normal e se seus valores são convertidos em escores-z subtraindo a média e dividindo pelo desvio padrão, então os escores-z têm a distribuição normal padrão.

Suponha que convertemos cada escore-y do SAT usando $z = (y - 500)/100$. Por exemplo, $y = 650$ se converte em $z = 1{,}50$ e $y = 350$ se converte a $z = -1{,}50$. Então, todo o conjunto de escores-z tem uma distribuição normal com uma média de 0 e um desvio padrão de 1. Isso é a distribuição normal padrão.

Muitos métodos inferenciais convertem os valores das estatísticas em escores-z e, então, em probabilidades na curva normal. Usaremos os escores-z e probabilidades obtidas da normal, com frequência, no decorrer do livro.

4.4 AS DISTRIBUIÇÕES AMOSTRAIS DESCREVEM COMO AS ESTATÍSTICAS VARIAM

Vimos que as distribuições de probabilidade fornecem as probabilidades dos possíveis resultados de uma variável. Até agora, este capítulo tratou essas distribui-



☑ **Figura 4.8** A distribuição normal padrão tem média 0 e desvio padrão 1. Seus escores ordinários são os mesmos que os seus **escores-z**.

ções como conhecidas. Na prática, elas raramente são conhecidas. Usamos dados amostrais para fazer inferências sobre os parâmetros destas distribuições. Entretanto, as distribuições de probabilidades com valores dos parâmetros fixos são úteis para muitos desses métodos inferenciais. Vamos olhar um exemplo que ilustra a conexão entre inferência estatística e cálculos de probabilidade com valores conhecidos dos parâmetros.

EXEMPLO 4.6 Prevendo o resultado de uma eleição a partir de uma pesquisa de boca de urna

As redes de televisão amostram eleitores no dia da eleição para ajudá-las a prever com antecedência os candidatos vencedores. Na eleição de 2006 para o governo da Califórnia, a CNN[2] relatou os resultados de uma pesquisa de boca de urna com 2705 eleitores. A CNN declarou que 56,5% declararam que votaram no candidato Republicano, Arnold Schwarzenegger. Neste exemplo, a distribuição de probabilidade para o voto de uma pessoa iria informar a probabilidade de que um eleitor, aleatoriamente selecionado, votou em Schwarzenegger. Isso é igual à proporção da população de eleitores que votaram nele. Quando a pesquisa de boca de urna foi feita, tratava-se de um parâmetro desconhecido da população.

Para julgar se isso é informação suficiente para prever o resultado da eleição, a rede de televisão pode perguntar: "Suponha que metade da população tenha votado em Schwarzenegger. Seria, então, surpreendente que 56,5% dos indivíduos amostrados votaram nele?". Se isso seria muito improvável, a rede de televisão infere que Schwarzenegger recebeu mais do que metade dos votos da população. A inferência sobre o resultado da eleição tem como base a determinação da probabilidade do resultado amostral sob a suposição de que o parâmetro da população, o percentual dos eleitores que preferem Schwarzenegger, é igual a 50%. ■

Aproximadamente 7 milhões de pessoas votaram nessa eleição. Uma pesquisa de boca de urna amostrou somente 2705 eleitores, e, mesmo assim, as redes de televisão usaram isso para prever que Schwarzenegger iria vencer. Como poderia haver informação suficiente nessa pesquisa para fazer a previsão? Veremos a seguir a justificativa para isso.

Simulando o processo de estimativa

Uma **simulação** pode nos dizer quão bem os resultados de uma pesquisa de boca de urna se aproximam da proporção da população que vota em um candidato. Simulamos o voto de um eleitor retirado aleatoriamente da população pela seleção de um número de dois dígitos de uma tabela de números aleatórios (como a Tabela 2.2) ou com o uso de um *software*. Suponha que exatamente 50% da população tenha votado em Schwarzenegger e 50% tenha votado no candidato Democrata, Phil Angelides. Identifique todos os 50 números de dois dígitos entre 00 e 49 como votos para os Republicanos e todos os 50 números de dois dígitos entre 50 e 99 como votos para os Democratas. Então, cada candidato tem 50% de chance de ser selecionado em cada retirada de um valor aleatório de dois dígitos.

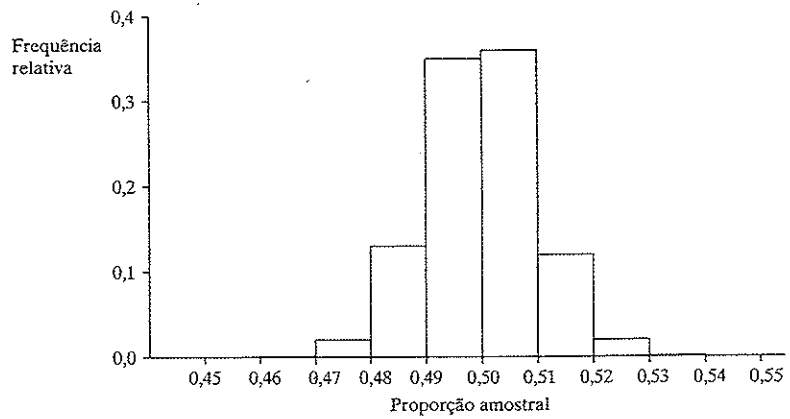
Por exemplo, os dois primeiros dígitos da primeira coluna da Tabela 2.2 fornecem os números aleatórios 10, 22, 24, 42, 37, 77 e assim por diante. Assim, dos seis primeiros eleitores selecionados, cinco votaram no Republicano (isto é, os números entre 00 e 49). Selecionar 2705 números de dois dígitos simula o processo de observação dos votos de uma amostra aleatória de 2705 eleitores de uma população muito maior (que é, na verdade, tratada como de tamanho infinito).

Usando um computador, selecionamos 2705 números aleatórios de dois dígitos e obtemos 1334 votos Republicanos

e 1371 votos Democratas. (Você pode tentar isto por si utilizando *applets* disponíveis na internet. Veja o Exercício 4.41.) A proporção amostral dos votos Republicanos foi de 1334/2705 = 0,493, bem próximo da proporção da população de 0,50. Essa estimativa em particular foi boa. Foi meramente sorte? Repetimos o processo e selecionamos mais 2705 números aleatórios de dois dígitos. Desta vez a proporção amostral dos votos dos Republicanos foi de 0,511, também muito boa. A seguir, programamos o computador para executar esse processo selecionando 2705 pessoas um milhão de vezes para que pudéssemos procurar por um padrão nos resultados. A Figura 4.9 mostra um histograma do milhão dos valores da proporção amostral obtidos. Aproximadamente todas as proporções simuladas estão entre 0,47 e 0,53; isto é, dentro de 0,03 da proporção da população de 0,50. Aparentemente, um tamanho da amostra de 2705 fornece uma boa estimativa da proporção da população.

Em resumo, se metade da população dos eleitores votou em Schwarzenegger, esperaríamos que entre aproximadamente 47% e 53% dos eleitores em uma pesquisa de boca de urna do tamanho de 2705 tenham votado nele. Assim teria sido muito incomum observar que 56,5% votaram nele, como aconteceu na boca de urna real. Se *menos da metade* da população votou no Schwarzenegger, teria sido ainda mais incomum observar isto. Isso é o básico da previsão da rede de televisão, da sua boca de urna, que Schwarzenegger venceu a eleição.

É possível executar essa simulação usando qualquer valor da proporção da população. Por exemplo, poderíamos simular a amostra quando a proporção da população que votou nos Republicanos é de 0,45 deixando os 45 números aleatórios entre 00 e 44 representar os votos dos Republicanos e os 55 entre 45 e 99 representar os votos dos Democratas. Da mesma forma, poderíamos mudar o tamanho de cada amostra aleatória da simulação para estudar o impacto do tamanho da amostra. Dos resultados da próxima seção, para uma amostra aleatória do tamanho de 2705, é muito provável a proporção amos-



☑ **Figura 4.9** Resultados da simulação da proporção amostral favorecendo o candidato Republicano para uma amostra aleatória de 2705 sujeitos de uma população na qual metade votou em cada candidato. Em aproximadamente todos os casos, a proporção amostral está dentro de 0,03 da proporção da população de 0,50.

tral estar dentro de 0,03 da proporção da população, apesar do seu valor.

Representando a variabilidade amostral por uma distribuição amostral

A preferência dos eleitores é uma variável, variando entre os eleitores. Da mesma forma, assim é a proporção amostral para uma variável atribuída a um determinado candidato: antes de ser obtida a amostra, seu valor é desconhecido, e aquele valor varia de amostra para amostra. Se várias amostras do tamanho de $n = 2705$ fossem selecionadas, certa quantia previsível de variação iria ocorrer nos valores da proporção amostral. Uma distribuição de probabilidade com aparência similar à Figura 4.9 descreve a variação que ocorre de selecionar repetidamente amostras de certo tamanho n e determinar uma estatística em particular. Essa distribuição é chamada de *distribuição amostral*. Ela também fornece as probabilidades dos possíveis valores da estatística para uma única amostra de tamanho n.

☑ **Distribuição amostral**

Uma **distribuição amostral** de uma estatística é a distribuição da probabilidade que especifica as probabilidades para valores possíveis que a estatística pode assumir.

Cada estatística da amostra tem uma distribuição amostral. Existe uma distribuição amostral da média amostral, uma distribuição amostral da proporção amostral, uma distribuição amostral da mediana amostral, e assim por diante. Uma distribuição amostral é meramente um tipo de distribuição de probabilidade. Ao contrário das distribuições estudadas até agora, uma distribuição amostral não especifica as probabilidades para observações individuais, mas para valores possíveis de uma estatística calculada das observações. Uma distribuição amostral nos permite calcular, por exemplo, as probabilidades sobre a proporção amostral de indivíduos que votaram nos Republicanos em uma pesquisa de boca de urna. Antes de os eleitores serem selecionados para a pesquisa de boca de urna, isso é uma variável. Ela tem uma distribuição amostral que fornece as probabilidades dos possíveis valores.

Uma distribuição amostral é importante na estatística inferencial porque nos ajuda a prever quão próximo uma estatística está de um parâmetro que ela estima. Da Figura 4.9, por exemplo, com uma amostra de tamanho 2705, a probabilidade é aparentemente alta de que a proporção amostral esteja dentro de 0,03 da proporção da população.

EXEMPLO 4.7 Construindo uma distribuição amostral

Às vezes, é possível construir uma distribuição amostral sem recorrer à simulação ou derivações matemáticas complexas. Para ilustrar, construímos a distribuição amostral da proporção da amostra para uma pesquisa de boca de urna de $n = 4$ eleitores de uma população na qual metade votou em cada candidato. Para cada eleitor, defina a variável y representando o voto, como segue:

$y = 1$, voto para o Republicano

$y = 0$, voto para o Democrata

Usamos uma sequência ordenada com quatro valores para representar os valores-y assumidos em uma amostra potencial de tamanho 4. Por exemplo, (1, 0, 0, 1) representa uma amostra que são o primeiro e quarto sujeitos que votaram no Republicano e o segundo e terceiro sujeitos que votaram no Democrata. As 16 amostras possíveis são:

(1, 1, 1, 1) (1, 1, 1, 0) (1, 1, 0, 1) (1, 0, 1, 1)
(0, 1, 1, 1) (1, 1, 0, 0) (1, 0, 1, 0) (1, 0, 0, 1)
(0, 1, 1, 0) (0, 1, 0, 1) (0, 0, 1, 1) (1, 0, 0, 0)
(0, 1, 0, 0) (0, 0, 1, 0) (0, 0, 0, 1) (0, 0, 0, 0)

Visto que metade da população votou em cada candidato, as 16 amostras são igualmente prováveis.

Agora, vamos construir a distribuição amostral da proporção da amostra que votou no candidato Republicano. Para um tamanho da amostra 4, essa proporção pode ser 0; 0,25; 0,50; 0,75 ou 1,0. A proporção 0 ocorre com somente uma das 16 amostras possíveis, (0, 0, 0, 0), assim sua probabilidade é igual a 1/16 = 0,0625. A proporção de 0,25 ocorre para quatro amostras, (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1,0) e (0, 0, 0, 1), assim sua probabilidade é igual a 4/16 = 0,25. Com base nesse raciocínio, a Tabela 4.3 mostra a probabilidade para cada um dos possíveis valores da proporção amostral.

A Figura 4.10 descreve a distribuição amostral da proporção para uma amostra de tamanho $n = 4$. Ela é muito mais dispersa do que a da Figura 4.9, para amostras de tamanho $n = 2705$, que está quase toda entre 0,47 e 0,53. Com uma amostra tão pequena (n =4), a proporção da amostra não precisa estar próxima da proporção da população. Isto não é surpreendente. Na prática, as amostras são sempre maiores do que $n = 4$. Usamos um valor pequeno nesse exemplo para ser mais simples escrever todas as amostras potenciais e encontrar as probabilidades para a distribuição amostral. ■

Com os dois resultados possíveis representados por 0 e 1, a Seção 3.2 observou que a proporção das vezes que 1 ocorre é a média amostral dos dados. Por exemplo, para a amostra (0, 1, 0, 0) na qual somente o segundo sujeito votou no Republicano, a média amostral é igual a (0 + 1 + 0 + 0)/4 = ¼ = 0,25, a proporção da amostra que votou no Republicano. Assim, a Figura 4.10 é também um exemplo de uma distribuição amostral da média amostral. A Seção 4.5 fornece resultados gerais sobre a distribuição amostral da média amostral.

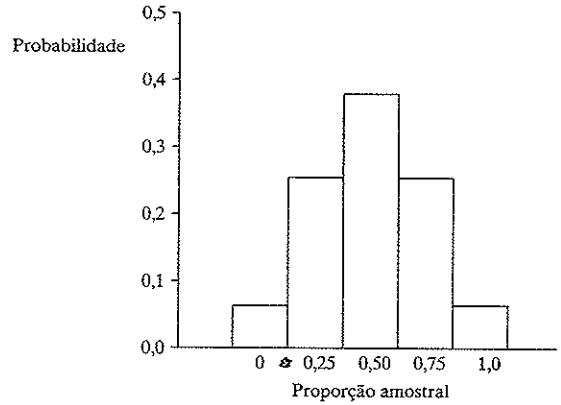
Interpretação da amostragem repetida das distribuições amostrais

As distribuições amostrais retratam a variabilidade amostral que ocorre na coleta dos dados e usa estatísticas amostrais para estimar os parâmetros. Se organizações de pesquisa de opinião pública diferentes, cada uma, fazem sua própria pesquisa de boca de urna e estimam a proporção da população que vota no candidato Republicano, elas obterão diferentes estimativas porque as amostras têm pessoas diferentes. Da mesma forma, a Figura 4.9 descreve a variabilidade nos valores da proporção amostral que ocorre selecionando um grande número de amostras do tamanho $n = 2705$ e na construção de um histograma das proporções amostrais. Ao contrário, a Figura 4.10 descreve a variabilidade para um grande número de amostras do tamanho $n = 4$.

Uma distribuição amostral de uma estatística baseada em n observações é a

☑ **Tabela 4.3** Distribuição amostral da proporção, para uma amostra de tamanho $n = 4$, quando a proporção populacional é 0,50. Por exemplo, uma proporção de 0,0 ocorre para somente 1 das 16 amostras possíveis, a saber, (0, 0, 0, 0), e assim sua probabilidade é 1/16 = 0,0625

Proporção da amostra	Probabilidade
0,00	0,0625
0,25	0,2500
0,50	0,3750
0,75	0,2500
1,00	0,0625



☑ **Figura 4.10** Distribuição amostral da proporção da amostra, para amostras aleatórias de tamanho $n = 4$ quando a proporção da população é 0,50.

distribuição de frequência relativa para aquela estatística resultante de coletar repetidamente amostras do tamanho n e a cada vez calcular o valor da estatística. É possível formar tal distribuição empiricamente, como na Figura 4.9, por repetidas amostragens ou por meio de simulação. Na prática, isto não é necessário. A forma de amostrar distribuições é geralmente conhecida teoricamente, como foi mostrado no exemplo anterior e na próxima seção. Podemos encontrar probabilidades para os valores de uma estatística amostral para uma amostra de um tamanho n dado.

4.5 DISTRIBUIÇÕES AMOSTRAIS DE MÉDIAS AMOSTRAIS

Pelo fato de que a média amostral $\bar{y}$ é muito usada, sua distribuição amostral merece atenção especial. Na prática, quando analisamos dados e encontramos $\bar{y}$, não sabemos quão perto ela está da média da população μ, porque não conhecemos o valor de μ. Usando a informação sobre a dispersão da distribuição amostral, podemos prever quão próximo ela está. Por exemplo, a distribuição amostral pode nos dizer com alta probabilidade se $\bar{y}$ está dentro de 10 unidades distante de μ.

Nesta seção iremos ver dois resultados principais sobre a distribuição amostral da média amostral. Um fornece fórmulas para o centro e a dispersão da distribuição amostral, enquanto o outro descreve a sua forma.

Média e erro padrão da distribuição amostral de $\bar{y}$

A média amostral $\bar{y}$ é uma variável porque seu valor varia de amostra para amostra. Para amostras aleatórias, ela flutua em torno da média da população μ, algumas vezes sendo menor e algumas vezes sendo maior. Na verdade, a média da distribuição amostral de $\bar{y}$ é igual a μ. Se, repetidamente, coletarmos amostras, então, em um grande número de repetições, a média das médias amostrais será igual à média da população μ.

A dispersão da distribuição amostral de $\bar{y}$ é descrita por seu desvio padrão, que é chamado de *erro padrão* de $\bar{y}$.

☑ Erro padrão

O desvio padrão da distribuição amostral de $\bar{y}$ é chamado de **erro padrão** de $\bar{y}$. O erro padrão de $\bar{y}$ é representado por $\sigma_{\bar{y}}$.

O erro padrão descreve como $\bar{y}$ varia de amostra para amostra. Suponha que, repetidamente, selecionamos amostras do tamanho n da população, encontrando $\bar{y}$ para cada conjunto de n observações. Então, em um grande número de repetições, o desvio padrão dos valores-$\bar{y}$ será igual ao erro padrão. O símbolo $\sigma_{\bar{y}}$ (em vez de σ) e a terminologia *erro padrão* (em vez de *desvio padrão*) distinguem essa medida do desvio padrão σ da população.

Na prática, não precisamos coletar amostras repetidamente para encontrar o erro padrão de $\bar{y}$ porque uma fórmula está disponível. Para uma amostra aleatória do tamanho n, o erro padrão de $\bar{y}$ depende de n e do desvio padrão da população σ por:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}.$$

A Figura 4.11 mostra a distribuição de uma população com $\sigma = 10$ e mostra, também, a distribuição amostral de $\bar{y}$ para $n = 100$, para a qual o erro padrão é $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 10/\sqrt{100} = 1{,}0$. A distribuição amostral tem somente um décimo da dispersão da distribuição da população. Isso significa que observações individuais tendem a variar muito mais do que as médias amostrais de amostra para amostra.

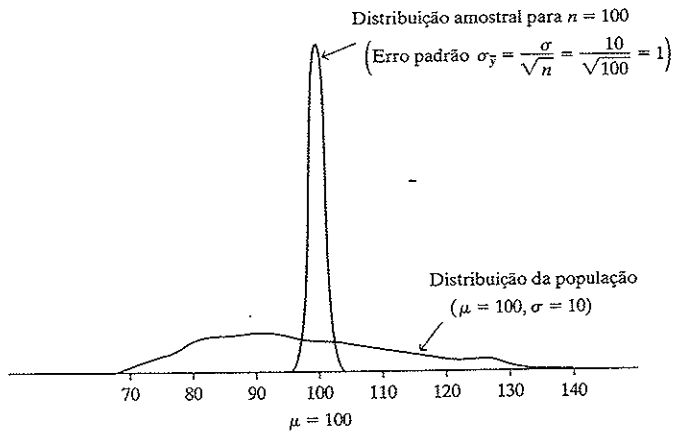
Em resumo, o resultado seguinte descreve o centro e a dispersão da distribuição amostral de $\bar{y}$:

☑ Média e erro padrão de $\bar{y}$

Considere uma amostra aleatória do tamanho n de uma população com média μ e desvio padrão σ. A distribuição amostral de $\bar{y}$, que fornece as probabilidades para os possíveis valores de $\bar{y}$, tem média μ e erro padrão $\sigma_{\bar{y}} = \sigma/\sqrt{n}$.

EXEMPLO 4.8 Erro padrão e proporção amostral na pesquisa de boca de urna de uma eleição

Seguindo o Exemplo 4.7, realizamos uma simulação para determinar quanta variabilidade esperar de amostra para amostra em uma pesquisa de boca de urna com 2705 eleitores. Em vez de conduzir uma si-



☑ **Figura 4.11** A distribuição da população e a distribuição amostral de $\bar{y}$ para $n = 100$.

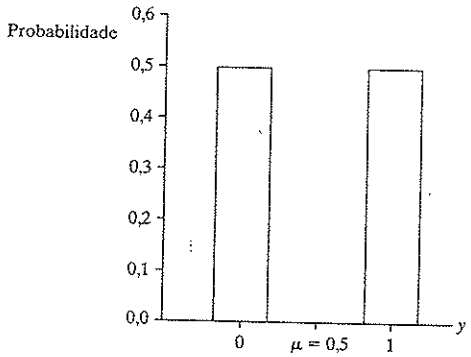
mulação, podemos obter uma informação similar encontrando diretamente o erro padrão. Conhecer o erro padrão nos ajuda a responder à seguinte pergunta: se metade da população votou em cada candidato, quanto uma proporção amostral de uma pesquisa de boca de urna com 2705 eleitores irá variar de amostra para amostra?

Como no Exemplo 4.8, considere a variável y igual a 1 para um voto no Republicano e 0 para um voto no Democrata. A Figura 4.12 mostra a distribuição para a qual metade da população votou no Republicano, assim que $P(1) = 0{,}50$ e $P(0) = 0{,}50$. A média da distribuição é igual a 0,50, que é a proporção da população que votou no Republicano. (Ou, da fórmula, $\mu = \Sigma y P(y) = 0(0{,}50) + 1(0{,}50) = 0{,}50$.) O desvio entre y e a média, ao quadrado, $(y - \mu)^2$, é igual a $(0 - 0{,}50)^2 = 0{,}25$ quando $y = 0$ e ele é igual a $(1 - 0{,}50)^2 = 0{,}25$ quando $y = 1$. A variância é o valor esperado desses quadrados dos desvios. Portanto, ela é igual a $\sigma^2 = 0{,}25$. Assim, o desvio padrão da distribuição da população de y é igual a $\sigma = \sqrt{0{,}25} = 0{,}50$.

Para uma amostra, a média dos valores 0 e 1 é a proporção da amostra dos eleitores que votaram no Republicano. Sua distribuição amostral tem uma média que é a média da distribuição da população de y, a saber, $\mu = 0{,}50$. Para amostras repetidas de um tamanho fixo de n, a proporção amostral flutua em torno de 0,50, sendo maior aproximadamente a metade das vezes e menor outra metade das vezes. O desvio padrão da distribuição amostral é o erro padrão. Para um tamanho da amostra de 2705, ele vale:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0{,}50}{\sqrt{2705}} = 0{,}01.$$

Um resultado, que veremos mais tarde nesta seção, diz que essa distribuição amostral tem forma de sino. Assim, com a probabilidade próxima a 1,0, a proporção amostral estará entre três erros padrão a partir de μ, isto é, dentro de $3(0{,}01) = 0{,}03$ de 0,50, ou entre 0,47 e 0,53. Para uma amostra aleatória de tamanho 2705 retirada de uma população na qual 50% votou no Republicano, seria extremamente surpreendente se fosse encontrado que menos do que 47% ou mais do que 53% votaram no Republicano. Vimos, agora, como obter esse resultado usando simulação, como mostra a Figura 4.9, ou usando a informação sobre a média e o erro padrão da distribuição amostral. ■



☑ **Figura 4.12** Distribuição da população quando $y = 0$ ou 1, com probabilidade 0,50 cada. Esta é a distribuição para um voto, com 1 = voto para o candidato Republicano e 0 = voto para o candidato Democrata.

Efeito do tamanho da amostra na distribuição amostral e precisão das estimativas

O erro padrão fica menor à medida que o tamanho da amostra n fica maior. A razão para isto é que o denominador ($\sqrt{n}$), da fórmula do erro padrão $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, aumenta à medida que n aumenta. Por exemplo, quando o desvio padrão da população é $\sigma = 0{,}50$, recém vimos que o erro padrão é 0,01 quando $n = 2705$. Quando $n = 100$, um tamanho menos típico para uma pesquisa, o erro padrão é igual a:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0{,}50}{\sqrt{n}} = \frac{0{,}50}{\sqrt{100}} = 0{,}05.$$

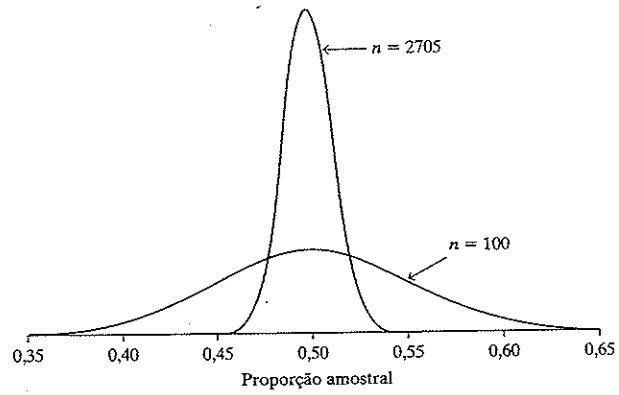
Com $n = 100$, visto que três erros padrão são iguais a $3(0{,}05) = 0{,}15$, a probabilidade é muito alta de que a proporção amostral esteja dentro de $0{,}50 \pm 0{,}15$ ou entre 0,35 e 0,65.

A Figura 4.13 mostra as distribuições amostrais da proporção da amostra quando $n = 100$ e quando $n = 2705$. À medida que n aumenta, o erro padrão diminui e a distribuição amostral fica mais limitada. Isso significa que a proporção da amostra tende a estar próxima da proporção da população. É mais provável que a proporção da amostra aproxime uma proporção desconhecida da população quando $n = 2705$ do que quando $n = 100$. Isso está de acordo com a nossa intuição de que amostras maiores fornecem estimativas mais precisas das características da população.

Em resumo, o erro resulta da estimativa de μ por $\bar{y}$ porque amostramos somente parte da população. Esse erro, que é o **erro amostral**, tende a diminuir à medida que o tamanho da amostra n aumenta. O erro padrão é fundamental para procedimentos de inferência que preveem o erro amostral usando $\bar{y}$ para estimar μ.

A distribuição amostral da média amostral é aproximadamente normal

Para a distribuição da população para o voto em uma eleição, exibido na Figura 4.12, o resultado tem somente dois valores possíveis, pois é altamente discreto. No entanto, as duas distribuições amostrais exibidas na Figura 4.13 têm forma de sino. É a consequência do segundo resultado principal desta seção, que descreve a *forma* da distribuição amostral de $\bar{y}$. Esse resultado pode ser provado matematicamente e é geralmente chamado de *Teorema Central do Limite*.



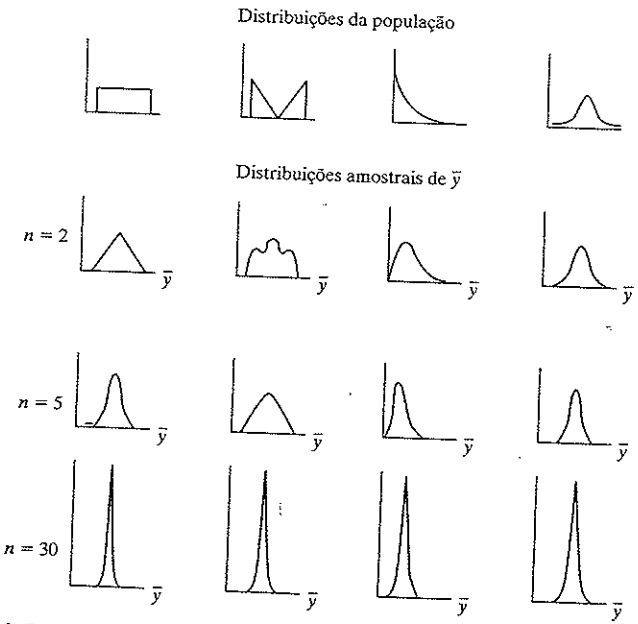
☑ **Figura 4.13** As distribuições amostrais da proporção amostral, quando n = 100 e quando n = 2705. Ambas se referem à amostragem realizada sobre a população na Figura 4.12.

☑ Teorema Central do Limite

Para amostragem aleatória com um tamanho da amostra n grande, a distribuição amostral da média da amostra, $\bar{y}$, é aproximadamente uma distribuição normal.

Aqui estão algumas implicações e interpretações desse resultado:

- A normalidade aproximada da distribuição amostral ocorre *não importa* a forma da distribuição da população. Isto é extraordinário. Para amostras aleatórias grandes, a distribuição amostral de $\bar{y}$ é aproximadamente normal mesmo se a distribuição da população for altamente assimétrica, com a forma de U, ou altamente discreta como a distribuição binária na Figura 4.12. Veremos que isso permite fazer inferências mesmo quando a distribuição da população é altamente irregular. Isto é útil porque muitas variáveis das ciências sociais são assimétricas ou altamente discretas.
  A Figura 4.14 exibe as distribuições amostrais de $\bar{y}$ para quatro formas diferentes de distribuições populacionais mostradas no topo da figura. Na parte de baixo delas estão retratadas as distribuições amostrais para amostras aleatórias de tamanhos $n = 2$,5 e 30. À medida que n aumenta, a distribuição amostral tem mais a forma de sino.
- O tamanho que n deve ter antes que a distribuição amostral tenha a forma de sino depende amplamente da assimetria da distribuição da população. Se a distribuição da *população* tem a forma de sino, então, a distribuição amostral terá a forma de sino para *todos* os tamanhos de amostra. O painel mais à direita na Figura 4.14 ilustra isso. As distribuições mais assimétricas requerem tamanhos de amostra maiores. Para a maioria dos casos, um tamanho da amostra de aproximadamente 30 é suficiente (embora não possa ser grande o suficiente para uma inferência precisa). Assim, na prática, para a amostragem aleatória, a distribuição amostral tem quase sempre aproximadamente a forma de sino.



☑ **Figura 4.14** Quatro distribuições populacionais diferentes e as correspondentes distribuições amostrais de $\bar{y}$. À medida que n aumenta, as distribuições amostrais ficam mais estreitas e mais se aproximam da forma de sino.

- Poderíamos verificar o Teorema Central do Limite empiricamente selecionando repetidamente amostras aleatórias, calculando $\bar{y}$ para cada amostra de n observações. Então, o histograma dos valores-$\bar{y}$ seria aproximadamente uma curva normal em torno de μ com o erro padrão igual a $\sigma/\sqrt{n}$, o desvio padrão da população dividido pela raiz quadrada do tamanho da amostra utilizada.
- Sabendo que a distribuição amostral de $\bar{y}$ é aproximadamente normal nos ajuda a encontrar as probabilidades para os possíveis valores de $\bar{y}$. Por exemplo, $\bar{y}$ quase certamente está dentro de $3\sigma_{\bar{y}} = 3\sigma/\sqrt{n}$ de μ. Veremos que um raciocínio dessa natureza é vital para os métodos de estatística inferencial.

EXEMPLO 4.9 A média amostral da renda de trabalhadores migrantes está próxima à média da população?

Para a população de trabalhadores migrantes na Califórnia, suponha que a renda semanal tem uma distribuição que é assimétrica à esquerda com uma média de $\mu = \$380$ e um desvio padrão de $\sigma = \$80$. Um pesquisador, desconhecendo esses valores, planeja amostrar aleatoriamente 100 trabalhadores migrantes e usar a média amostral da renda $\bar{y}$ para estimar μ. Qual é a distribuição amostral da média amostral? Qual é a probabilidade de que $\bar{y}$ esteja acima de \$400?

Pelo Teorema Central do Limite, a distribuição amostral da média amostral $\bar{y}$ é aproximadamente normal, embora a distribuição da população seja assimétrica. A distribuição amostral tem a mesma média da distribuição da população, a saber, $\mu =$ \$380. Seu erro padrão é

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8{,}0 \text{ dólares.}$$

Assim, é altamente improvável que $\bar{y}$ esteja entre \$24 (três erros padrão) de μ.

Para a distribuição amostral normal com média 380 e erro padrão de 8, um valor de $\bar{y} = 400$ tem um escore-z de:

$$z = (400 - 380)/8 = 2{,}5.$$

De uma tabela da distribuição normal padrão (como a Tabela A), a probabilidade da cauda direita acima de 400 é 0,0062. É muito improvável que seja encontrada uma média amostral acima de \$400. ■

Este último exemplo não é realista, porque usou o valor da média da população μ. Na prática, esse valor seria desconhecido. Entretanto, a distribuição amostral de $\bar{y}$ fornece a probabilidade de que a média amostral esteja dentro de certa distância da média da população μ, mesmo quando μ é desconhecido. Ilustramos para o estudo as rendas dos trabalhadores migrantes da Califórnia. Vamos calcular a probabilidade de que a média amostral da renda mensal $\bar{y}$ esteja dentro de \$10 da renda média verdadeira μ para todos esses trabalhadores.

Agora, a distribuição amostral de $\bar{y}$ é aproximadamente normal na forma e está centrada aproximadamente em μ. Vimos

no exemplo anterior que quando $n = 100$, o erro padrão é $\sigma_{\bar{y}} = \$8{,}0$. Portanto, a probabilidade de que $\bar{y}$ esteja dentro de \$10 de μ é a probabilidade de que uma variável normalmente distribuída esteja dentro de $10/8 = 1{,}25$ desvios padrão da sua média. Isto é, o número de erros padrão que $\mu + 10$ (ou $\mu - 10$) está afastado de μ é

$$z = \frac{(\mu + 10) - \mu}{8} = \frac{10}{8} = 1{,}25,$$

como mostra a Figura 4.15. De uma tabela da normal padrão, a probabilidade de que $\bar{y}$ esteja *mais do que* 1,25 erros padrão de μ (em qualquer direção) é $2(0{,}1056) = 0{,}21$. Portanto, a probabilidade de que $\bar{y}$ esteja a não mais do que \$10 de μ é igual a $1 - 0{,}21 = 0{,}79$.

Esse exemplo ainda não é realista porque usou o desvio padrão da população σ. Na prática, nós estimaríamos este valor. O próximo capítulo mostra que, para conduzir uma inferência, podemos estimar σ pelo desvio padrão da amostra s.

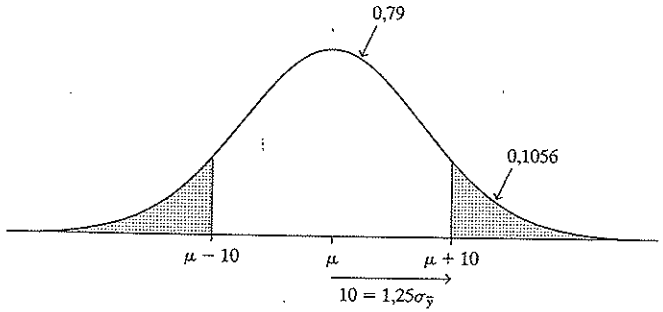
Para conhecer o Teorema Central do Limite e como a distribuição amostral tem mais e mais a forma de sino à medida que n aumenta, pode ser muito útil usar um *applet* da internet. Recomendamos fortemente que você tente fazer os Exercícios 4.41 e 4.42.

4.6 REVISÃO: POPULAÇÃO, DADOS AMOSTRAIS E DISTRIBUIÇÕES AMOSTRAIS

As distribuições amostrais são fundamentais para a inferência estatística e para a metodologia apresentada no restante deste livro. Por causa disso, revisaremos e detalharemos a distinção entre a distribuição amostral e os dois tipos de distribuições apresentados na Seção 3.1 – a distribuição da **população** e a distribuição dos **dados amostrais**.

Aqui está uma descrição condensada dos três tipos de distribuição:

- **Distribuição da população:** é a distribuição da qual selecionamos a amostra. Geralmente é desconhecida. Fazemos inferências sobre características, como os parâmetros μ e σ, que descrevem seu centro e dispersão. Representamos o tamanho da população por N.
- **Distribuição dos dados amostrais:** é a distribuição dos dados que realmente observamos; isto é, as observações da amostra $y_1, y_2, \ldots, y_n$. Podemos descrevê-la por estatísticas tais como a média amostral $\bar{y}$ e o desvio padrão da amostra s. Quanto maior o tamanho da amostra n, mais a distribuição dos dados amostrais se assemelha à distribuição da população e mais perto a estatística amostral, como $\bar{y}$, estará dos parâmetros da população, como μ.
- **Distribuição amostral** de uma estatística: é a distribuição da probabilidade para os possíveis valores de uma estatística amostral, tal como $\bar{y}$. Uma distribuição amostral descreve a variabilidade que ocorre com a estatística entre amostras de determinado tamanho. Essa distribuição determina a probabilidade de que a estatística não se afaste mais do que um determinado valor do parâmetro da população que ela está estimando.



☑ **Figura 4.15** Distribuição amostral de $\bar{y}$ para μ desconhecido e erro padrão $\sigma_{\bar{y}} = 8$.

EXEMPLO 4.10 Três distribuições para um item da Pesquisa Social Geral

Em 2006, a PSG perguntou sobre o número de horas por semana que se passa navegando na rede (*web*), excluindo o *e-mail* (variável representada por "WWWHR"). A *distribuição dos dados amostrais* para $n = 2778$ sujeitos da amostra era bem assimétrica à direita. Ela é descrita pela média amostral $\bar{y} = 5{,}7$ e o desvio padrão da amostra $s = 10{,}5$.

Pelo fato de que a PSG não pode amostrar toda a população de norte-americanos adultos (N de aproximadamente 200 milhões), não conhecemos a *distribuição da população*. Pelo fato de a distribuição dos dados amostrais ter uma amostra tamanho grande, provavelmente a distribuição da população se parece com ela. Muito provavelmente, a distribuição da população seria altamente assimétrica à direita. Sua média e seu desvio padrão seriam similares aos valores da amostra. Valores como $\mu = 6{,}0$ e $\sigma = 10{,}3$ seriam realistas.

Se a PSG, repetidamente, coletasse amostras aleatórias de 2778 norte-americanos adultos, o tempo médio da amostra $\bar{y}$ que se passa na rede iria variar de pesquisa para pesquisa. A *distribuição amostral* descreve como $\bar{y}$ variaria. Por exemplo, se a população tem uma média $\mu = 6{,}0$ e um desvio padrão de $\sigma = 10{,}3$, então a distribuição amostral de $\bar{y}$ também teria uma média de 6,0 e ela teria um desvio padrão (erro padrão) de:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{10{,}3}{\sqrt{2778}} = 0{,}20.$$

Diferente da população e das distribuições amostrais de dados, a distribuição amostral teria a forma de sino e seria estreita. Aproximadamente toda a distribuição estaria dentro de $3(0{,}20) = 0{,}6$ da média da população de 6,0. Assim, seria bem provável que toda a amostra de tamanho 2778 tivesse uma média amostral dentro de $6{,}0 \pm 0{,}6$. Em resumo, os dados amostrais e a distribuição da população são altamente assimétricos e dispersos, enquanto a distribuição amostral de $\bar{y}$ é aproximadamente normal e tem aproximadamente todos os valores concentrados em um intervalo estreito. ■

Na realidade, a PSG usa uma amostra por conglomerados multiestágio, em vez de uma amostra aleatória simples. Por causa disso, o erro padrão verdadeiro é, na verdade, um pouco maior do que o obtido por essa fórmula. (Isto é discutido no Apêndice A do dicionário de dados em http://sda.berkeley.edu/GSS.) Está além do alcance deste livro ajustar os erros padrão para efeitos de agrupamento. Com a finalidade de ilustração, trataremos os dados da PSG como se eles viessem de uma amostra aleatória simples, lembrando que, na prática, alguns ajustes podem ser necessários como está explicado no *site* da PSG.

EXEMPLO 4.11 Três distribuições para o exemplo de pesquisa de boca de urna

Consideramos, mais uma vez, a variável y = voto na eleição governamental da Califórnia em 2006 para um eleitor aleatoriamente selecionado. Seja $y = 1$ para o candidato Republicano e $y = 0$ para outro candidato. Na verdade, dos 6921442 adul-

tos residentes da Califórnia que votaram, 55,9% votaram em Schwarzenegger. Assim, a distribuição da probabilidade para y tem uma probabilidade de 0,559 em $y = 1$ e uma probabilidade de 0,441 em $y = 0$. A média dessa distribuição é $\mu = 0{,}559$, que é a proporção da população que votou em Schwarzenegger. De uma fórmula que estudaremos no próximo capítulo, o desvio padrão dessa distribuição é igual a $\sigma = 0{,}497$.

A distribuição populacional para a preferência eleitoral consiste em $N = 6921442$ valores de y, 44,1% dos quais são 0 e 55,9% são 1. Essa distribuição é descrita pelo parâmetro $\mu = 0{,}559$ e $\sigma = 0{,}497$. A Figura 4.16 retrata esta distribuição, que é altamente discreta (binária) e não tem a forma de sino.

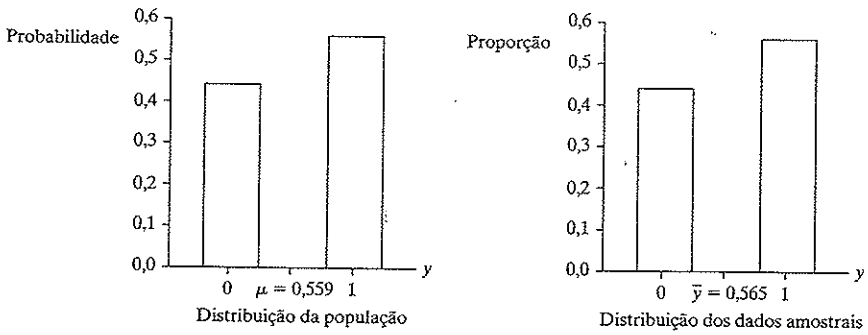
Antes de todos os votos terem sidos contados, a distribuição da população era desconhecida. Quando a votação foi encerrada, a CNN relatou os resultados de uma pesquisa de boca de urna de tamanho $n = 2705$ para prever o resultado. Um histograma dos 2705 votos da amostra descreve a distribuição dos dados amostrais. Dos 2705 eleitores, 56,5% disseram que votaram em Schwarzenegger (isto é, $y = 1$) e 43,5% disseram que votaram em outro candidato ($y = 0$). A Figura 4.16 também exibe o histograma dos valores dos dados amostrais. Como a distribuição da população, a distribuição dos dados amostrais se concentra em $y = 0$ e $y = 1$. Ela é descrita pela estatística amostral como, por exemplo, $\bar{y} = 0{,}565$, que é a proporção amostral que votou em Schwarzenegger. Quanto maior o tamanho da amostra, mais esta distribuição dos dados amostrais tende a assemelhar-se à distribuição da população, visto que as observações da amostra são um subconjunto dos valores da população. Se toda a população é amostrada, quando todos os votos são contados, então as duas distribuições são idênticas.

Para uma amostra aleatória de tamanho $n = 2705$, a distribuição amostral de $\bar{y}$ é aproximadamente normal. Sua média é $\mu = 0{,}559$ e seu erro padrão é:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0{,}497}{\sqrt{2705}} = 0{,}01.$$

A Figura 4.17 apresenta essa distribuição amostral, relativa à distribuição da população de votos.

Em contraposição, a distribuição da população e a distribuição dos dados amostrais dos votos estão concentradas nos valores 0 e 1. A distribuição amostral é completamente diferente delas, sendo muito menos dispersa e com forma de sino. A população e as distribuições dos dados amostrais do voto não têm a forma de sino. Elas são altamente discretas, concentradas em 0 e 1. Com $n = 2705$, a proporção da amostra pode assumir um grande número de valores entre 0 e 1 e sua distribuição amostral é aparentemente contínua, sendo aproximadamente normal pelo Teorema Central do Limite. ■

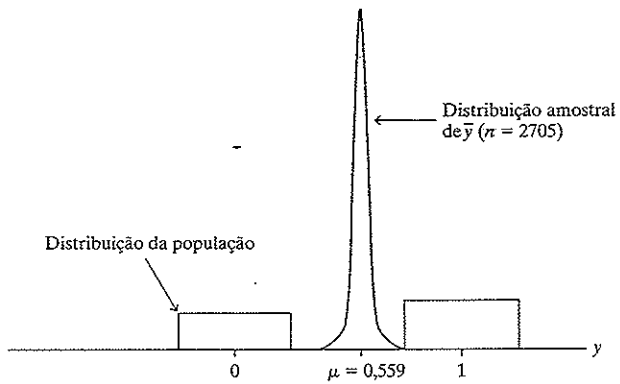


☑ **Figura 4.16** As distribuições da população ($N = 6921442$) e dos dados amostrais ($n = 2705$) dos votos na eleição governamental da Califórnia em 2006, onde 1 = Schwarzenegger e 0 = outros candidatos.

Efeito do tamanho da amostra em dados amostrais e distribuições amostrais

Vimos que a distribuição amostral se aproxima da normal na forma para valores grandes de n. Amostrando apenas uma observação ($n = 1$), $\bar{y} = y_1$ e a distribuição amostral de $\bar{y}$ é a mesma da distribuição de probabilidade de uma observação de y. Isso é simplesmente a distribuição da população y, que não precisa nem mesmo se parecer com uma normal. À medida que n aumenta, a distribuição amostral de $\bar{y}$ assume cada vez mais a forma de sino. Para $n \geq 30$, a aproximação é geralmente boa o bastante. À medida que o tamanho da amostra n se aproxima do tamanho da população N, a distribuição amostral normal de $\bar{y}$ fica cada vez mais estreita convergindo a um valor único μ. Quando toda a população é amostrada, $\bar{y} = \mu$ com a probabilidade 1 (isto é, as duas medidas são as mesmas) e a distribuição amostral se concentra no ponto μ.

A Figura 4.17 mostrou uma grande diferença entre a distribuição da população e a distribuição amostral de $\bar{y}$ (para $n = 2705$). Observe, também, a grande diferença entre a distribuição dos dados amostrais (Figura 4.16) e a distribuição amostral (Figura 4.17). A distribuição dos dados amostrais se assemelha mais com a distribuição da população, principalmente quando o tamanho da amostra aumenta. A distribuição amostral, por outro lado, tem uma aparência em forma de sino e fica mais estreita à medida que n aumenta. Como mostra a Figura 4.16, os valores amostrais de y podem ser somente 0 ou 1. Por outro lado, as médias das amostras (que são proporções amostrais) podem assumir somente valores entre 0 e 1. De acordo com a distribuição amostral de $\bar{y}$ para $n = 2705$, é praticamente impossível que uma amostra aleatória desse tamanho tenha uma média



☑ **Figura 4.17** A distribuição da população (onde $y = 1$ é voto para Schwarzenegger e $y = 0$ é voto para outro candidato) e a distribuição amostral de $\bar{y}$ para $n = 2705$.

amostral em algum lugar próximo de 0 ou de 1; aproximadamente todos os valores estão entre 0,53 e 0,59 (dentro de aproximadamente três desvios padrão da média da distribuição amostral).

A função básica das distribuições amostrais na inferência estatística

Vimos, pelo Teorema Central do Limite, que, geralmente, podemos usar a distribuição normal para encontrar probabilidades sobre $\bar{y}$. Nos próximos dois capítulos, mostraremos que as inferências estatísticas se baseiam nesse teorema.

O resultado de que as médias amostrais têm distribuições amostrais aproximadamente normais é importante também para grandes amostras porque resultados similares valem para muitas outras estatísticas. Por exemplo, muitas estatísticas amostrais que são utilizadas para estimar parâmetros populacionais têm distribuições amostrais aproximadamente normais, para amostras aleatórias grandes. A razão principal para o papel básico da distribuição normal é que muitas estatísticas têm distribuições amostrais aproximadamente normais.

4.7 RESUMO DO CAPÍTULO

Para uma observação em uma amostra aleatória ou um experimento aleatório, a **probabilidade** de um resultado particular é a proporção das vezes em que este resultado iria ocorrer em uma longa sequência de observações.

- Uma **distribuição de probabilidade** especifica as probabilidades para os possíveis valores de uma variável. Consideramos $P(y)$ como representação da probabilidade do valor y. As probabilidades são não negativas e somam 1,0.
- As distribuições de probabilidade têm parâmetros resumo, como, por exemplo, a média μ e o desvio padrão σ. A média para a distribuição de probabilidade de uma variável discreta é:

$$\mu = \sum yP(y).$$

Esse resultado é também denominado de **valor esperado** de y.

- A **distribuição normal** tem um gráfico que é uma curva simétrica em forma de sino especificada pela média μ e pelo desvio padrão σ. Para todo z, a probabilidade que está entre z desvios padrão da média é a mesma para qualquer distribuição normal.
- O **escore-z** para uma observação y é igual a

$$z = (y - \mu)/\sigma.$$

Ele mensura o número de desvios padrão que y está da média μ. Para uma distribuição normal, os escores-z têm uma *distribuição normal padrão* que tem uma média = 0 e desvio padrão = 1.

- Uma **distribuição amostral** é uma distribuição de probabilidade de uma estatística amostral, como, por exemplo, a média amostral ou a proporção da amostra. Ela especifica probabilidades para todos os possíveis valores da estatística considerando todas as amostras possíveis.
- A distribuição amostral da média da amostra $\bar{y}$ está centrada na média da população μ. Seu desvio padrão, chamado de **erro padrão**, se relaciona ao desvio padrão σ da população por $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. À medida que o tamanho da amostra n aumenta, o erro padrão diminui, assim, a média da amostra se aproxima da média da população.
- O **Teorema Central do Limite** declara que, para amostras aleatórias grandes, a distribuição amostral da média da amostra é aproximadamente normal, não importando a forma da distribuição da população. O resultado se aplica também às proporções, visto que a proporção da amostra é um caso especial da média da amostra para observações codificadas como 0 e 1 (como para dois candidatos em uma eleição).

A forma de sino para a distribuição amostral de muitas estatísticas é a razão principal para a importância da distribuição normal. Os próximos dois capítulos mostram como o Teorema Central do Limite é a base dos métodos da inferência estatística.

EXERCÍCIOS

Praticando o básico

4.1 Em uma PSG, em resposta à pergunta: "Você acredita em vida após a morte?", 907 pessoas responderam *sim* e 220 responderam *não*. Baseado neste levantamento de dados, estime a probabilidade de que um adulto norte-americano, aleatoriamente selecionado, acredite na vida após a morte.

4.2 Uma PSG estima a probabilidade de que um adulto norte-americano acredite no paraíso em 0,85.

(a) Estime a probabilidade de que um adulto norte-americano não acredite no paraíso.

(b) Daqueles que acreditam no paraíso, aproximadamente 84% acreditam no inferno. Estime a probabilidade de que um norte-americano adulto, aleatoriamente selecionado, acredite em ambos, paraíso e inferno.

4.3 Em 2000, a PSG perguntou a sujeitos se eles eram membros de um grupo ambientalista (variável "GRNGROUP") e se eles estariam dispostos a pagar preços mais altos para proteger o meio ambiente (variável "GRNPRICE"). A Tabela 4.4 mostra os resultados.

(a) Explique porque 96/1117 = 0,086 estima a probabilidade de que um adulto norte-americano, aleatoriamente selecionado, seja membro de um grupo ambientalista.

(b) Mostre que a probabilidade estimada de estar disposto a pagar preços mais altos para proteger o meio ambiente é (i) 0,312, dado que a pessoa é um membro de um grupo ambientalista, (ii) 0,086, dado que ela não é um membro de um grupo ambientalista.

(c) Mostre que a probabilidade estimada de que uma pessoa é tanto membro de um grupo ambientalista *e* muito disposta a pagar preços muito mais altos para proteger o meio ambiente é de 0,027, (i) diretamente usando as frequências da tabela, (ii) usando as estimativas da probabilidade de (a) e (b).

(d) Mostre que a probabilidade estimada de que uma pessoa responda *sim* a ambas as perguntas ou não a ambas as questões é de 0,862.

☑ **Tabela 4.4**

		Paga preços altos		
		Sim	Não	Total
Membro de um grupo ambientalista	Sim	30	66	96
	Não	88	933	1021
Total		118	999	1117

4.4 Considere y = número de línguas nas quais a pessoa é fluente. De acordo com Statistics Canada, para residentes no Canadá, a distribuição da probabilidade é $P(0) = 0{,}02$, $P(1) = 0{,}81$, $P(2) = 0{,}17$, com uma probabilidade insignificante para valores altos de y.

(a) O y é uma variável discreta ou contínua? Por quê?

(b) Construa uma tabela mostrando a distribuição da probabilidade de y.

(c) Encontre a probabilidade de um canadense *não* ser poliglota.

(d) Encontre a média dessa distribuição da probabilidade.

4.5 Considere y a representação do número de pessoas conhecidas pessoalmente que foram vítimas de homicídio den-

tro dos últimos 12 meses. De acordo com resultados de uma Pesquisa Social Geral recente, para uma pessoa aleatoriamente escolhida nos Estados Unidos, a distribuição da probabilidade de y é aproximadamente $P(0) = 0{,}91$, $P(1) = 0{,}06$, $P(2) = 0{,}02$, $P(3) = 0{,}01$.

(a) Explique por que não é válido encontrar a média dessa distribuição da probabilidade como $(0 + 1 + 2 + 3)/4 = 1{,}5$.

(b) Encontre a média correta da distribuição da probabilidade.

4.6 Um bilhete de loteria estadual custa \$1,00. Com uma probabilidade de 0,0000001, você ganha um milhão de dólares (\$1000000) e com a probabilidade de 0,9999999 você não ganha nada. Considere y como sendo o prêmio pela compra de um bilhete. Construa a distribuição de probabilidade para y. Mostre que a média da distribuição é igual a 0,10, correspondendo a um retorno esperado de 10 centavos para o dólar pago.

4.7 Seja y o resultado da seleção de um único dígito de uma tabela de números aleatórios.

(a) Construa a distribuição da probabilidade para y. (Esse tipo de distribuição é chamado de distribuição *uniforme* por causa da dispersão uniforme das probabilidades por meio dos resultados possíveis.)

(b) Encontre a média dessa distribuição da probabilidade.

(c) O desvio padrão σ dessa distribuição é um dos seguintes valores: 0,4, 2,9, 7,0, 12,0. Qual você acha que é o correto? Por quê?

4.8 Para uma distribuição normal, encontre a probabilidade de que uma observação esteja:

(a) Pelo menos 1 desvio padrão acima da média.

(b) Pelo menos 1 desvio padrão abaixo da média.

(c) Pelo menos 0,67 desvios padrão acima da média.

4.9 Para uma variável distribuída normalmente, verifique que a probabilidade entre

(a) $\mu - \sigma$ e $\mu + \sigma$ é igual a 0,68.

(b) $\mu - 1{,}96\sigma$ e $\mu + 1{,}96\sigma$ é igual a 0,95.

(c) $\mu - 3\sigma$ e $\mu + \sigma$ é igual a 0,997.

(d) $\mu - 0{,}67\sigma$ e $\mu + 0{,}67\sigma$ é igual a 0,50.

4.10 Encontre o valor-z para o qual a probabilidade de que uma variável normal exceda $\mu + z\sigma$ seja igual a:

(a) 0,01

(b) 0,025

(c) 0,05

(d) 0,10

(e) 0,25

(f) 0,50

4.11 Encontre o valor-z tal que, para uma distribuição normal, o intervalo de $\mu - z\sigma$ a $\mu + z\sigma$ contenha:

(a) 50%

(b) 90%

(c) 95%

(d) 98%

(e) 99% dos valores da variável.

4.12 Encontre os valores-z correspondentes ao:

(a) $90^{\underline{o}}$

(b) $95^{\underline{o}}$

(c) $98^{\underline{o}}$ e

(d) $99^{\underline{o}}$ percentil de uma distribuição normal.

4.13 Mostre que se z é o número tal que o intervalo de $\mu - z\sigma$ a $\mu + z\sigma$ contém 90% dos valores de uma distribuição normal, então $\mu + z\sigma$ é igual ao $95^{\underline{o}}$ percentil.

4.14 Se z é o número positivo tal que o intervalo de $\mu - z\sigma$ a $\mu + z\sigma$ contém 50% dos valores de uma distribuição normal, então:

(a) Que percentil corresponde a (i) $\mu + z\sigma$? (ii) $\mu - z\sigma$?

(b) Encontre este valor de z.

(c) Usando esse resultado, explique por que o quartil superior e o inferior de uma distribuição normal são $\mu + 0{,}67\sigma$ e $\mu - 0{,}67\sigma$, respectivamente.

4.15 Que proporção de valores de uma distribuição normal está nos seguintes intervalos?

(a) Acima de um escore-z de 2,10.

(b) Abaixo de um escore-z de −2,10.

(c) Acima de um escore-z de −2,10.

(d) Entre os escores-z de −2,10 e 2,10.

4.16 Encontre o escore-z tal que menos do que 1% dos valores de uma distribuição normal estejam abaixo dele.

4.17 A Mensa é uma sociedade de pessoas com QI alto cujos membros têm um escore em um teste de QI no percentil $98^{\underline{o}}$ ou acima.

(a) Quantos desvios padrão acima da média está o $98^{\underline{o}}$ percentil?

(b) Para a distribuição normal do QI com média 100 e desvio padrão 16, qual é o escore do QI para o $98^{\underline{o}}$ percentil?

4.18 De acordo com um recente Current Population Reports (Relatório Atual da População), indivíduos autônomos nos Estados Unidos trabalham uma média de 45 horas por semana, com um desvio padrão de 15. Se essa variável for aproximadamente distribuída normalmente, qual a proporção que tem uma média maior do que 40 horas por semana?

4.19 O Mental Development Index – MDI (Índice do Desenvolvimento Mental) das Escalas Bayley de Desenvolvimento Infantil é uma medida padronizada usada em estudos com crianças de alto risco. Ela tem uma distribuição aproximadamente normal com média de 100 e desvio padrão 16.

(a) Qual a proporção das crianças que têm um MDI de pelo menos 120?

(b) Encontre o escore do MDI que corresponde ao $90^{\underline{o}}$ percentil.

(c) Encontre e interprete o quartil inferior, o médio e o superior do MDI.

4.20 Para as 5459 mulheres grávidas que se trataram no Aarhus University Hospital na Dinamarca em um período de dois anos e que tiveram registro sobre a duração da gestação, a média foi de 281,9 dias com um desvio padrão de 11,4 dias.[3] Um bebê é classificado como prematuro se o período de gestação é de 258 dias ou menos.

(a) Se o período de gestação é normalmente distribuído, qual é a proporção de bebês nascidos prematuramente?

(b) A proporção real de nascidos prematuramente durante esse período foi de 0,036. Com base nessa informação, como você esperaria que a distribuição do período da gestação diferisse da normal?

4.21 Suponha que o uso semanal de gasolina para viagens de carro por adultos norte-americanos seja aproximadamente normal com uma média de 16 galões e desvio padrão de 5 galões.

(a) Qual a proporção de adultos que usam mais do que 20 galões de gasolina por semana?

(b) Assumindo que o desvio padrão e a forma normal permaneçam constantes, a que nível a média deve ser reduzida para que somente 5% usem mais do que 20 galões por semana?

(c) Se a distribuição do uso da gasolina não é na verdade normal, como você esperaria que ela se desviasse da normal?

4.22 No exame do meio do semestre de estatística introdutória, um professor sempre dá grau B para os estudantes que têm um escore entre 80 e 90. Em um determinado ano, os escores têm aproximadamente uma distribuição normal com um escore médio de 83 e um desvio padrão de 5. Aproximadamente, qual é a proporção de alunos que obtêm um B?

4.23 Para uma distribuição do SAT ($\mu = 500$, $\sigma = 100$) e uma do ACT ($\mu = 21$, $\sigma = 4{,}7$), qual o escore que é relativamente mais alto, um SAT = 600 ou um ACT = 29. Explique.

4.24 Suponha que o imposto predial de casas na cidade de Iowa tem aproximadamente uma distribuição normal com uma média de \$2500 e um desvio padrão de \$1500. O imposto predial para uma casa em particular é de \$5000.

(a) Encontre o escore-z correspondente a esse valor.

(b) Qual é a proporção dos impostos prediais que excedem \$5000?

(c) Se a distribuição verdadeira não é normal, como você acha que ela se desvia da normal? Por quê?

4.25 Um estudo sobre energia em Gainesville, Flórida, descobriu que, em março de 2006, o uso de eletricidade por domicílio tinha uma média de 673 kWh (quilowatt-hora) com um desvio padrão de 556 kWh.

(a) Se a distribuição fosse normal, qual o percentual de domicílios que teriam um uso acima de 1000 kWh?

(b) Você acha que a distribuição é verdadeiramente normal? Por que sim ou por que não?

4.26 Cinco estudantes, Ann, Betty, Clint, Douglas e Edward, foram igualmente avaliados na classificação para a admissão na faculdade de Direito, à frente de outros candidatos. Entretanto, todas as posições menos duas foram preenchidas pela turma de ingressantes. Visto que o comitê de ingresso aceita somente mais dois estudantes, ele decide aleatoriamente selecionar dois destes cinco candidatos. Para essa estratégia, seja y = o número de mulheres aceitas. Usando a primeira letra do nome para representar um estudante, as diferentes combinações que poderiam ser aceitas são (A, B), (A, C), (A, D), (A, E), (B, C), (B, D), (B, E), (C, D), (C, E) e (D, E).

(a) Construa a distribuição da probabilidade para y.

(b) Construa a distribuição amostral da *proporção* da amostra dos estudantes selecionados que são mulheres.

4.27 Construa a distribuição amostral da proporção amostral de caras (CA) ao se lançar uma moeda equilibrada:

(a) Uma vez.

(b) Duas vezes. (*Dica:* as amostras possíveis são (CA, CA), (CA, CO), (CO, CA), (CO, CO).)

(c) Três vezes. (*Dica:* existem 8 amostras possíveis.)

(d) Quatro vezes. (*Dica:* existem 16 amostras possíveis.)

(e) Descreva como a forma da distribuição amostral parece mudar à medida que o número de lançamentos aumenta.

4.28 A distribuição de probabilidade associada ao resultado de lançar um dado tem a probabilidade de 1/6 vinculada a cada inteiro, {1, 2, 3, 4, 5, 6}. Considere (y_1, y_2) como o resultado de lançar o dado duas vezes.

(a) Enumere os 36 pares (y_1, y_2) possíveis (por exemplo, (2, 1) representa um 2 seguido de um 1).

(b) Tratando os 36 pares como igualmente prováveis, construa a distribuição amostral para a média da amostral $\bar{y}$ dos dois números obtidos.

(c) Construa um histograma da (i) distribuição de probabilidade de cada lançamento, (ii) da distribuição amostral de $\bar{y}$ em (b). Descreva as formas dessas distribuições.

(d) Quais são as médias das duas distribuições em (c)? Por que elas são as mesmas?

(e) Explique por que a distribuição amostral de $\bar{y}$ tem relativamente mais valores próximos ao meio do que próximos aos valores mínimo e máximo. (*Dica:* observe que existem muito mais pares (y_1, y_2) que têm uma média da amostra próxima ao meio do que próxima do mínimo e máximo.)

4.29 Uma pesquisa de boca de urna de 2293 eleitores na eleição para senador em Ohio, em 2006, indicou que 44% votaram no candidato Republicano, Mike DeWine, e 56% votaram no candidato Democrata, Sherrod Brown.

(a) Se realmente 50% da população votou em DeWine, encontre o erro padrão da proporção amostral de quem votou nele, para esta pesquisa de boca de urna. (Lembre-se do Exemplo 4.8 da página 111 e que o desvio padrão da população é 0,50.)

(b) Se realmente 50% da população votou em DeWine, seria surpreendente obter os resultados dessa pesquisa de boca de urna? Por quê?

(c) Baseado na sua resposta em (b), você estaria disposto a prever o resultado desta eleição? Explique?

4.30 De acordo com o *Current Population Reports* (Relatório Atual da População), a distribuição da população em números de anos de escolaridade, para indivíduos autônomos nos Estados Unidos, tem uma média de 13,6 e um desvio padrão de 3,0. Encontre a média e o desvio padrão da distribuição amostral de $\bar{y}$ para uma amostra aleatória de:

(a) 9 residentes,

(b) 36 residentes,

(c) 100 residentes. Descreva o padrão à medida que n aumenta.

4.31 Considere o Exercício 4.6. A média e o desvio padrão da distribuição de probabilidade para os prêmios da loteria y são $\mu = 0{,}10$ e $\sigma = 316{,}23$. Suponha que você jogue na loteria um milhão de vezes. Deixe $\bar{y}$ representar sua média de prêmios.

(a) Encontre a média e o erro padrão da distribuição amostral de $\bar{y}$.

(b) Aproximadamente, quão provável é que você obtenha para a sua média de ganhos o valor de $1 além da quantia que você pagou para jogar cada vez?

4.32 De acordo com uma recente Pesquisa Social Geral (variável "PARTNERS"), nos Estados Unidos, a distribuição de y = número de parceiros sexuais que você teve nos últimos 12 meses tem uma média e um desvio padrão de aproximadamente 1,1. Suponha que esses valores sejam a média e o desvio padrão da população.

(a) A variável y tem uma distribuição normal? Explique.

(b) Para uma amostra aleatória de 2400 adultos (o tamanho da PSG de 2006 para esta variável), descreva a distribuição amostral de $\bar{y}$ fornecendo sua forma, média e erro padrão.

(c) Considerando o item (b). Informe um intervalo dentro do qual a média da amostra estaria quase certamente.

4.33 Os escores do Psychomotor Development Index – PDI (Índice do Desenvolvimento Psicomotor), uma escala do desenvolvimento infantil, são aproximadamente normais com média 100 e desvio padrão 15.

(a) Uma criança é selecionada ao acaso. Encontre a probabilidade de que o PDI esteja abaixo de 90.

(b) Um estudo usa uma amostra aleatória de 25 crianças. Especifique a distribuição amostral da média amostral do PDI e encontre a probabilidade de que a média da amostra esteja abaixo de 90.

(c) Você ficaria surpreso em observar um escore do PDI de 90? Você ficaria surpreso em observar uma média da amostra do PDI de 90? Por quê?

(d) Esboce a distribuição da população para o PDI. Sobreponha um diagrama da distribuição amostral para $n = 25$.

4.34 Um estudante planeja amostrar aleatoriamente 100 registros do governo de fazendas em Ontário para estimar a área média, em acres, das fazendas naquela província. Os resultados de um estudo anterior sugerem que 200 acres é uma hipótese razoável para o desvio padrão do tamanho populacional das fazendas.

(a) Encontre a probabilidade de que a média da amostra da área em acres esteja dentro de 10 acres da média da área em acres da população.

(b) Se na realidade o desvio padrão da população for maior do que 200, a probabilidade seria maior ou menor do que a que você encontrou em (a)?

4.35 De acordo com a Agência do Censo dos Estados Unidos, em 2000, o número de pessoas em cada domicílio tinha uma média de 2,6 e um desvio padrão de 1,5. Suponha que a Agência Censitária, ao contrário, tinha estimado esta média usando uma amostra aleatória de 225 domicílios e que a amostra tinha uma média de 2,4 e um desvio padrão de 1,4.

(a) Identifique a variável y.

(b) Descreva o centro e a dispersão da distribuição da população.

(c) Descreva o centro e a dispersão da distribuição dos dados amostrais.
(d) Descreva o centro e a dispersão da distribuição amostral da média da amostra para 225 domicílios. O que esta distribuição descreve?

4.36 A distribuição do tamanho de uma família, em uma sociedade tribal em particular, é assimétrica à esquerda, com $\mu = 5{,}2$ e $\sigma = 3{,}0$. Esses valores são desconhecidos para uma antropóloga que amostra famílias para estimar o seu tamanho médio. Para uma amostra aleatória de 36 famílias, ela obtém uma média de 4,6 e um desvio padrão de 3,2.
(a) Identifique a distribuição da população. Informe a sua média e o seu desvio padrão.
(b) Identifique a distribuição dos dados amostrais. Determine a média e o desvio padrão.
(c) Identifique a distribuição amostral de $\bar{y}$. Determine a sua média e erro padrão e explique o que ela descreve.

4.37 Considere o exercício anterior.
(a) Encontre a probabilidade de que a média da amostra esteja dentro de 0,5 da média da população.
(b) Suponha que a antropóloga colete uma amostra aleatória do tamanho 100. Encontre a probabilidade de que a média da amostra esteja dentro de 0,5 da média verdadeira e compare essa resposta com a item (a).
(c) Considere o item (b). Se a amostra fosse verdadeiramente aleatória, você ficaria surpreso se a antropóloga obtivesse $\bar{y} = 4{,}0$? Por quê? (Isto muito bem poderia acontecer se a amostra não fosse aleatória.)

4.38 Em uma universidade, 60% dos 7400 estudantes são mulheres. O jornal dos estudantes relata um levantamento de dados de uma amostra aleatória de 50 estudantes sobre vários tópicos incluindo o abuso de álcool, como a participação em consumo desenfreado de bebidas (bebedeiras). Eles informam que a amostra utilizada continha 26 mulheres.
(a) Explique como você pode estabelecer uma variável y para representar o sexo.
(b) Identifique a distribuição da população do sexo nesta universidade.
(c) Identifique a distribuição da amostra dos dados do sexo para esta amostra.
(d) A distribuição amostral da proporção amostral de mulheres é aproximadamente uma distribuição normal com média 0,60 e erro padrão 0,07. Explique o que isto significa.

4.39 Sunshine City foi projetada para atrair pessoas aposentadas. Sua população atual de 50000 residentes tem uma idade média de 60 anos e um desvio padrão de 16 anos. A distribuição das idades é assimétrica à esquerda, refletindo o predomínio de indivíduos mais velhos. Uma amostra aleatória de 100 residentes de Sunshine City tem $\bar{y} = 58{,}3$ e $s = 15{,}0$.
(a) Descreva o centro e a dispersão da distribuição da população.
(b) Descreva o centro e a dispersão da distribuição da amostra dos dados. Que forma ela provavelmente tem?
(c) Encontre o centro e a dispersão da distribuição amostral de $\bar{y}$ para $n = 100$. Que forma ela tem e o que ela descreve?
(d) Explique por que não seria incomum observar uma pessoa com a idade de 40 anos em Sunshine City e seria altamente incomum observar uma média amostral de 40, para um tamanho da amostra de 100.

4.40 Considere o exercício anterior.
(a) Descreva a distribuição amostral de $\bar{y}$ para uma amostra aleatória do tamanho $n = 1$.
(b) Descreva a distribuição amostral de $\bar{y}$ se você amostrar todos os 50000 residentes.

Conceitos e aplicações

4.41 Você pode usar um ***applet*** em um computador ou na internet para gerar repetidamente amostras aleatórias de uma população artificial e analisá-las para estudar as propriedades dos métodos estatísticos. Para tentar isto, vá a www.grupoa.com.br e use o *applet* da *distribuição amostral*. Selecione *binário* para a população de origem, ajustando a proporção da população para 0,50. Selecione para o tamanho da amostra $n = 100$.
(a) Simule uma vez (ajustando o número de simulações $N = 1$ e clicando em [*Sample*]) e relate as frequências e proporções para as duas categorias. Você obteve uma proporção amostral próxima a 0,50? Execute essa simulação de uma amostra aleatória de tamanho 100, dez vezes, observando nos gráficos a cada vez as frequências e a proporção amostral de votos *sim*. Resuma.
(b) Agora faça um gráfico da simulação de uma amostra aleatória de tamanho 100 e determine a proporção amostral 1000 vezes, ajustando o $N = 1000$ no menu. Como este gráfico reflete o Teorema Central do Limite?

4.42 Considere o exercício anterior.
(a) Para esse *applet*, selecione a distribuição da população assimétrica. Colete 1000 amostras de tamanho 30, cada uma. Como a distribuição empírica das médias amostrais se compara à distribuição da população? O que isto reflete?
(b) Repita, escolhendo, dessa vez, um tamanho da amostra de somente dois. Por que a distribuição amostral não é simétrica e com a forma de sino?

4.43 (*Exercício para a Turma*) Considere os Exercícios 1.11 e 1.12 (páginas 25-26). Usando a população definida por sua turma ou usando um levantamento de dados com os alunos, o professor irá selecionar uma variável, como, por exemplo, horas por semana que são passadas vendo televisão.
(a) Construa um histograma ou um diagrama de caule e folhas da distribuição da população dessa variável para a turma.
(b) Usando uma tabela de números aleatórios, cada estudante deve selecionar outros nove estudantes ao acaso e calcular a resposta média da amostra para esses estudantes. (Cada estudante deve usar números aleatórios diferentes.) Faça um histograma das médias das amostras obtidas por todos os estudantes. Como a dispersão e a forma se comparam ao histograma em (a)? O que isto ilustra?

4.44 (*Exercício para a Turma*) A Tabela 4.5 fornece as idades de todos os responsáveis por um domicílio em um pequeno vilarejo de pescadores da Nova Escócia. A distribuição dessas idades é caracterizada por $\mu = 47{,}18$ e $\sigma = 14{,}74$.
(a) Construa um diagrama de caule e folhas para a distribuição da população.
(b) Usando uma tabela de números aleatórios, cada estudante deve selecionar nove números aleatórios entre 01 e 50. Usando esses números, cada estudante deve amostrar nove responsáveis por domicílios e calcular sua idade média amostral. Faça um gráfico da distribuição amostral empírica dos valores de $\bar{y}$. Compare-o à distribuição em (a).
(c) O que você espera para a média dos valores-$\bar{y}$ com muitas repetições utilizando amostras de tamanho 9?
(d) O que você espera para o desvio padrão dos valores-$\bar{y}$ com muitas repetições utilizando amostras de tamanho 9?

4.45 (*Exercício para a Turma*) Para um único arremesso de uma moeda, considere $y = 1$ para cara e $y = 0$ para coroa. Isto simula o voto em uma eleição com dois candidatos igualmente preferidos.
(a) Construa a distribuição da probabilidade de y e encontre a sua média.
(b) A moeda é arremessada dez vezes, gerando seis caras e quatro coroas. Construa a distribuição dos dados amostrais.
(c) Cada estudante da turma deve arremessar uma moeda 10 vezes e calcu-

☑ **Tabela 4.5**

Nome	Idade	Nome	Idade	Nome	Idade	Nome	Idade
Alexander	50	Griffith	66	McTell	49	Staines	33
Bell	45	Grosvenor	51	MacLeod	30	Stewart	36
Bell	23	Ian	57	McNeil	28	Stewart	25
Bok	28	Jansch	40	McNeil	31	Thames	29
Clancy	67	Keelaghan	36	McNeil	45	Thomas	57
Cochran	62	Lavin	38	McNeil	43	Todd	39
Fairchild	41	Lunny	81	Mitchell	43	Trickett	50
Finney	68	MacColl	27	Muir	54	Trickett	64
Fisher	37	McCusker	37	Oban	62	Tyson	76
Francey	60	McCusker	56	Reid	67	Watson	63
Fricker	41	McDonald	71	Renbourn	48	Young	29
Gaughan	70	McDonald	39	Rogers	32		
Graham	47	McDonald	46	Rush	42		

lar a proporção amostral de caras. Resuma a distribuição amostral empírica fazendo um gráfico das proporções para todos os estudantes. Descreva a forma e a dispersão da distribuição dos dados da amostra comparada com a distribuição em (a) e (b).

(d) Se executarmos o experimento de arremessar a moeda 10 vezes, um grande número de vezes, o que obteríamos para (i) a média, (ii) desvio padrão dos valores da proporção amostral? Você pode usar 0,50 como o desvio padrão da distribuição em (a).

4.46 (a) Com qual distribuição a distribuição dos dados da amostra tende a se parecer mais – a amostral ou a da população? Explique.

(b) Explique cuidadosamente a diferença entre *a distribuição dos dados da amostra* e a *amostral* de $\bar{y}$. Ilustre sua resposta para a variável y que pode somente assumir os valores 0 e 1.

4.47 A Palestinian Central Bureau of Statistics (Agência Central de Estatística da Palestina) (www.pcbs.gov.ps) perguntou a mães com idade entre 20 e 24 anos sobre o número ideal de filhos. Para as mães morando na Faixa de Gaza, a distribuição da probabilidade foi aproximadamente $P(1) = 0{,}01$, $P(2) = 0{,}10$, $P(3) = 0{,}09$, $P(4) = 0{,}31$, $P(5) = 0{,}19$ e $P(6 \text{ ou mais}) = 0{,}29$.

(a) Visto que a última categoria é aberta, não é possível calcular exatamente a média. Encontre um limite inferior para a média.

(b) Explique por que você pode determinar a mediana da distribuição e encontre-a.

4.48 Para uma distribuição normal, mostre que:

(a) O quartil superior é igual a $\mu + 0{,}67\sigma$.

(b) De acordo com o critério 1,5(IIQ), um valor atípico é uma observação que está a mais do que 2,7 desvios padrão abaixo ou acima da média, e isso acontece com somente 0,7% dos dados.

4.49 Em uma pesquisa de boca de urna, com 1336 eleitores, da eleição de 2006 para o senado do estado de Nova Iorque, 67% disseram que votaram em Hillary Clinton. Baseado nessa informação, você estaria disposto a prever o vencedor da eleição? Explique o seu raciocínio.

4.50 Para uma pesquisa de boca de urna de uma eleição para o senado, encontre o erro padrão da proporção amostral dos que votaram em um candidato para o qual a proporção da população é de 0,50, quando $n = 100$, 1000, 10000. Em cada caso, determine o intervalo dentro do qual a proporção amostral estará quase certamente. Observe que o intervalo diminui na largura à medida que o tamanho da amostra aumenta. Isto é a consequência da **lei dos grandes números**, que declara que a proporção amostral tende a ficar mais próxima da proporção da população à medida que n aumenta.

Selecione a(s) resposta(s) correta(s) nas questões de múltipla escolha em 4.51-4.52. (Pode haver mais de uma resposta correta.)

4.51 O erro padrão de uma estatística descreve:

(a) O desvio padrão da distribuição amostral daquela estatística.

(b) O desvio padrão dos dados amostrais.

(c) Quão próximo é provável que aquela estatística esteja do parâmetro que ela estima.

(d) A variabilidade nos valores da estatística para amostras aleatórias repetidas de tamanho n.

(e) O erro que ocorre devido a não resposta e erros de mensuração.

4.52 O Teorema Central do Limite implica que:

(a) Todas as variáveis têm distribuições dos dados amostrais em forma de sino se uma amostra aleatória contém pelo menos 30 observações.

(b) As distribuições da população são normais sempre que o tamanho da população for grande.

(c) Para amostras aleatórias grandes, a distribuição amostral de $\bar{y}$ é aproximadamente normal, apesar da forma da distribuição da população.

(d) A distribuição amostral se parece mais com a distribuição da população à medida que o tamanho da amostra aumenta.

(e) Todas as respostas acima.

4.53 Verdadeiro ou falso: à medida que o tamanho da amostra aumenta, o erro padrão da distribuição amostral de $\bar{y}$ também aumenta. Explique a sua resposta.

***4.54** A Lake Wobegon Junior College admite somente estudantes que tenham escores acima de 400 pontos em um teste padrão de desempenho. Os candidatos do grupo A têm uma média de 500 e um desvio padrão de 100; nesse teste, e os candidatos do grupo B apresentam uma média de 450 e um desvio padrão de 100. Os dois grupos têm o mesmo tamanho.

(a) Encontre as proporções de não selecionados para cada grupo.

(b) Dos estudantes que foram selecionados que proporção pertence ao grupo B?

(c) Um membro do governo propõe que o estabelecimento baixe o ponto de corte de admissão para 300, supondo que a proporção de estudantes não classificados do grupo B diminuiria. Se esta política for implementada, determine o efeito na resposta (b) e comente.

***4.55** O desvio padrão de uma distribuição da probabilidade discreta é

$$\sigma = \sqrt{\sum (y - \mu)^2 P(y)}.$$

(a) Suponha que $y = 1$ com probabilidade 0,50 e $y = 0$ com probabilidade 0,50, como no Exemplo 4.8 (página 111). Mostre que $\sigma = 0{,}50$.

(b) Suponha que $y = 1$ com probabilidade π e $y = 0$ com probabilidade $1 - \pi$, onde π representa um número entre 0 e 1. Mostre que $\mu = \pi$ e que $\sigma = \sqrt{\pi(1 - \pi)}$.

(c) Mostre que o erro padrão da proporção amostral para uma amostra aleatória do tamanho n é igual à $\sqrt{\pi(1 - \pi)/n}$.

***4.56** A curva para uma distribuição normal com média μ e desvio padrão σ pode ser determinada pela seguinte expressão:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/(2\sigma^2)}.$$

Mostre que esta curva é assimétrica, mostrando que para qualquer constante c, a curva tem o mesmo valor tanto em $y = \mu + c$ como em $y = \mu - c$. (A integral

de $f(y)$ para y entre $\mu + z\sigma$ e ∞ é igual a probabilidade da cauda apresentada na Tabela A.)

*4.57 A fórmula do erro padrão $\sigma_{\bar{y}} = \sigma/\sqrt{n}$ trata o tamanho da população N como infinitamente grande relativo ao tamanho da amostra n. A fórmula $\sigma_{\bar{y}}$ para uma população finita de tamanho N é:

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right).$$

O termo $\sqrt{(N-n)/(N-1)}$ é chamado de **correção de população finita**.

(a) Quando $n = 300$ estudantes são selecionados de um corpo docente de uma faculdade de tamanho $N = 30000$, mostre que $\sigma_{\bar{y}} = 0.995\sigma/\sqrt{n}$. (Em geral, n é bem menor quando comparado a N; assim a correção, na maioria das vezes, tem pouca influência prática.)

(b) Se $n = N$ (isto é, amostramos toda a população), mostre que $\sigma_{\bar{y}} = 0$. Em outras palavras, não ocorre um erro amostral porque $\bar{y} = \mu$.

(c) Para $n = 1$, explique por que a distribuição amostral de $\bar{y}$ e seu erro padrão são idênticos à distribuição da população e seu desvio padrão.

NOTAS

1 *Journey to Work*, editado em 2004 pela Agência do Censo dos Estados Unidos.
2 www.cnn.com/ELECTION/2006.
3 *British Medical Journal*. V. 307, 1993, p. 234.

5 INFERÊNCIA ESTATÍSTICA: ESTIMAÇÃO

Este capítulo mostra como usar os dados amostrais para estimar os parâmetros da população. Com variáveis quantitativas estimamos a média da população. Um estudo que trata de assuntos do sistema de saúde, por exemplo, pode estimar os parâmetros da população como a quantia média de dinheiro gasta em medicamentos prescritos durante o último ano e o número médio de visitas ao médico. Com variáveis categóricas, estimamos as proporções da população para as categorias. O estudo do sistema de saúde pode estimar as proporções das pessoas que (têm, não têm) seguro de saúde e as proporções que (estão satisfeitas, não estão satisfeitas) com seu plano de saúde.

Inicialmente aprenderemos sobre dois tipos de estimativas dos parâmetros. Após, nas Seções 5.2 e 5.3, as aplicaremos às médias e proporções da população. A Seção 5.4 encontra o tamanho da amostra necessário para alcançar a precisão desejada da estimativa. A Seção 5.5 discute a estimativa da mediana e de outros parâmetros.

5.1 ESTIMAÇÃO POR PONTO E INTERVALAR

Existem dois tipos de estimativas de parâmetros:

- Uma **estimativa por ponto** é um único número que é a melhor avaliação do parâmetro.
- Uma **estimativa intervalar** é um conjunto de números em torno da estimativa por ponto dentro do qual o valor do parâmetro deve estar.

Por exemplo, uma PSG perguntou: "Você acredita em vida após a morte?". Para 1958 sujeitos amostrados, a estimativa por ponto para a proporção de todos os norte-americanos que iriam responder *sim* é igual a 0,73. Uma estimativa intervalar prevê que a proporção da população que respondeu *sim* está entre 0,71 e 0,75. Isto é, ela prevê que a estimativa por ponto de 0,73 está dentro de *uma margem de erro* de 0,02 do valor real. Portanto, uma estimativa intervalar nos ajuda a avaliar a precisão provável de uma estimativa por ponto.

O termo *estimativa*, apenas, é geralmente usado como uma abreviação de *estimativa por ponto*. O termo *estimador* se refere a um tipo específico de estatística para estimar um parâmetro e *estimativa* se refere ao seu valor para uma amostra específica. Por exemplo, a proporção amostral é um estimador da proporção populacional. O valor 0,73 é a estimativa para a proporção da população que acredita em vida após a morte.

Estimativa por ponto

Qualquer parâmetro em particular tem muitos estimadores possíveis. Para uma população normalmente distribuída, por exemplo, o centro é a média e a mediana, uma vez que ela é simétrica. Assim, com dados amostrais, dois estimadores possíveis do centro da população são a média e a mediana da amostra.