*5.73 Você conhece a média amostral de n observações. Uma vez que você conhece $(n-1)$ das observações, mostre que você pode encontrar a remanescente. Em outras palavras, para um dado valor de $\bar{y}$, os valores de $(n-1)$ observações determinam a remanescente. Ao obter escores de uma variável quantitativa, ter $(n-1)$ graus de liberdade significa que apenas aquela quantidade de observações é independente.

*5.74 Encontre o erro padrão da proporção amostral quando $\pi = 0$ ou $\pi = 1$. O que isto reflete?

*5.75 Seja π a probabilidade de que um eleitor aleatoriamente selecionado prefira o candidato Republicano. Você amostra 2 pessoas e nenhuma delas prefere o Republicano. Encontre a estimativa por ponto de π. Esta estimativa parece lógica? Por quê? (A estimativa *bayesiana* é uma estimativa alternativa que utiliza uma abordagem *subjetiva*, combinando os dados da amostra com opinião prévia sobre π antes de ver os dados. Por exemplo, se você achava que π era igualmente provável de estar entre 0 e 1, a estimativa bayesiana adiciona duas observações, uma de cada tipo, assim gerando a estimativa ¼.)

*5.76 Para encorajar os sujeitos a dar respostas sobre questões delicadas, o método de **resposta aleatorizada** é geralmente usado. É solicitado ao sujeito jogar uma moeda em segredo. Se for cara, o sujeito joga a moeda uma vez mais e relata o resultado, cara ou coroa. Se, ao contrário, o primeiro arremesso for coroa, o sujeito relata, então, a resposta a uma questão delicada; por exemplo, relata a resposta *cara* se a verdadeira resposta for *sim* e relata a resposta *coroa* se a verdadeira resposta for *não*. Considere π a representação da probabilidade verdadeira da resposta *sim* à questão delicada.

(a) Explique por que os números na Tabela 5.5 são as probabilidades dos quatro resultados possíveis.

(b) Considere p a representação da proporção amostral de sujeitos que relataram *cara* para a segunda resposta. Explique por que $\hat{\pi} = 2p - 0{,}5$ estima π.

(c) Usando esta abordagem, 200 sujeitos foram perguntados se nunca tinham fraudado intencionalmente seu imposto de renda. Relate a estimativa de π se o número de caras relatadas é igual a (i) 50, (ii) 70, (iii) 100, (iv) 150.

☑ Tabela 5.5

	Segunda resposta	
Primeira moeda	Cara	Coroa
Cara	0,25	0,25
Coroa	$\pi/2$	$(1-\pi)/2$

*5.77 Para construir um intervalo de confiança para amostras grandes para a proporção π, não é necessário substituir $\hat{\pi}$ para o valor desconhecido de π na fórmula para o erro padrão de $\hat{\pi}$. Um método menos aproximado encontra os extremos de um intervalo de 95% de confiança determinando os valores de π que estão a 1,96 erros padrão da proporção amostral, solucionando para π a equação

$$|\hat{\pi} - \pi| = 1{,}96\sqrt{\frac{\pi(1-\pi)}{n}}.$$

Para o Exemplo 5.8 (página 153) com nenhum vegetariano em uma amostra de tamanho 20, substitua $\hat{\pi}$ e n nesta equação e mostre que a equação é satisfeita em $\pi = 0$ e em $\pi = 0{,}161$. Portanto, o intervalo de confiança é (0; 0,161).

☑ NOTAS

1 Aqui, π *não* é a constante matemática, 3,1415...

2 Veja sda.berkeley.edu/D3/GSS06/Doc/gs06.htm.

3 Cortesia do Prof. Brian Everitt, Instituto de Psiquiatria, Londres.

4 Veja o artigo de A. Agresti e B. Coull (que propôs esse intervalo de confiança), *American Statistician*, v. 52, p. 119-26, 1998.

☑ 6 INFERÊNCIA ESTATÍSTICA: TESTES DE SIGNIFICÂNCIA

O objetivo de muitos estudos é verificar se os dados concordam com certas previsões. As previsões geralmente resultam da teoria que leva à pesquisa. Essas previsões são *hipóteses* sobre a população em estudo.

☑ **Hipótese**

Na estatística, uma **hipótese** é uma afirmação sobre a população. Ela é geralmente uma previsão na qual um parâmetro que descreve uma característica de uma variável assume um valor numérico particular ou está em certo intervalo de valores.

Exemplos de hipóteses são os seguintes: "Para prestadores de serviço, a renda média é a mesma tanto para mulheres quanto para homens", "Não existe diferença em termos probabilísticos entre Democratas e Republicanos em relação ao voto seguindo a liderança do seu partido" e "A metade ou mais dos adultos canadenses está satisfeita com seu serviço nacional de saúde".

Um **teste de significância** usa dados para resumir a evidência sobre uma hipótese, comparando as estimativas por pontos dos parâmetros aos valores previstos pela hipótese. O seguinte exemplo ilustra os conceitos por trás dos testes de significância.

EXEMPLO 6.1 Testando a tendenciosidade de gênero na seleção de gerentes

Uma grande rede de supermercados na Flórida selecionou alguns empregados para receber treinamento para a gerência. Um grupo de funcionários femininos alegou que os homens foram selecionados em uma taxa desproporcionalmente alta para tal treinamento. A empresa negou esta alegação.[1] Uma alegação similar de tendenciosidade de gênero foi feita sobre as promoções e salários para mulheres que trabalham na Wal-Mart.[2] Como as funcionárias poderiam respaldar estatisticamente sua declaração?

Suponha que o grupo potencial de funcionários para uma seleção para o treinamento para gerente é formado meio a meio por homens e mulheres. Portanto, a alegação de não tendenciosidade de gênero é uma hipótese. Ela afirma que, mantendo as demais condições constantes, a cada escolha, a probabilidade de selecionar uma mulher é igual a ½ e a probabilidade de selecionar um homem é igual a ½. Se os funcionários realmente são selecionados aleatoriamente para o treinamento para gerente em termos de gênero, aproximadamente metade dos funcionários selecionados deveria ser de mulheres e metade deveria ser de homens. A alegação das mulheres é uma hipótese alternativa de que a probabilidade de selecionar um homem é maior do que ½.

Suponha que nove dos 10 funcionários escolhidos para o treinamento para gerente foram homens. Podemos ficar inclinados a acreditar na alegação das mulheres. Entretanto, devemos analisar se es-

ses resultados seriam improváveis, se *não* existisse tendenciosidade de gênero. Seria altamente incomum que 9/10 dos funcionários escolhidos teriam o mesmo gênero se eles fossem realmente selecionados ao acaso do grupo de funcionários? Devido à variação amostral, não exatamente ½ da amostra precisa ser masculina. Quanto acima de ½ deve estar a proporção amostral de homens escolhidos para acreditarmos na alegação das mulheres?

Este capítulo introduz métodos estatísticos para resumir evidências e tomar decisões sobre hipóteses. Inicialmente apresentamos as partes que todos os testes de significância têm em comum. O restante do capítulo apresenta testes de significância para médias e proporções populacionais. Aprenderemos, também, como encontrar e controlar a probabilidade de uma decisão incorreta a respeito de uma hipótese.

6.1 AS CINCO ETAPAS DE UM TESTE DE SIGNIFICÂNCIA

Agora, vamos observar com mais detalhes o método do teste de significância, também chamado de *teste de hipótese* ou simplesmente *teste*. Todos os testes têm cinco partes: suposição, hipóteses, estatística-teste, valor-*p* e conclusão.

Suposições

Cada teste faz certas suposições ou tem certas condições para ser válido. Elas dizem respeito ao seguinte:

- **Tipo de dados:** como outros métodos estatísticos, cada teste se aplica tanto para os dados quantitativos quanto para os dados categóricos.
- **Aleatorização:** como o método de intervalo de confiança da inferência estatística, um teste assume que os dados foram obtidos usando a aleatorização, isto é, como uma amostra aleatória.
- **Distribuição da população:** para alguns testes, assume-se que a variável tenha uma distribuição particular, como a distribuição normal.
- **Tamanho da amostra:** a validade de muitos testes melhora à medida que o tamanho da amostra aumenta.

Hipóteses

Cada teste de significância tem duas hipóteses sobre o valor de um parâmetro.

☑ **Hipótese nula, hipótese alternativa**

A **hipótese nula** é uma afirmação de que o parâmetro assume um valor em particular. A **hipótese alternativa** declara que o parâmetro está em um intervalo alternativo de valores. Geralmente o valor na hipótese nula corresponde, em certo sentido, a *sem efeito*. Os valores na hipótese alternativa, então, representam um efeito de certo tipo.

☑ **Notação para hipóteses**

O símbolo H_0 representa a hipótese nula. O símbolo H_a(*) representa a hipótese alternativa.

Considere o Exemplo 6.1 sobre a possível discriminação de gênero na seleção de *trainees* para gerente. Considere π a da probabilidade de que qualquer seleção em particular seja um homem. A empresa afirma que $\pi = ½$. Isto é um exemplo de uma hipótese nula, *sem efeito* referindo-se à ausência de tendenciosidade de gênero. A hipótese alternativa reflete a opinião cética das funcionárias de que essa probabilidade realmente exceda a ½. Assim, as hipóteses são H_0: $\pi = ½$ e H_a: $\pi > ½$. Observe que H_0 tem um *único* valor enquanto H_a tem um intervalo de valores.

* N. de T. T.: Essa é uma preferência desses autores, contudo a maioria dos textos de estatística utiliza a notação H_1.

Um teste de significância analisa a evidência amostral sobre a hipótese nula, H_0. O teste investiga se os dados contradizem H_0, portanto sugerindo que H_a é verdadeira. A abordagem feita é a indireta da *prova por contradição*. A hipótese nula é presumida ser verdadeira. Sob esta suposição, se os dados observados fossem muito incomuns, a evidência suportaria a hipótese alternativa. No estudo da discriminação potencial de gênero, assumimos que o valor da hipótese nula, $\pi = ½$, é verdadeiro. Então, determinamos se o resultado da amostra na seleção de 9 homens para o treinamento de gerente em 10 escolhas seria incomum, sob esta suposição. Se for assim, então poderemos estar inclinados a acreditar na reclamação das mulheres. Mas, se a diferença entre a proporção amostral de homens escolhidos (9/10) e o valor de H_0 de ½ pudesse facilmente ser atribuída à variabilidade amostral usual, então não teríamos evidência suficiente para aceitar a reclamação das mulheres.

Um pesquisador geralmente conduz um teste para avaliar o grau de suporte para a hipótese alternativa. Assim, H_a é, algumas vezes, chamada de **hipótese de pesquisa**. As hipóteses são formuladas *antes* de se coletar e analisar os dados.

Estatística-teste

O parâmetro ao qual a hipótese se refere tem uma estimativa por ponto. A **estatística-teste** resume quão longe essa estimativa está do valor suposto do parâmetro em H_0. Geralmente a diferença é expressa em números de erros padrão que a estimativa dista do valor de H_0.

Valor-*p*

Para interpretar um valor da estatística-teste, criamos um resumo probabilístico da evidência contra H_0. Para isso é utilizada a distribuição amostral da estatística-teste sob a suposição de que a hipótese nula seja verdadeira. A finalidade é determinar quão incomum o valor da estatística-teste observado é quando comparado com o que H_0 prevê.

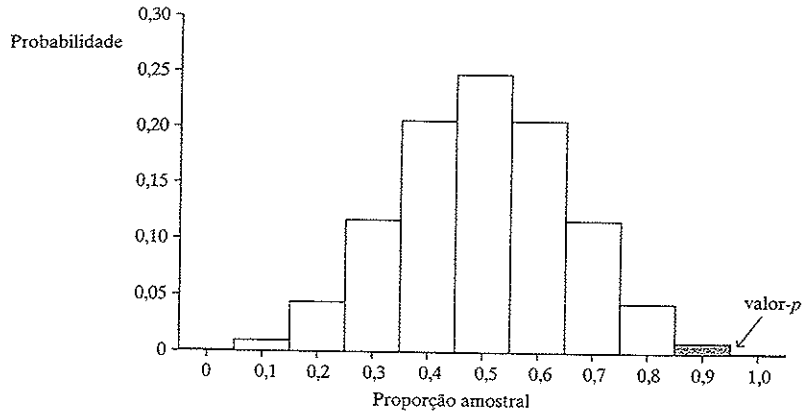
Especificamente, se a estatística-teste está bem no final da cauda da distribuição amostral em uma direção prevista por H_a, então está longe do que H_0 prevê. Podemos resumir quão distante a estatística-teste está na cauda pela probabilidade da cauda daquele valor e pelos valores mais extremos. Estes são os valores possíveis da estatística-teste que fornecem *pelo menos tanta evidência contra H_0 quanto a estatística-teste observada*, na direção prevista pela H_a. Essa probabilidade é denominada **valor-*p***.

☑ **Valor-*p***

O **valor-*p*** é a probabilidade de que a estatística-teste seja igual ou mais extrema que o valor observado na direção prevista pela H_a. Ele é calculado presumindo que H_0 seja verdadeira. O valor-*p* é representado por *p*.

Um valor-*p* pequeno (como $p = 0{,}01$) significa que os dados observados seriam incomuns, se H_0 fosse verdadeira. *Quanto menor o valor-*p*, mais forte a evidência contra H_0.*

Para o Exemplo 6.1 da discriminação potencial de gênero na escolha de *trainees* para gerente, π é a probabilidade da seleção de um homem. Testamos H_0: $\pi = ½$ contra H_a: $\pi > ½$. Uma estatística-teste possível é a proporção amostral de homens selecionados, que é 9/10 = 0,90. Os valores para a proporção amostral que fornecem esta evidência ou uma evidência ainda mais extrema contra H_0: $\pi = ½$ e a favor de H_a: $\pi > ½$ são os valores da proporção amostral à direita de 0,90. Veja a Figura 6.1. A fórmula da Seção 6.7 calcula esta probabilidade como sendo 0,01, assim o valor-*p* é igual a $p = 0{,}01$. Se as seleções tivessem sido realmente aleatórias com respeito ao



☑ **Figura 6.1** O valor-*p* é igual à probabilidade do valor observado ou de resultados mais extremos. Ele é calculado sob a suposição de que a H_0 seja verdadeira, assim um valor-*p* bem pequeno fornece uma forte evidência contra a H_0.

gênero, a probabilidade seria de somente 0,01 de um resultado amostral extremo, a saber, de que nove ou 10 seleções seriam homens. Mantendo-se os demais valores constantes, este valor-*p* pequeno fornece evidência considerável contra H_0: $\pi = ½$ e suporta a alternativa H_a: $\pi > ½$ de discriminação contra as mulheres.

Em contraposição, um valor-*p* de moderado a grande significa que os dados são consistentes com H_0. Um valor-*p* como 0,26 ou 0,83 indica que, se H_0 for verdadeira, os dados observados não seriam incomuns.

Conclusão

O valor-*p* resume a evidência contra H_0. Nossa conclusão deveria também *interpretar* o que o valor-*p* nos diz sobre a questão que motivou o teste. Algumas vezes, é necessário tomar uma decisão sobre a validade de H_0. Se o valor-*p* for suficientemente pequeno, nós rejeitamos H_0 e aceitamos H_a.

Muitos estudos requerem valores-*p* bem pequenos, como $p \leq 0{,}05$, para rejeitar H_0. Nesses casos, os resultados são ditos serem *significativos no nível 0,05*. Isso quer dizer que, se H_0 fosse verdadeira, a chance de obter resultados tão extremos como nos dados amostrais não seria maior do que 0,05.

Tomar uma decisão de rejeitar ou não a hipótese nula é uma parte opcional do teste de significância. Continuaremos com este assunto até a Seção 6.4. A Tabela 6.1 resume as etapas do teste de significância.

6.2 TESTE DE SIGNIFICÂNCIA PARA UMA MÉDIA

Para as variáveis quantitativas, os testes de significância geralmente se referem à média da população μ. Seguem as cinco etapas do teste de significância.

As cinco etapas de um teste de significância para uma média

1. Suposições

O teste assume que os dados são obtidos por meio de uma amostra aleatória. A variável quantitativa é presumida ter uma distribuição populacional normal. Veremos que isso é relevante principalmente para tamanhos de amostra pequenos e certos tipos de H_a.

☑ **Tabela 6.1** As cinco etapas de um teste de significância estatística

1. **Suposição**
Tipos de dados, aleatorização, distribuição populacional, condição do tamanho da amostra
2. **Hipóteses**
Hipótese nula, H_0 (valor do parâmetro para "sem efeito")
Hipótese alternativa, H_a (valores alternativos do parâmetro)
3. **Estatística-teste**
Compara a estimativa por ponto ao valor do parâmetro H_0
4. **Valor-*p***
Peso da evidência contra H_0; *p* pequeno é evidência forte
5. **Conclusão**
Relatar o valor-*p*
Decisão formal (opcional; veja a Seção 6.4)

2. Hipóteses

A hipótese nula para uma média populacional μ tem a forma:

$$H_0\text{: } \mu = \mu_0,$$

onde μ_0 é um valor particular para a média da população. Em outras palavras, o valor hipotético de μ em H_0 é um único valor. Esta hipótese geralmente se refere a *sem efeito* ou *sem mudança* comparada a um padrão. Como ilustração, o Exemplo 5.5 do capítulo anterior (página 144) estimou a mudança média do peso da população μ para meninas adolescentes após receberem um tratamento para anorexia. A hipótese de que o tratamento não teve *efeito* é a hipótese nula, H_0: $\mu = 0$. Aqui, o valor μ_0 da hipótese nula para o parâmetro μ é 0.

A hipótese alternativa contém valores alternativos ao parâmetro em H_0. A hipótese alternativa mais comum é:

$$H_a\text{: } \mu \neq \mu_0, \text{ como } H_a\text{: } \mu \neq 0.$$

Essa hipótese é denominada de **bilateral**, porque contém valores tanto acima como abaixo do valor fixado em H_0. Para o estudo de anorexia, H_a: $\mu \neq 0$ declara que o tratamento tem *algum efeito*, a média da população é um valor diferente de 0.

3. Estatística-teste

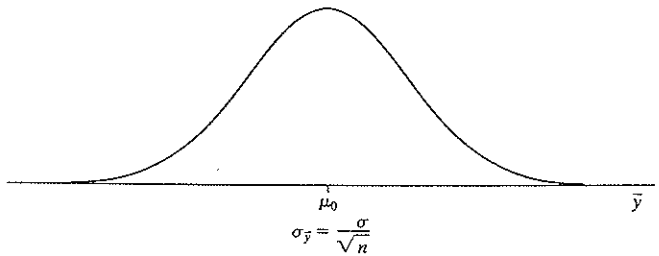
A média amostral $\bar{y}$ estima a média da população μ. Quando a distribuição populacional é normal, a distribuição amostral de $\bar{y}$ é normal em torno de μ. Isto é também aproximadamente verdadeiro quando a distribuição populacional *não* é normal, mas o tamanho da amostra aleatória é relativamente grande considerando o Teorema Central do Limite.

Sob a suposição de que H_0: $\mu = \mu_0$ é verdadeira, o centro da distribuição amostral de $\bar{y}$ é o valor μ_0, como a Figura 6.2 mostra. Um valor de $\bar{y}$ que esteja bem longe do centro em uma das caudas fornece forte evidência contra H_0, porque seria incomum se verdadeiramente $\mu = \mu_0$. A evidência sobre H_0 é resumida pelo número de erros padrão que $\bar{y}$ está distante do valor da hipótese nula μ_0.

Lembre que o erro padrão *verdadeiro* é $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. Como no Capítulo 5, substituímos o desvio padrão, σ desconhecido, da população pelo desvio padrão da amostra s para obter o erro padrão *estimado*, $ep = s/\sqrt{n}$. A estatística-teste é o escore-*t*

$$t = \frac{\bar{y} - \mu_0}{ep}, \text{ onde } ep = \frac{s}{\sqrt{n}}.$$

Quanto mais longe $\bar{y}$ está de μ_0, maior será o valor absoluto da estatística-teste *t*. Portanto, quanto maior for o valor de |*t*|, mais forte a evidência contra H_0.



☑ **Figura 6.2** Distribuição amostral de $\bar{y}$ se H_0: $\mu = \mu_0$ é verdadeira. Para grandes amostras aleatórias, ela é aproximadamente normal e centrada no valor da hipótese nula, μ_0.

Usamos o símbolo t em vez de z porque, como na construção de um intervalo de confiança, usar s para estimar σ no erro padrão introduz um erro adicional. A distribuição amostral nula da estatística-teste t é a *distribuição* t, introduzida na Seção 5.3 (página 140). Ela se parece com a distribuição padrão normal, tendo a média igual a 0, mas sendo mais dispersa quanto menor for o tamanho da amostra (n). Ela é especificada pelos seus graus de liberdade, $gl = n - 1$.

4. Valor-*p*

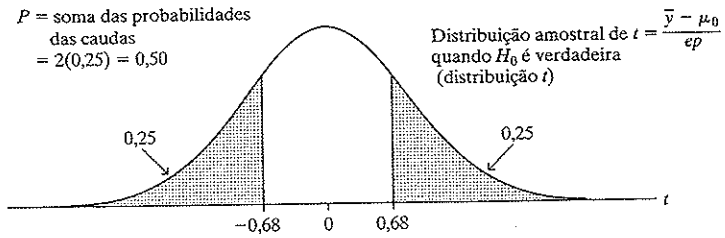
A estatística-teste resume quão longe os dados estão de H_0. Testes diferentes usam estatísticas diferentes e interpretações mais simples resultam de transformá-las em uma escala probabilística de 0 a 1. O valor-p faz isso.

Calculamos o valor-p sob a suposição de que H_0 é verdadeira. Isto é, damos o benefício da dúvida para H_0, analisando quão incomum os dados observados seriam se H_0 fosse verdadeira. O valor-p é a probabilidade de que a estatística-teste seja igual ao valor observado ou ainda mais extremo, fornecendo evidências ainda mais fortes contra H_0. Para H_a: $\mu \neq \mu_0$, os valores t mais extremos são aqueles mais distantes nas caudas da distribuição t. Assim, o valor-p é a soma das probabilidades de duas caudas de que a estatística-teste t seja, pelo menos, tão grande em valor absoluto quanto a estatística-teste observada. Isto é, também, a probabilidade de que $\bar{y}$ esteja, pelo menos, tão longe de μ_0, em *qualquer direção*, quanto o valor de $\bar{y}$ que foi observado.

A Figura 6.3 mostra a distribuição amostral da estatística-teste t quando H_0 é verdadeira. Um valor da estatística-teste $t = (\bar{y} - \mu_0)/ep = 0$ ocorre quando $\bar{y} = \mu_0$. Este seria o valor-t mais consistente com H_0. O valor-p é a probabilidade de que um valor da estatística-teste t esteja, pelo menos, tão longe deste valor consistente como o observado. Para ilustrar seu cálculo, suponha que $t = 0{,}68$, para um tamanho da amostra de 186. (Este é o resultado no Exemplo 6.2.) Esse escore-t quer dizer que a média amostral $\bar{y}$ está a 0,68 erros padrão estimados distante de μ_0.

O valor-p é a probabilidade de que $t \geq 0{,}68$ ou $t \leq -0{,}68$ (isto é, $|t| \geq 0{,}68$). Visto que $n = 186$, o $gl = n - 1 = 185$ é grande e a distribuição t é aproximadamente idêntica à normal padrão. Da Tabela A (página 650), a probabilidade em uma cauda acima de 0,68 é aproximadamente de 0,25, portanto, a probabilidade das duas caudas é aproximadamente $p = 2(0{,}25) = 0{,}50$.

Um cálculo mais preciso do valor-p com a distribuição t usando um *software* fornece um valor-$p = 0{,}4973545$. Arredonde tal valor, digamos, a 0,50, antes de relatá-lo. Relatar um valor-p com tantas casas decimais, como 0,4973545, parece ser mais preciso do



☑ **Figura 6.3** Cálculo do valor-p quando $t = 0{,}68$, para testar H_0: $\mu = \mu_0$ contra H_a: $\mu \neq \mu_0$. O valor-p é a soma das probabilidades das caudas de um resultado igual ou mais extremo do que o que foi observado.

que ele realmente é. Na prática, a distribuição amostral é somente *aproximadamente* igual à distribuição t, porque a distribuição da população não é exatamente normal como suposta na aplicação do teste t.

5. Conclusão

Finalmente, o estudo deverá interpretar o valor-p no contexto. Quanto menor o p, maior a evidência contra H_0 e em favor da H_a.

EXEMPLO 6.2 Conservadorismo e liberalismo político

Alguns comentaristas políticos têm observado que os cidadãos dos Estados Unidos estão cada vez mais conservadores, tanto que eles consideram a palavra *liberal* como uma palavra desprezível. Podemos estudar a ideologia política analisando a resposta a certos itens da PSG. Por exemplo, o levantamento de dados perguntou (com o item "POLVIEWS") onde você se enquadraria em uma escala de sete pontos sobre ideologia política variando de extremamente liberal, (1), a extremamente conservador, (7). A Tabela 6.2 mostra a escala e a distribuição das respostas para o levantamento realizado em 2006. Os resultados são mostrados separadamente para as três categorias da variável denominada "RACE" na PSG.

A ideologia política é uma escala ordinal. Geralmente, tratamos estas escalas de uma forma quantitativa atribuindo valores às categorias. Assim podemos usar resumos quantitativos, como as médias, para verificar a extensão na qual as observações gravitam na direção dos extremos da escala (conservador ou liberal).

☑ **Tabela 6.2** Resposta dos sujeitos entrevistados sobre ideologia política

	Raça		
Resposta	**Negro**	**Branco**	**Outra**
1. Extremamente liberal	10	36	1
2. Liberal	21	109	13
3. Um pouco liberal	22	124	13
4. Moderado (meio-termo)	74	421	27
5. Um pouco conservador	21	179	9
6. Conservador	27	176	7
7. Extremamente conservador	11	28	2
	$n = 186$	$n = 1073$	$n = 72$

Se atribuirmos escores às categorias exibidas na Tabela 6.2, então uma média abaixo de 4 mostra uma tendência em relação ao liberalismo, e uma média acima de 4 mostra a tendência em direção ao conservadorismo. Podemos testar se esses dados mostram evidências em ambos executando um teste de significância para ver como a média populacional se compara ao valor moderado de 4. Faremos isso aqui para a amostra negra e na Seção 6.5 para toda a amostra.

1. *Suposições:* a amostra é selecionada aleatoriamente. Estamos tratando a ideologia política como quantitativa com escores igualmente distribuídos. O teste t assume uma distribuição populacional normal para a ideologia política. Discutiremos essa suposição mais tarde no final desta seção.
2. *Hipóteses*: considere μ a representação da média populacional da ideologia para norte-americanos negros para esta escala de sete pontos. A hipótese nula contém um valor específico para μ. Visto que conduzimos a análise para verificar como, de alguma forma, a média populacional é diferente da resposta moderada 4, a hipótese nula é:

$$H_0: \mu = 4{,}0.$$

A hipótese alternativa é, então:

$$H_a: \mu \neq 4{,}0.$$

A hipótese nula afirma que, em média, a população é politicamente "moderada, meio-termo". A alternativa afirma que a média é liberal ($\mu < 4$) ou conservadora ($\mu > 4$).

3. *Estatística-teste:* as 186 observações na Tabela 6.2 para os negros está resumida por $\bar{y} = 4{,}075$ e $s = 1{,}512$. O erro padrão estimado da distribuição amostral de $\bar{y}$ é:

$$ep = \frac{s}{\sqrt{n}} = \frac{1{,}512}{\sqrt{186}} = 0{,}111.$$

O valor da estatística-teste é, então:

$$t = \frac{\bar{y} - \mu_0}{ep} = \frac{4{,}075 - 4{,}0}{0{,}111} = 0{,}68.$$

A média da amostra está 0,68 erros padrão estimados acima do valor da hipótese nula de uma média igual a 4. O valor do gl é $186 - 1 = 185$.

4. *O valor-*p: o valor-p é a soma das probabilidades das duas caudas, assumindo que a H_0 seja verdadeira, isto significa que t daria exceder 0,68 em valor absoluto. Da distribuição t com $gl = 185$ (ou sua aproximação normal padrão), essa probabilidade (soma das caudas) é $p = 0{,}50$. Se a ideologia política média da população fosse 4,0, então a probabilidade seria igual a 0,50 de que a média da amostra, para $n = 186$ sujeitos, estaria ao menos tão distante de 4,0 quanto o valor $\bar{y}$ observado de 4,075.
5. *Conclusão:* o valor-p de 0,50 não é pequeno, assim ele não contradiz H_0. Se H_0 fosse verdadeira, os dados observados não seriam incomuns. É plausível que a resposta média da população para norte-americanos negros em 2006 fosse 4,0, não sendo nem conservadora nem liberal. ■

Correspondência entre os testes bilaterais e os intervalos de confiança

As conclusões que usam testes de significância bilaterais são consistentes com conclusões que usam intervalos de confiança. Se um teste diz que um valor em particular é possível para o parâmetro, então assim também o faz o intervalo de confiança.

EXEMPLO 6.3 Intervalo de confiança para a média da ideologia política

Para os dados do Exemplo 6.2, vamos construir um intervalo de 95% de confiança para a ideologia política da média da população. Com $gl = 185$, o múltiplo do erro padrão ($ep = 0{,}111$) é $t_{0{,}025} = 1{,}97$. Visto que $\bar{y} = 4{,}075$, o intervalo de confiança é:

$$\bar{y} \pm 1{,}97(ep) = 4{,}075 \pm 1{,}97(0{,}111)$$
$$= 4{,}075 \pm 0{,}219, \text{ ou } (3{,}9; 4{,}3).$$

Ao nível de 95% de confiança, estes são os valores plausíveis para μ.

Esse intervalo de confiança indica que μ pode ser igual a 4,0 uma vez que 4,0 está dentro do intervalo de confiança. Portanto, não é surpresa que o valor-p ($p = 0{,}50$) testando H_0: $\mu = 4{,}0$ contra H_a: $\mu \neq 4{,}0$ no Exemplo 6.2 não foi pequeno. Na verdade:

- Sempre que o valor-$p > 0{,}05$ em um teste bilateral, um intervalo de 95% de confiança para μ necessariamente contém o valor H_0 de μ.

Em contraposição, suponha que o valor-$p = 0{,}02$ no teste da H_0: $\mu = 4{,}0$. Então, um intervalo de 95% de confiança nos diria que 4,0 não é plausível para μ, pois ele estaria *fora* do intervalo de confiança.

- Sempre que $p \leq 0{,}05$ em um teste bilateral, um intervalo de 95% de confiança para μ não conterá o valor H_0 de μ.

A Seção 6.4 discute mais a conexão entre os dois métodos. ■

Testes de significância unilaterais

Uma hipótese alternativa diferente é algumas vezes usada quando um pesquisador prevê um desvio da H_0 em uma direção em particular. Ela tem a forma:

$$H_a: \mu > \mu_0 \text{ ou } H_a: \mu < \mu_0.$$

A alternativa H_a: $\mu > \mu_0$ é usada para detectar se μ é *maior* do que o valor particular μ_0, ao passo que H_a: $\mu < \mu_0$ é usado para detectar se μ é *menor* do que esse valor. Essas hipóteses são chamadas de **unilaterais.** Em contraposição, a H_a *bilateral* é usada para detectar *qualquer* tipo de desvio de H_0. Esta escolha é feita antes da análise dos dados.

Para H_a: $\mu > \mu_0$, o valor-p é a probabilidade (assumindo que H_0 seja verdadeira) de se obter um valor-t *acima* do escore-t observado; isto é, à direita na linha dos números reais. Esses escores-t fornecem mais evidência extrema do que o valor observado em favor de H_a: $\mu > \mu_0$. Assim, p é igual à probabilidade da cauda direita sob a curva t, como a Figura 6.4 mostra. Um escore-t de 0,68 resulta em $p = 0{,}25$ para esta alternativa.

Para H_a: $\mu < \mu_0$, o valor-p é a probabilidade da cauda esquerda, (*abaixo*) do escore-t observado. Um escore-$t = -0{,}68$ resulta em $p = 0{,}25$ para esta alternativa. Um escore-$t = +0{,}68$ resulta em $p = 1 - 0{,}25 = 0{,}75$, nesse caso.

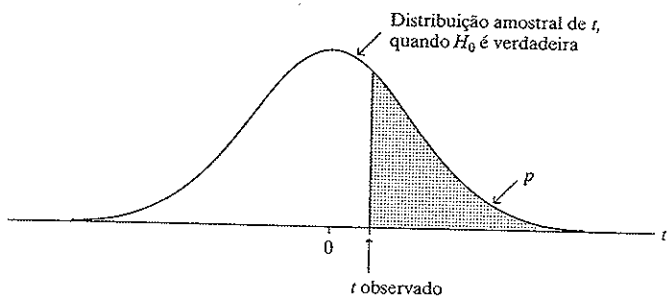
EXEMPLO 6.4 Mudança do peso médio em meninas anoréxicas

O Exemplo 5.5 do Capítulo 5 (página 144) analisou os dados de um estudo que comparava tratamentos para meninas adolescentes sofrendo de anorexia. Para cada menina, o estudo observou sua mudança de peso enquanto recebia o tratamento. Considere μ a representação da mudança média do peso da população para o tratamento cognitivo-comportamental. Se o tratamento tem efeito benéfico, como esperado, então μ é positivo. Para verificar a falta de efeito do tratamento *versus* um valor médio positivo para a mudança do peso, testamos H_0: $\mu = 0$ contra H_a: $\mu > 0$.

O *software* (SPSS) usado para analisar os relatórios dos dados relata:

Variável	Número de Casos	Média	DP	EP da Média
MUDANÇA	29	3,007	7,309	1,357

Para $n = 29$ meninas, a média amostral da alteração do peso foi 3,007 libras com um erro padrão estimado de $ep = 1{,}357$. A estatística-teste é, então, igual a:



☑ **Figura 6.4** Cálculo do valor-*p* no teste de H_0: $\mu = \mu_0$ contra H_a: $\mu > \mu_0$. O valor-*p* é a probabilidade dos valores à direita da estatística-teste observada.

$$t = \frac{\bar{y} - \mu_0}{ep} = \frac{3{,}007 - 0}{1{,}357} = 2{,}22.$$

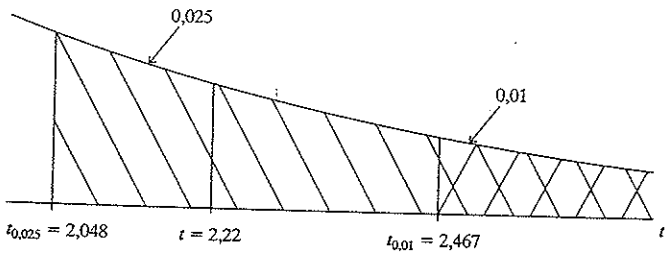
Para essa H_a observada, o valor-*p* é a probabilidade da cauda direita acima de 2,22. Por que usamos a cauda direita? Porque H_a: $\mu > 0$ tem valores *acima* (isto é, à direita) do valor da hipótese nula. São os valores positivos de *t* que suportam esta hipótese alternativa.

Agora, para $n = 29$, $gl = n - 1 = 28$. Da Tabela B (p. 651), $t = 2{,}048$ gera $p = 0{,}025$ para uma H_a unilateral e $t = 2{,}467$ gera $p = 0{,}01$. O *t* observado = 2,22 está entre 2,048 e 2,467, assim o valor-*p* está entre 0,01 e 0,025. A Figura 6.5 ilustra esse fato. A Tabela B não é detalhada o suficiente para fornecer o valor-*p* exato. Quando um *software* executa uma análise, a saída apresenta o valor-*p* real em vez de aproximado. Muitos *softwares* fornecem o valor-*p* para uma alternativa bilateral como padrão a não ser que você solicite de outra forma. O SPSS apresenta os resultados para o teste bilateral e o intervalo de confiança como:

Média	IC 95%				
	Inferior	Superior	valor t	gl	Sig. bilateral
3,01	0,227	5,787	2,22	28	0,035

O SPSS relata $p = 0{,}035$ (sob "Sig. bilateral"). O valor-*p* unilateral é metade disto ou aproximadamente $p = 0{,}02$. Existe uma evidência relativamente forte contra H_0. Parece que o tratamento tem um efeito.

O teste de significância conclui que o ganho médio de peso não foi igual a 0. Mas o intervalo de 95% de confiança de (0,2; 5,8), mostrado também na saída do SPSS, é mais informativo. Ele mostra apenas quão diferente de 0 a mudança de peso médio da população é provável de



☑ **Figura 6.5** Para $gl = 28$, $t = 2{,}22$ tem uma probabilidade unicaudal situada entre 0,01 e 0,025.

ser. O efeito poderia ser muito pequeno. Também, tenha em mente que esse estudo experimental (como muito estudos com orientação médica) utilizou uma amostra voluntária. Portanto, esses resultados são altamente experimentais, outra razão do por que é simplório para estudos como esse apresentar valores-*p* com várias casas decimais. ■

H_0 unilateral implícita para uma H_a unilateral

Do Exemplo 6.4, o valor-*p* unilateral = 0,017. Assim, se $\mu = 0$, a probabilidade é igual a 0,017 de observar uma média amostral de ganho de peso de 3,01 ou maior. Agora, suponha que $\mu < 0$; isto é, a mudança média do ganho de peso da população é negativa. Então, a probabilidade de observar $\bar{y} \geq 3{,}01$ seria ainda menor do que 0,017. Por exemplo, um valor amostral de $\bar{y} = 3{,}01$ é ainda menos provável quando $\mu = -5$ do que quando $\mu = 0$, visto que 3,01 está bem mais longe na cauda da distribuição amostral de $\bar{y}$ quando $\mu = -5$ do que quando $\mu = 0$. Portanto, a rejeição de H_0: $\mu = 0$ em favor de H_a: $\mu > 0$ também, basicamente, rejeita um leque maior de valores para a hipótese nula, ou seja, H_0: $\mu \leq 0$. Em outras palavras, podemos concluir que $\mu = 0$ é falso e também que $\mu < 0$ é igualmente falso.

A escolha dos testes unilaterais e bilaterais

Na prática, os testes bilaterais são mais comuns do que os testes unilaterais. Mesmo se um pesquisador prevê a direção de um efeito, os testes bilaterais podem também detectar um efeito que está na direção oposta. Na maioria dos artigos de pesquisa, os testes de significância usam valores-*p* bilaterais. Parte disto reflete uma abordagem objetiva à pesquisa que reconhece que um efeito poderia ir a ambas as direções. Usando os valores-*p* bilaterais, os pesquisadores evitam a suspeita de que eles escolheram H_a quando viram a direção na qual os dados ocorreram. Isso não seria ético.

Os testes bilaterais coincidem com a abordagem usual da estimação. Os intervalos de confiança são bilaterais, obtidos pela adição e subtração de algum valor da estimativa por ponto. Qualquer um pode formar intervalos de confiança unilaterais; por exemplo, concluindo que a média da população é *pelo menos* igual a 7 (isto é, entre 7 e ∞). Na prática, contudo, intervalos unilaterais são raramente usados.

Na decisão de usar uma H_a unilateral ou bilateral em um exercício em particular ou na prática, considere o contexto. Um exercício que diz: "Teste se a média *mudou*" sugere uma alternativa bilateral, para permitir aumento ou diminuição. "Teste se a média *aumentou*" sugere uma H_a: $\mu > \mu_0$ unilateral.

Tanto no caso unilateral quanto bilateral, as hipóteses sempre se referem aos parâmetros da população, não a estatísticas amostrais. Assim, *nunca* expresse uma hipótese usando uma notação estatística amostral, como H_0: $\bar{y} = 0$. Não existe incerteza ou necessidade de conduzir uma inferência estatística sobre a estatística amostral como $\bar{y}$, porque podemos calcular seus valores exatamente dos dados.

O nível α: usando o valor-*p* para tomar uma decisão

Um teste de significância analisa a força da evidência contra a hipótese nula, H_0. Iniciamos presumindo que H_0 é verdadeira. Analisamos se os dados seriam incomuns se H_0 fosse verdadeira encontrando o valor-*p*. Se o valor-*p* for pequeno, os dados contradizem H_0 e apoiam H_a. Geralmente, os pesquisadores não consideram a evidência contra H_0 tão forte a não ser que *p* seja muito pequeno, digamos $p < 0{,}05$ ou $p < 0{,}01$.

Por que valores-p menores indicam uma evidência mais forte contra H_0? Porque os dados seriam, então, mais incomuns se H_0 fosse verdadeira. Quando H_0 é verdadeira, o valor-p é mais provável de estar próximo de 1. Em contraposição, quando H_0 é falsa, o valor-p é mais provável de estar próximo a 0 do que próximo a 1.

Na prática, é algumas vezes necessário decidir se a evidência contra H_0 é forte o suficiente para rejeitá-la. A decisão é baseada em verificar se o valor-p está abaixo de um ponto de corte pré-especificado. É mais comum rejeitar H_0 se $p \leq 0{,}05$ e concluir que a evidência não é forte o suficiente para rejeitar H_0 se $p > 0{,}05$. O valor limite 0,05 é chamado de **nível** α do teste.

☑ **Nível** α

O **nível** α é um número tal que rejeitamos H_0 se o valor-p é menor ou igual a ele. O nível α é também chamado de **nível de significância**. Na prática, os níveis α aceitos mais comuns são 0,05 e 0,01.

Como a escolha de um nível de confiança para um intervalo de confiança, a escolha de α reflete quão cuidadoso você quer ser. Quanto menor o nível α, mais forte a evidência deve ser para rejeitar H_0. Para evitar tendenciosidade no processo de tomada de decisão, você deve selecionar α *antes* de analisar os dados.

EXEMPLO 6.5 Acrescentando decisões aos exemplos anteriores

Vamos usar $\alpha = 0{,}05$ para nos guiar na tomada de decisão sobre H_0 para os exemplos desta seção. O Exemplo 6.2 testou H_0: $\mu = 4{,}0$ para a média da ideologia política. Com a média amostral $\bar{y} = 4{,}075$, o valor-p foi de 0,50. O valor-p não é pequeno, assim se verdadeiro for $\mu = 4{,}0$, não será incomum observar uma média amostral como $\bar{y} = 4{,}075$. Visto que $p = 0{,}50 > 0{,}05$, não existe evidência o suficiente para rejeitar H_0. É possível que a ideologia média da população seja 4,0*.

O Exemplo 6.4 testou H_0: $\mu = 0$ como o ganho médio de peso para meninas adolescentes sofrendo de anorexia. O valor-p foi de 0,017. Visto que $p = 0{,}017 < 0{,}05$, existe evidência o suficiente para rejeitar a H_0 em favor da H_a: $\mu > 0$. Concluímos, assim, que o tratamento resulta em um aumento no peso médio. Tal conclusão é, algumas vezes, redigida como "O aumento no peso médio é *estatisticamente significativo* no nível 0,05". Visto que $p = 0{,}017$ *não* é menor do que 0,010, o resultado *não* é significativo ao nível de 0,010. Na verdade, *o valor-*p *é o menor nível de* α *para o qual os resultados são significativos*. Portanto, com o valor-$p = 0{,}017$, rejeitamos H_0 se $\alpha = 0{,}02$ ou 0,05 ou 0,10, mas não se $\alpha = 0{,}015$ ou 0,010 ou 0,001. ■

A Tabela 6.3 resume os testes de significância para médias populacionais.

Robustez para violações da suposição de normalidade

O teste t para uma média supõe que a distribuição populacional é normal. Isso assegura que a distribuição amostral da média amostral $\bar{y}$ é normal (mesmo para n pequeno) e após usarmos s para estimar σ para encontrar o ep, a estatística-teste t tem uma distribuição t. À medida que o tamanho da amostra aumenta, a suposição de normalidade se torna menos importante. Vimos que, quando n é aproximadamente 30 ou mais, uma distribuição amostral normal aproximada ocorre para $\bar{y}$ independentemente da distribuição da população segundo o Teorema Central do Limite (Seção 4.5, na página 110).

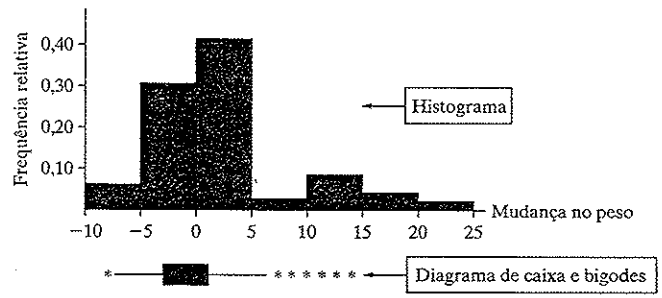
* N. de T. T.: Note-se que o autor está dizendo que é possível a média ser 4 e não que estamos provando que ela é 4. De fato, quando se aceita H_0, não provamos que ela é verdadeira, mas apenas que não foi possível rejeitá-la, isso porque existem outras infinitas possibilidades (em torno de 4) que poderiam ser igualmente verdadeiras.

☑ **Tabela 6.3** As cinco etapas dos testes de significância para médias populacionais

1. **Suposições**
   Variável quantitativa
   Aleatorização
   População normal (robusta, especialmente para H_a bilateral e n grande)
2. **Hipóteses**
   H_0: $\mu = \mu_0$
   H_a: $\mu \neq \mu_0$ (ou H_a: $\mu > \mu_0$ ou H_a: $\mu < \mu_0$)
3. **Estatística-teste**
   $t = \dfrac{\bar{y} - \mu_0}{se}$ onde $se = \dfrac{s}{\sqrt{n}}$
4. **Valor-*p***
   Na curva t utilize:
   p = probabilidade bilateral para H_a: $\mu \neq \mu_0$
   p = probabilidade à direita do valor observado t para H_a: $\mu > \mu_0$
   p = probabilidade à esquerda do valor observado t para H_a: $\mu < \mu_0$
5. **Conclusão**
   Relate o valor-p. Valores p pequenos fornecem evidências mais fortes contra H_0 e suportam H_a. Pode rejeitar H_0 se $p \leq \alpha$.

Da Seção 5.3 (página 140), um método estatístico é **robusto** se ele tem um desempenho adequado mesmo quando uma suposição é violada. Os estatísticos têm mostrado que as inferências *bilaterais* para uma média usando a distribuição t são robustas contra as violações da suposição de população normal. Mesmo se a população não for normal, os testes t bilaterais e intervalos de confiança ainda funcionam muito bem. O teste não funciona tão bem para uma hipótese unilateral com um n pequeno quando a distribuição da população for altamente assimétrica.

A Figura 6.6 mostra um histograma e um diagrama de caixa e bigodes dos dados do estudo da anorexia do Exemplo 6.4 (página 177). A Figura 6.6 sugere que existe uma assimetria à direita. O diagrama de caixa e bigodes destaca (como valores atípicos) seis meninas que tiveram ganhos consideráveis de peso. Como acabou de ser mencionado, um teste t bilateral funciona muito bem se a distribuição da população é assimétrica. Entretanto, este diagrama nos torna cautelosos sobre o uso do teste unilateral, visto que o tamanho da amostra não é grande ($n = 29$).



☑ **Figura 6.6** Histograma e diagrama de caixa e bigodes da mudança de peso para vítimas de anorexia.

Considerando isso e a discussão na subseção anterior sobre testes unilaterais *versus* testes bilaterais, estamos mais seguros com aquele estudo para relatar um valor-p bilateral de 0,035. Também, como o Exemplo 5.5 (página 144) observou, a mediana pode ser uma medida mais relevante para esses dados.

6.3 TESTE DE SIGNIFICÂNCIA PARA UMA PROPORÇÃO

Para uma variável categórica, o parâmetro é a proporção da população em uma categoria. Por exemplo, um teste de significância poderia analisar se a maioria da população apoia o uso de células-tronco embrionárias testando H_0: $\pi = 0{,}50$ contra H_a: $\pi > 0{,}50$, onde π é a proporção da população que apoia o uso das células-tronco embrionárias. O teste para uma proporção é como o teste para uma média, encontra o valor-p para uma estatística-teste mensurando o número de erros padrão que uma estimativa por ponto está do valor H_0.

As cinco etapas de um teste de significância para uma proporção

1. Suposições

Como os outros testes, esse assume que os dados são obtidos utilizando uma amostra aleatória. O tamanho da amostra deve ser suficientemente grande para que a distribuição amostral de $\hat{\pi}$ seja aproximadamente normal. Para o caso mais comum, no qual o valor H_0 de π é 0,50, um tamanho da amostra de pelo menos 20 é suficiente. Daremos uma orientação precisa na Seção 6.7, que apresenta um pequeno teste amostral.

2. Hipóteses

A hipótese nula de um teste sobre a proporção da população tem a forma:

$$H_0\text{: } \pi = \pi_0, \text{ tal como } H_0\text{: } \pi = 0{,}50.$$

Aqui, π_0 representa um valor particular da proporção entre 0 e 1, como 0,50. A hipótese alternativa mais comum é:

$$H_a\text{: } \pi \neq \pi_0, \text{ tal como } H_a\text{: } \pi \neq 0{,}50.$$

Esta alternativa *bilateral* declara que a proporção da população difere do valor em H_0. As alternativas *unilaterais*

$$H_a\text{: } \pi > \pi_0 \text{ e } H_a\text{: } \pi < \pi_0$$

se aplicam quando o pesquisador prevê um desvio em uma certa direção do valor H_0.

3. Estatística-teste

Da Seção 5.2 (página 134), vimos que a distribuição amostral de $\hat{\pi}$ tem média π e o erro padrão $\sigma_{\hat{\pi}} = \sqrt{\pi(1-\pi)/n}$. Quando H_0 é verdadeira, $\pi = \pi_0$, assim o erro padrão é $ep_0 = \sqrt{\pi_0(1-\pi_0)/n}$. Usamos a representação ep_0 para indicar que esse é o erro padrão sob a suposição de que H_0 é verdadeira.

A estatística-teste é:

$$z = \frac{\hat{\pi} - \pi_0}{ep_0},$$

$$\text{onde } ep_0 = \sqrt{\pi_0(1-\pi_0)/n}.$$

Isso avalia o número de erros padrão que a proporção amostral $\hat{\pi}$ está de π_0. Para amostras grandes, se H_0 é verdadeira, a estatística-teste z tem uma distribuição normal padrão.

☑ **Forma da estatística-teste**

$$z = \frac{\text{Estimativa do parâmetro} - \text{valor do parâmetro da hipótese nula}}{\text{Erro padrão do estimador}}$$

Aqui, a estimativa $\hat{\pi}$ da proporção substitui a estimativa $\bar{y}$ da média e a proporção da hipótese nula π_0 substitui a média da hipótese nula μ_0.

4. Valor-p

O valor-p é a probabilidade de uma ou duas caudas, como nos testes para uma média, com exceção, de que usamos a distribuição normal, em vez da distribuição t. Para H_a: $\pi \neq \pi_0$, p é a probabilidade de duas caudas. Veja a Figura 6.7. Esta probabilidade é o dobro da probabilidade de uma cauda além do valor-z observado.

Para alternativas unilaterais, o valor-p é uma probabilidade unilateral. Visto que H_a: $\pi > \pi_0$ prevê que a proporção da população é maior do que o valor H_0, seu valor-p é a probabilidade *acima* (isto é, à direita) do valor-z observado. Para H_a: $\pi < \pi_0$, o valor-p é a probabilidade *abaixo* (isto é, à esquerda) do valor-z observado.

5. Conclusão

Como de costume, quanto menor for o valor-p mais fortemente os dados contradizem H_0 e suportam H_a. Quando precisarmos tomar uma decisão, rejeitamos H_0 se $p \leq \alpha$ para um nível α especificado como 0,05.

EXEMPLO 6.6 Reduzir serviços ou aumentar as taxas?

Nos dias de hoje, tanto a nível local, estadual ou nacional, o governo geralmente se depara com o problema de não ter dinheiro suficiente para pagar vários serviços que fornece. Uma forma de tratar desse problema é aumentar as taxas. Outra forma é reduzir os serviços. Qual delas você iria preferir? Quando a Florida Poll[3] perguntou a uma amostra aleatória de 1200 residentes em 2006, 52% disseram aumento das taxas e 48% disseram redução dos serviços.

Considere π a representação da proporção da população da Flórida que iria escolher o aumento das taxas. Se $\pi < 0{,}50$, isto é a minoria da população, enquanto $\pi > 0{,}50$ é a sua maioria. Para analisar se π está em um destes intervalos, testamos H_0: $\pi = 0{,}50$ contra H_a: $\pi \neq 0{,}50$.

A estimativa de π é $\hat{\pi} = 0{,}52$. Assumindo que H_0: $\pi = 0{,}50$ é verdadeira, o erro padrão de $\hat{\pi}$ é:

$$ep_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

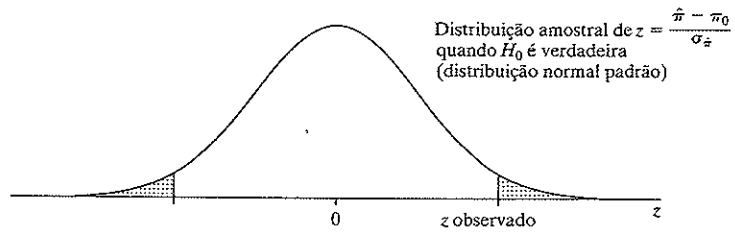
$$= \sqrt{\frac{(0{,}50)(0{,}50)}{1200}} = 0{,}0144.$$

O valor da estatística-teste é:

$$z = \frac{\hat{\pi} - \pi_0}{ep_0} = \frac{0{,}52 - 0{,}50}{0{,}0144} = 1{,}39.$$

Da Tabela A, o valor-p bilateral é $2(0{,}0823) = 0{,}16$. Se H_0 é verdadeira (isto é, se $\pi = 0{,}50$), a probabilidade é igual a 0,16 de que os resultados da amostra fossem tão extremos, em uma ou outra direção, como nesta amostra.

Este valor-p não é pequeno, assim não existe muita evidência contra H_0. Parece possível que $\pi = 0{,}50$. Com um nível α como 0,05, visto que $p = 0{,}16 > 0{,}05$, não iríamos rejeitar H_0. Não podemos



☑ **Figura 6.7** Cálculo do valor-p testando H_0: $\pi = \pi_0$ contra H_a: $\pi \neq \pi_0$. A hipótese alternativa bilateral usa a probabilidade das duas caudas.

determinar se aqueles a favor do aumento das taxas são a maioria ou minoria da população. ■

Na fórmula do erro padrão, $\sqrt{\pi(1-\pi)/n}$, observe que substituímos o valor da hipótese nula $\pi_0 = 0{,}50$ para a proporção da população π. Os valores do parâmetro das distribuições amostrais dos testes presumem que H_0 é verdadeira, visto que o valor-p tem por base essa suposição. Isto é porque, para os testes, usamos $ep_0 = \sqrt{\pi_0(1-\pi_0)/n}$ em vez do erro padrão estimado, $ep = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$. Com o ep estimado, a aproximação normal para a distribuição amostral de z é menos satisfatória. Isto é especialmente verdadeiro para proporções próximas a 0 e 1. A validade do valor-p é, então, insatisfatória. Em contraposição, o método do intervalo de confiança não tem um valor hipotético para π, portanto ele usa o ep estimado em vez do valor H_0.

Nunca "aceite H_0"

No Exemplo 6.6, sobre o aumento de taxas ou redução dos serviços, o valor-p de 0,16 não era pequeno. Assim H_0: $\pi = 0{,}50$ é aceitável. Neste caso, a conclusão é, algumas vezes, relatada como: "não rejeite H_0", visto que os dados não contradizem H_0.

Dizemos "não rejeite H_0" em vez de "aceite H_0". A proporção populacional tem muitos valores aceitáveis além do valor da H_0. Por exemplo, um intervalo de 95% de confiança para a proporção da população π que apoia o aumento das taxas em vez da redução dos serviços é:

$$\hat{\pi} \pm 1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0{,}52 \pm 1{,}96\sqrt{\frac{(0{,}52)(0{,}48)}{1200}} \text{ ou } (0{,}49;\ 0{,}55)$$

Este intervalo mostra um conjunto de valores plausíveis para π. Embora não exista evidência suficiente para concluir que $\pi \neq 0{,}50$, não é adequado concluir que $\pi = 0{,}50$.

Em resumo, H_0 contém um único valor para o parâmetro. Quando o valor-p é maior do que o nível α, dizendo "não rejeite H_0" em vez de "aceite H_0", enfatiza-se que aquele valor é meramente um dos muitos valores aceitáveis. Por causa da variabilidade amostral, existe um intervalo de valores aceitáveis, assim nunca podemos, de fato, aceitar H_0. A razão pela qual a terminologia "aceite H_a" é permitida para H_a é que quando o valor-p é suficientemente pequeno, todo o conjunto de valores possíveis para o parâmetro está dentro do intervalo de valores que H_a especifica.

O efeito do tamanho da amostra nos valores-p

No Exemplo 6.6, do aumento das taxas ou o corte dos serviços, suponha que $\hat{\pi} = 0{,}52$ foi baseado em $n = 4800$ em vez de $n = 1200$. O erro padrão diminui para 0,0072 (metade do tamanho) e você pode verificar que a estatística-teste é $z = 2{,}77$. Isto tem um valor-p bilateral igual a 0,006. Esse valor-p fornece uma forte evidência contra H_0: $\pi = 0{,}50$ e sugere que a maioria apoia o aumento das taxas em vez do corte nos serviços. Neste caso, porém, o intervalo de 95% de confiança para π é igual a (0,506; 0,534). Isto indica que π está bem próximo a 0,50, em termos práticos.

Uma dada diferença entre uma estimativa e o valor de H_0 tem um valor-p menor à medida que o tamanho da amostra aumenta. Quanto maior o tamanho da amostra, maior é a probabilidade de que os desvios amostrais de H_0 indiquem desvios na população. Em particular, observe que mesmo uma pequena diferença entre $\hat{\pi}$ e π_0 (ou entre $\bar{y}$ e μ_0) pode gerar um valor-p pequeno se o tamanho da amostra for muito grande.

6.4 DECISÕES E TIPOS DE ERROS EM TESTES

Quando precisamos decidir se a evidência contra H_0 é forte o suficiente para rejeitá-la, vimos que H_0 é rejeitada se $p \leq \alpha$, para um nível α pré-especificado. A Tabela 6.4 resume as duas conclusões possíveis para o nível $\alpha = 0{,}05$. A hipótese nula é tanto *rejeitada* ou *não rejeitada*. Se H_0 for rejeitada, então H_a é aceita. Se H_0 não for rejeitada, então H_0 é aceitável, mas outros valores dos parâmetros são também possíveis. Portanto, H_0 nunca é *aceita*. Neste caso, os resultados são inconclusivos e o teste não identifica nenhuma das hipóteses como a mais válida.

É melhor relatar o valor-p do que indicar simplesmente se o resultado é "estatisticamente significativo". Relatar o valor-p tem a vantagem de que o leitor pode dizer se o resultado é significativo em qualquer nível. Os valores-p de 0,049 e 0,001 são ambos "significativos no nível 0,05", mas o segundo caso fornece evidências mais fortes do que o primeiro caso. Da mesma forma, os valores-p de 0,049 e 0,051 fornecem, em termos práticos, o mesmo grau de evidência sobre H_0. É um pouco artificial chamar um resultado de "significativo" e o outro de "não significativo".

Erros do Tipo I e do Tipo II nas decisões

Por causa da variabilidade amostral, as decisões em testes sempre têm alguma incerteza. A decisão poderia estar errada. Existem dois tipos de erros potenciais, convencionalmente chamados de erros do *Tipo I* e do *Tipo II*.

☑ Erros do Tipo I e do Tipo II

Quando H_0 é verdadeira, ocorre um ***erro do Tipo I*** se H_0 for rejeitada.
Quando H_0 é falsa, ocorre um ***erro do Tipo II*** se H_0 não for rejeitada.

Existem quatro resultados possíveis. Eles se referem às duas decisões possíveis de classificação cruzada com as duas possibilidades de H_0 ser ou não verdadeira. Veja a Tabela 6.5.

☑ **Tabela 6.4** Decisões possíveis em um teste de significância com nível $\alpha = 0{,}05$

	Decisão	
Valor-p	H_0	H_a
$p \leq 0{,}05$	Rejeitar	Aceitar
$p > 0{,}05$	Não rejeitar	Não aceitar

☑ **Tabela 6.5** Os quatro resultados possíveis da tomada de decisão em um teste de hipóteses. Os erros do Tipo I e do Tipo II são as duas decisões incorretas possíveis

	Decisão	
Realidade	***Rejeitar H_0***	***Não rejeitar H_0***
H_0 é verdadeira	Erro do Tipo I	Decisão correta
H_0 é falsa	Decisão correta	Erro do Tipo II

Regiões de rejeição

A coleção dos valores da estatística-teste para a qual o teste rejeita H_0 é chamada de **região de rejeição**. Por exemplo, a região de rejeição para um teste de nível $\alpha = 0{,}05$ é o conjunto dos valores da estatística-teste para o qual $p \leq 0{,}05$.

Para testes bilaterais sobre uma proporção, o valor-p bilateral será menor ou igual a 0,05 sempre que a estatística-teste $|z| \geq 1{,}96$. Em outras palavras, a região da rejeição consiste em valores de z resultantes de a estimativa estar a, pelo menos, 1,96 erros padrão do valor H_0.

O nível α é a probabilidade do erro do Tipo I

Quando H_0 for verdadeira, vamos encontrar a probabilidade do erro do Tipo I. Suponha que $\alpha = 0{,}05$. Acabamos de ver que, para o teste bilateral sobre uma proporção, a região de rejeição é $|z| \geq 1{,}96$. Assim, a probabilidade de rejeitar H_0 é exatamente 0,05 porque a probabilidade dos valores nessa região sob a curva da normal padrão é 0,05. Mas isto é precisamente o nível α.

☑ A probabilidade de um erro do Tipo I é o nível α para o teste.

Com $\alpha = 0{,}05$, se H_0 for verdadeira, a probabilidade é igual a 0,05 de cometer um erro do Tipo I e rejeitar aquela H_0 (verdadeira). Controlamos o erro do Tipo I pela escolha de α. Quanto mais sérias as consequências de um erro do Tipo I, menor deveria ser α. Na prática, $\alpha = 0{,}05$ é mais comum, como a probabilidade de um erro de 0,05 é mais comum com intervalos de confiança (isto é, 95% de confiança). Contudo, isso pode ser muito alto quando a decisão tem sérias implicações.

Por exemplo, considere um processo criminal legal de um acusado. Considere H_0 a representação de inocente e H_a a representação de culpado. O júri rejeita H_0 e julga o acusado culpado se decidir que a evidência é suficiente para condenar. Um erro do Tipo I, rejeitando H_0 verdadeira, ocorre na condenação de um acusado que é realmente inocente. Em um processo por homicídio, suponha que um acusado condenado tenha a pena de morte. Então, se o acusado for realmente inocente, esperaríamos que a probabilidade de condenação fosse muito menor do que 0,05.

Quando tomamos uma decisão, não sabemos se cometemos um erro do Tipo I ou do Tipo II, assim como não sabemos se um intervalo de confiança em particular verdadeiramente contém o valor do parâmetro. Entretanto, podemos controlar a probabilidade de uma decisão incorreta para ambos os tipos de inferência.

À medida que a P(erro do Tipo I) diminui, a P(erro do Tipo II) aumenta

Em um mundo ideal, os erros do Tipo I e do Tipo II não iriam ocorrer. Na prática, erros acontecem. Já lemos sobre acusados que foram condenados, mas mais tarde foi verificado que eram inocentes. Quando tomamos uma decisão, por que não usamos uma P(erro do Tipo I) extremamente pequena, tal como $\alpha = 0{,}000001$? Por exemplo, por que não tornamos quase impossível condenar alguém que é realmente inocente?

Quando tornamos um α menor em um teste de significância, precisamos de um valor-p menor para rejeitar H_0. Então, se torna mais difícil rejeitar H_0. Mas isso também significa que será mais difícil até mesmo se H_0 for falsa. Quanto mais forte a evidência requerida para condenar alguém, mais provavelmente iremos falhar na condenação de acusados que realmente são culpados. Em outras palavras, quanto menor nós tornamos P(erro do Tipo I) maior se torna a P(erro do Tipo II); isto é, a probabilidade de falhar na rejeição de H_0 mesmo que ela seja falsa.

Se tolerássemos somente uma P(erro do Tipo I) extremamente pequena, como $\alpha = 0{,}000001$, pode ser improvável de que o teste rejeite H_0 mesmo se for falso – por exemplo, poderia ser improvável condenar alguém mesmo se ele ou ela for culpado(a). Este raciocínio reflete a relação fundamental:

- Quanto menor a P(erro do Tipo I), maior é a P(erro do Tipo II).

A Seção 6.6 mostra que a P(erro do Tipo II) depende somente de quão longe o parâmetro verdadeiro está de H_0. Se o parâmetro for aproximadamente igual ao valor em H_0 a P(erro do Tipo II) é relativamente alta. Se ele estiver longe de H_0, então a P(erro do Tipo II) é relativamente baixa. Quanto mais longe o parâmetro está do valor H_0, menos provável é que a amostra nos levará a cometer um erro do Tipo II.

Para uma P(erro do Tipo I) fixa, a P(erro do Tipo II) depende também do tamanho da amostra n. Quanto maior o tamanho da amostra, mais provável é que iremos rejeitar um H_0 falso. Para manter tanto a P(erro do Tipo I) quanto a P(erro do Tipo II) a níveis baixos, poderá ser necessário usar um tamanho da amostra muito grande. A P(erro do Tipo II) pode ser bem grande quando o tamanho da amostra é pequeno, a não ser que o parâmetro esteja bem longe do valor de H_0.

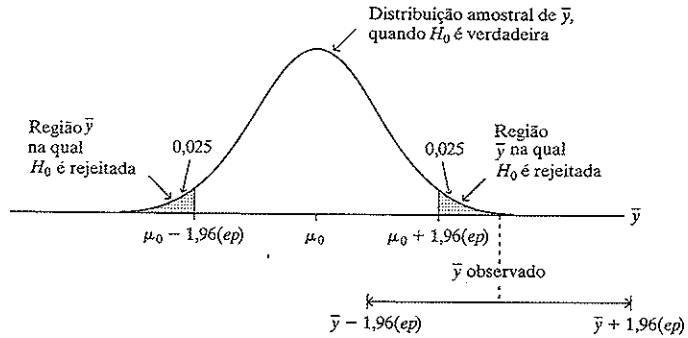
Com exceção da Seção 6.6, não devemos calcular P(erro do Tipo II) porque tais cálculos são complexos. Na prática, tomar uma decisão requer estabelecer somente α, a P(erro do Tipo I).

Equivalência entre intervalos de confiança e decisões de testes

Agora, elaboraremos a equivalência entre decisões de testes bilaterais e conclusões dos intervalos de confiança, mencionados inicialmente no Exemplo 6.3 (página 176). Considere o teste com um grande tamanho amostral:

$$H_0: \mu = \mu_0 \quad versus \quad H_a: \mu \neq \mu_0.$$

Quando $p < 0{,}05$, H_0 é rejeitada ao nível $\alpha = 0{,}05$. Isto acontece quando a estatística-teste $t = (\bar{y} - \mu_0)/ep$ é maior do que aproximadamente 1,96 em valor absoluto (quando n for grande), o que significa que $\bar{y}$ está a mais do que $1{,}96(ep)$ de μ_0. Mas se isto acontece, então, o intervalo de 95% de confiança para μ, a saber, $\bar{y} \pm 1{,}96(ep)$, não contém o valor da hipótese nula μ_0. Veja a Figura 6.8. Esses dois procedimentos de inferência são consistentes.



☑ **Figura 6.8** Relacionamento entre o intervalo de confiança e o teste de significância. O intervalo de 95% de confiança não contém o valor da H_0 (μ_0) quando a média da amostra está a mais do que 1,96 erros padrão de μ_0, neste caso a estatística-teste é $|z| > 1{,}96$ e o valor-p é menor do que 0,05.

☑ Testando H_0: $\mu = \mu_0$ contra H_a: $\mu \neq \mu_0$, suponha que rejeitamos H_0 no nível α 0,05. Então, o intervalo de 95% de confiança para μ não contém μ_0. O intervalo de 95% de confiança para μ consiste naqueles valores de μ_0 para os quais não rejeitamos H_0: $\mu = \mu_0$ no nível α 0,05.

No Exemplo 6.2, sobre a ideologia política, o valor-p para testar H_0: $\mu = 4{,}0$ contra H_a: $\mu \neq 4{,}0$ era $p = 0{,}50$. ao nível $\alpha = 0{,}05$, não rejeitamos H_0: $\mu = 4{,}0$. É aceitável que $\mu = 4{,}0$. O Exemplo 6.3 mostrou que um intervalo de 95% de confiança para μ é (3,9; 4,3), que contém $\mu_0 = 4{,}0$.

Rejeitar H_0 a um nível α particular é equivalente ao intervalo de confiança para μ com a mesma probabilidade de erro de não conter μ_0. Por exemplo, se um intervalo de 99% de confiança não contém 0, então rejeitaríamos H_0: $\mu = 0$ em favor de H_a: $\mu \neq 0$ no nível $\alpha = 0{,}01$ para o teste. O nível α é a P(erro do Tipo I) tanto para o teste quanto para a probabilidade de que o método do intervalo de confiança não contenha o parâmetro.

Tomar decisões *versus* relatar o valor-*p*

A abordagem formal para o teste de hipóteses que esta seção discutiu foi desenvolvida pelos estatísticos Jerzy Neyman e Egon Pearson no final da década de 1920 e início da de 1930. Em resumo, esta abordagem formula as hipóteses nula e alternativa, seleciona um nível α para a P(erro do Tipo I), determina a região de rejeição dos valores da estatística-teste e fornece evidências suficientes para rejeitar H_0 e, então, toma uma decisão sobre rejeitar H_0 de acordo com o que é realmente observado para o valor da estatística-teste. Com esta abordagem, não é nem mesmo necessário encontrar o valor-p. A escolha do nível α determina a região de rejeição, que junto com a estatística-teste determina a decisão.

A abordagem alternativa de encontrar um valor-p e usá-lo para resumir a evidência contra uma hipótese é devida ao grande estatístico britânico R. A. Fisher. Ele defendia somente relatar o valor-p em vez de usá-lo para tomar uma decisão formal sobre H_0. Ao longo do tempo, esta abordagem foi preferida, especialmente desde que o *software* pode, agora, relatar valores-p precisos para uma grande variedade de testes de significância.

Este capítulo apresentou uma combinação das duas abordagens (a abordagem baseada na decisão usando um nível α e a abordagem do valor-p), assim você pode interpretar um valor-p, mas também sabe como usá-lo para tomar uma decisão, se for necessário. Nos dias de hoje, muitos artigos de pesquisa simplesmente relatam o valor-p em vez de fornecer uma decisão sobre a rejeição ou não de H_0. Do valor-p, os leitores podem observar a força da evidência contra H_0 e tomar a sua própria decisão, se necessário.

6.5 LIMITAÇÕES DOS TESTES DE SIGNIFICÂNCIA

Um teste de significância faz inferência sobre se um parâmetro difere do valor H_0 e sobre sua direção. Na prática, também queremos saber se o parâmetro é suficientemente diferente do valor H_0 para ser de fato importante. A seguir, veremos que um teste não nos diz tanto quanto um intervalo de confiança sobre a importância prática.

Significância estatística *versus* significância prática

É importante distinguir entre *significância estatística* e *significância prática*. Um valor-p pequeno, como $p = 0{,}001$, tem um alto grau de significância estatística. Ele fornece forte evidência contra H_0. Ele não indica, entretanto, uma descoberta *importante* em qualquer sentido prático. O valor-p pequeno simplesmente significa que, se H_0 fosse verdadeira, os dados observados seriam muito incomuns. Isso não significa que o valor do parâmetro verdadeiro está longe de H_0 em termos práticos.

EXEMPLO 6.7 Ideologia política média para todos os norte-americanos

A ideologia política média de 4,08 no Exemplo 6.2 se refere a uma amostra de norte-americanos negros. A tabela também mostrou resultados para *brancos* e *outras* categorias. Para um escore de 1,0 a 7,0 para as categorias da ideologia com 4,0 = moderado, toda a amostra de 1331 observações tem uma média de 4,12 e um desvio padrão de 1,38. Parece que, em média, o conservadorismo era somente um pouco maior para a amostra combinada do que somente para os negros (4,12 *versus* 4,08).

Como no Exemplo 6.2, testamos H_0: $\mu = 4{,}0$ contra H_a: $\mu \neq 4{,}0$ para analisar se a média populacional difere do escore da ideologia moderada de 4,0. Agora, $ep = s/\sqrt{n} = 1{,}38/\sqrt{1331} = 0{,}038$, e:

$$t = \frac{\bar{y} - \mu_0}{ep} = \frac{4{,}12 - 4{,}0}{0{,}038} = 3{,}2.$$

O valor-p bilateral é igual a 0,001. Existe uma evidência *muito* forte de que a média verdadeira exceda 4,0; isto é, de que a média verdadeira está no lado conservador. Mas, em uma escala de 1,0 a 7,0, o valor 4,12 está bem próximo do escore moderado de 4,0. Embora a diferença de 0,12 entre a média amostral de 4,12 e a média de H_0 de 4,0 seja estatisticamente altamente significativa, a magnitude desta diferença é pequena em termos práticos. A resposta média sobre a ideologia política para todos os norte-americanos é essencialmente moderada. ■

No Exemplo 6.2, a média amostral da ideologia de 4,08 para $n = 186$ norte-americanos negros tinha valor-$p = 0{,}50$, não muita evidência contra H_0. Mas, se $\bar{y} = 4{,}08$ tivesse sido baseado em $n = 18600$ (isto é, 100 vezes maior do que n era), junto com $s = 1{,}51$, teríamos, então, encontrado $z = 6{,}79$ e um valor-p bilateral de 0,00000000001. Isto é estatisticamente altamente significativo, mas não praticamente significativo. Para finalidades práticas, uma média de 4,08 em uma escala de 1,0 a 7,0 para a ideologia política não difere de 4,00.

Vimos que, com tamanhos de amostras grandes, os valores-p podem ser pequenos mesmo quando a estimativa por ponto está próxima do valor de H_0. A magnitude do valor-p simplesmente resume a extensão da evidência sobre H_0 e não quão longe o parâmetro está de H_0. Sempre inspecione a diferença entre a estimativa e o valor de H_0 para avaliar as implicações práticas de um resultado de um teste.

Os testes de significância são menos úteis do que os intervalos de confiança

As hipóteses nulas contendo valores únicos raramente são verdadeiras. Isto é, raramente o parâmetro é *exatamente* igual ao valor listado em H_0. Com amostras suficientemente grandes, de forma que um erro do Tipo II seja improvável, estas hipóteses serão geralmente rejeitadas. O que é mais relevante é se o parâmetro é suficientemente diferente do valor de H_0 para ter importância prática.

Embora os testes de significância possam ser úteis, a maioria dos estatísticos acredita que eles foram enfatizados demais na pesquisa em Ciências Sociais. É preferível construir intervalos de confiança para parâmetros em vez de executar somente testes de significância. Um teste simplesmente indica se o valor em particular de H_0 é viável. Ele não nos diz que outros valores em potencial são viáveis. O intervalo de confiança, ao contrário, exibe todo o conjunto de valores possíveis. Ele mostra a extensão na qual H_0 pode ser falsa mostrando se os

valores no intervalo estão longe do valor de H_0. Portanto, ele nos ajuda a determinar se a rejeição de H_0 tem importância prática.

Para ilustrar, consideremos os dados da ideologia política do exemplo anterior, um intervalo de 95% de confiança para μ é $\bar{y} \pm 1{,}96(ep) = 4{,}12 \pm 1{,}96(0{,}038)$ ou (4,05; 4,20). Isso indica que a diferença entre a média populacional e o escore moderado de 4,0 é pequeno. Embora o valor-p = 0,001 forneça forte evidência contra H_0: μ = 4,0, em termos práticos o intervalo de confiança mostra que H_0 não está muito errada. Ao contrário, se $\bar{y}$ tivesse sido 6,125 (em vez de 4,125), o intervalo de 95% de confiança seria igual a (6,05; 6,20). Isto indicaria uma diferença substancial do valor moderado 4,0, com a resposta média estando mais próxima ao escore conservador do que do escore moderado.

Quando um valor-p não é pequeno, mas o intervalo de confiança é muito grande, isto nos força a perceber que o parâmetro pode muito bem-estar longe de H_0, embora não possamos rejeitá-lo. Isso também confirma por que não faz sentido "aceitar H_0", como a Seção 6.3 discutiu.

O restante do livro apresenta testes de significância para uma variedade de situações. É importante tornar-se familiar com esses testes, pelo seu uso frequente na pesquisa em Ciências Sociais. Contudo, também apresentaremos os intervalos de confiança que descrevem quão longe a realidade está do valor-H_0.

Interpretação errônea dos testes de significância e valores-*p*

Vimos que é impróprio "aceitar H_0". Vimos, também, que significância estatística não implica significância prática. Aqui estão outras interpretações errôneas possíveis dos testes de significância:

- **É equivocado relatar resultados somente se eles são estatisticamente significativos.** Alguns periódicos de pesquisa têm a política de publicar resultados de um estudo somente se o valor-p for menor do que 0,05. Aqui está o perigo dessa política: suponha que realmente não tenha efeito, mas 20 pesquisadores independentemente conduziram os estudos. Esperaríamos que aproximadamente 20(0,05) = 1 deles obteria um valor significativo no nível 0,05 simplesmente por acaso. (Quando H_0 é verdadeira, aproximadamente 5% das vezes obtemos um valor-p abaixo de 0,05.) Se aquele pesquisador, então, submeter os resultados a um periódico, mas os demais não, o artigo publicado será um erro do Tipo I. Ele irá relatar um efeito quando, na verdade, ele não existe.
- **Alguns testes podem ser estatisticamente significativos apenas por acaso.** Você nunca deveria escanear saídas do *software* para resultados que são estatisticamente significativos e relatá-los somente. Se você executar 100 testes, mesmo se todas as hipóteses nulas estão corretas, você esperaria obter valores-p inferiores ou equivalentes a 0,05 em aproximadamente 100(0,05) = 5 vezes. Seja cético com os relatórios de significância que simplesmente reflitam variabilidade aleatória ordinária.
- **É incorreto interpretar o valor-p como a probabilidade de que H_0 é verdadeira.** O valor-p é P(de que a estatística teste assuma um valor como o observado ou ainda mais extremo), presumindo que H_0 seja verdadeira. Ele não é P(H_0 ser verdadeira). Métodos estatísticos clássicos calculam probabilidades sobre variáveis e estatísticas (como a estatística-teste) que variam aleatoriamente de amostra para amostra, não sobre os parâmetros. As estatísticas têm distribuições amostrais; os parâmetros, não. Na realidade, não é uma questão de probabilidade. É verdadeira ou não é verdadeira, apenas não sabemos qual é o caso.

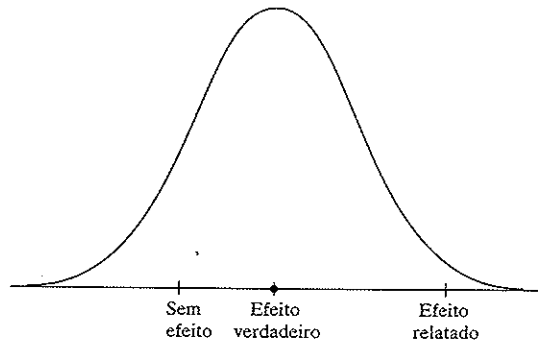
- **Efeitos verdadeiros podem ser menores do que as estimativas relatadas.** Mesmo se um resultado estatisticamente significativo for um efeito real, o verdadeiro efeito pode ser menor do que o relatado. Por exemplo, geralmente muitos pesquisadores executam estudos similares, mas os resultados que recebem atenção são os mais extremos. O pesquisador que decide publicá-los pode ser o que obteve o resultado amostral mais expressivo, talvez bem longe na cauda da distribuição amostral de todos os resultados possíveis. Veja a Figura 6.9.

EXEMPLO 6.8 Muitas das "descobertas" médicas são na verdade erros do Tipo I?

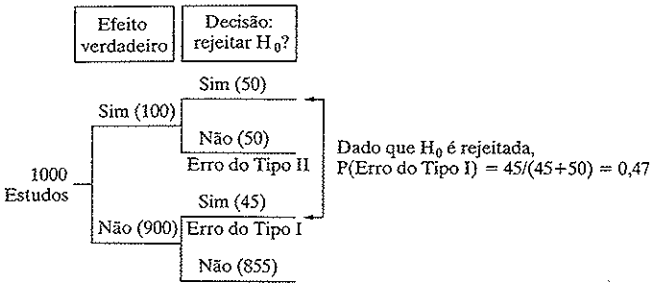
Nos estudos médicos, suponha que um efeito verdadeiro existe somente 10% das vezes. Suponha, também, que, se um efeito realmente existe, há uma chance de 50% de cometer um erro do Tipo II e não conseguir detectá-lo. Estes foram os percentuais hipotéticos usados em um artigo em um periódico médico.[4] Os autores observaram que muitos estudos médicos têm uma taxa alta de erros do Tipo II porque eles não são capazes de usar um tamanho da amostra grande. Assumindo essas taxas, um percentual substancial das "descobertas" médicas poderia ser realmente erros do Tipo I?

A Figura 6.10 é um **diagrama de árvore** mostrando o que iríamos esperar em 1000 estudos médicos que testam várias hipóteses. Se um efeito verdadeiro acontece somente 10% das vezes; este seria o caso para 100 dos 1000 estudos. Não obtemos um valor-p pequeno o suficiente para detectar este efeito verdadeiro 50% das vezes, isto é, em 50 destes 100 estudos. Um efeito será relatado para os outros 50 dos 100 que realmente têm um efeito. Para os 900 casos que na verdade não existe efeito, com o nível de significância usual de 0,05, esperamos que 5% dos 900 estudos erroneamente rejeitem H_0. Isso acontece para (0,05)900 = 45 estudos. Assim, dos 1000 estudos, esperamos que 50 relatem um efeito que realmente está lá, mas também esperamos que 45 relatem um efeito que, na verdade, não existe. Assim, uma proporção de 45/(45 + 50) = 0,47 de estudos médicos que relatam efeitos estão, de fato, apresentando erros do Tipo I.

A conclusão é ser cético. Quando você tiver notícias de novos avanços médicos, o efeito verdadeiro pode ser mais fraco do que o que for apresentado ou pode realmente não haver, de fato, qualquer efeito. ■



☑ **Figura 6.9** Quando muitos pesquisadores conduzem estudos, o resultado estatisticamente significativo publicado geralmente superestima o efeito verdadeiro.



☑ **Figura 6.10** Diagrama de árvore de 1000 estudos médicos hipotéticos. Ele assume que existe um efeito verdadeiro em 10% das vezes e que há probabilidade de 50% de ocorrência de um erro do Tipo II quando um efeito, de fato, existir.

6.6 CALCULANDO A P(ERRO DO TIPO II)*

Vimos que as decisões em testes de significância têm dois tipos de erros potenciais. Um erro do Tipo I que resulta da rejeição de H_0 quando ela é realmente verdadeira. Dado que H_0 é verdadeira, a probabilidade de um erro do Tipo I é o nível α do teste; quando $\alpha = 0,05$, a probabilidade de rejeitar H_0 é igual a 0,05.

Quando H_0 é falsa, um erro do Tipo II consiste em *não* rejeitá-la. Essa probabilidade tem mais do que um valor, porque H_a contém um intervalo de valores possíveis. Cada valor em H_a tem sua própria P(erro do Tipo II). Esta seção mostra como calcular P(erro do Tipo II) para um valor em particular.

EXEMPLO 6.9 Testando se a astrologia realmente funciona

Um teste científico da pseudociência, astrologia, usou o seguinte experimento:[5] para cada um de 116 sujeitos adultos, um astrólogo preparou um horóscopo baseado na posição dos planetas e da lua no momento do nascimento dessa pessoa. Cada sujeito preencheu, também, um levantamento de dados do California Personality Index (Índice de Personalidade da Califórnia). Para cada sujeito amostrado, sua data de nascimento e horóscopo foi mostrada a um astrólogo, bem como os resultados de sua personalidade em conjunto com outros dois adultos selecionados aleatoriamente do grupo experimental. Foi perguntado ao astrólogo qual era o mapa correto, entre os três, da personalidade para o sujeito sendo avaliado, tendo por base o seu horóscopo.

Considere π a representação da probabilidade de uma previsão correta feita pelo astrólogo. Se as previsões do astrólogo são aleatórias, como o suposto, então $\pi = 1/3$. Para testar isso contra a alternativa de que os acertos são melhores do que o acaso, podemos testar H_0: $\pi = 1/3$ contra H_a: $\pi > 1/3$. A hipótese alternativa reflete a crença dos astrólogos de que eles podem fazer melhor do que o acaso. Na verdade, o National Council for Geocosmic Research (Conselho Nacional da Pesquisa Geocósmica), que indicou os astrólogos para o experimento, afirmou que π seria 0,50 ou mais alto. Assim, vamos calcular P(erro do Tipo II) se realmente $\pi = 0,50$, para um teste de nível $\alpha = 0,05$. Isto é, se realmente $\pi = 0,50$, encontraremos a probabilidade de falhar na rejeição de H_0: $\pi = 1/3$.

Para determinar isso, primeiro vamos encontrar os valores da proporção amostral para os quais não rejeitamos H_0. Para o teste de H_0: $\pi = 1/3$, a distribuição amostral de $\hat{\pi}$ é a curva mostrada à esquerda na Figura 6.11. Com $n = 116$, esta curva tem o erro padrão de:

$$ep_0 = \sqrt{\pi_0(1 - \pi_0)/n}$$
$$= \sqrt{[(1/3)(2/3)]/116} = 0,0438.$$

Para H_a: $\pi > 1/3$, obtemos um valor-p de 0,05 se a estatística-teste $z = 1,645$. Isto é, 1,645 é o escore-z que tem uma probabilidade da cauda direita de 0,05. Portanto, não *conseguimos rejeitar* H_0, obtendo um valor-p *acima* de 0,05, se $z < 1,645$. Em outras palavras, não conseguimos rejeitar H_0: $\pi = 1/3$ se a proporção amostral $\hat{\pi}$ está a menos do que 1,645 erros padrão acima de 1/3, isto é, se

$$\hat{\pi} < 1/3 + 1.645(ep_0) =$$
$$1/3 + 1,645(0,0438) = 0,405.$$

Assim, a probabilidade da cauda direita acima de 0,405 é $\alpha = 0,05$ para a curva à esquerda na Figura 6.11.

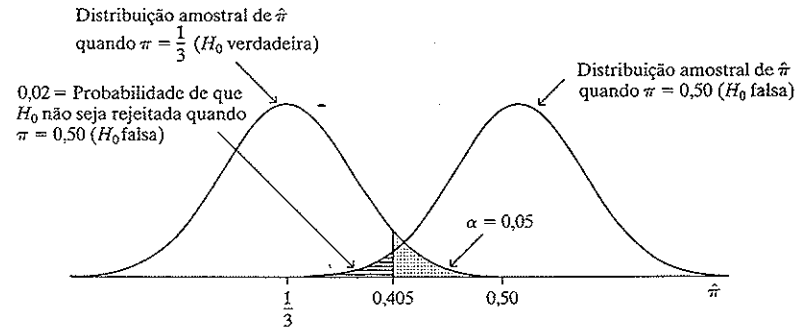
Para encontrar a P(erro do Tipo II) se π realmente é igual a 0,50, devemos encontrar $P(\hat{\pi} < 0,405)$ quando $\pi = 0,50$. Esta é a probabilidade da cauda esquerda *abaixo* de 0,405 para a curva à direita na Figura 6.11 (que é a curva correspondente a $\pi = 0,50$). Quando $\pi = 0,50$, o erro padrão para um tamanho da amostra de 116 é $\sqrt{[(0,50)(0,50)/116} = 0,0464$. (Isso difere um pouco do ep_0 para a estatística-teste, que usa 1/3 em vez de 0,50 para π.) Para a distribuição normal com uma média de 0,50 e erro padrão de 0,0464, o valor $\hat{\pi}$ de 0,405 tem um escore-z de

$$z = \frac{0,405 - 0,50}{0,0464} = -2,04.$$

A probabilidade de que $\hat{\pi} < 0,405$ é a probabilidade de que uma variável normal padrão esteja abaixo de $-2,04$. Da Tabela A, a probabilidade da cauda esquerda abaixo de $-2,04$ é igual a 0,02. Assim, para um tamanho da amostra de 116, a probabilidade de não rejeitar H_0: $\pi = 1/3$ é 0,02, se de fato $\pi = 0,50$.

Em outras palavras, se os astrólogos verdadeiramente têm o poder de previsão que eles afirmam ter, a chance de não detectar isto com esse experimento seria somente de aproximadamente 0,02. Para saber o que realmente aconteceu no experimento, veja o Exercício 6.17. ■

A probabilidade do erro do Tipo II aumenta quanto mais próximo o valor do parâmetro está da H_0. Para verificar isso, você pode verificar que P(erro do Tipo II)



☑ **Figura 6.11** Cálculo de P(erro do Tipo II) para testar H_0: $\pi = 1/3$ contra H_a: $\pi > 1/3$ no nível $\alpha = 0,05$, quando a proporção verdadeira é $\pi = 0,50$. Um erro do Tipo II ocorre se $\hat{\pi} < 0,405$, visto que, então, o valor-$p > 0,05$ embora H_0 seja falsa.

= 0,54 se $\pi = 0,40$. Assim, se o parâmetro está próximo do valor de H_0, pode haver uma chance substancial de não rejeitarmos H_0. Da mesma forma, quanto mais longe o parâmetro está da H_0, menor a probabilidade de um erro do Tipo II. A Figura 6.12 representa graficamente P(erro do Tipo II) para os vários valores π em H_a.

Para um nível α fixo e um valor alternativo do parâmetro, P(erro do Tipo II) diminui quando o tamanho da amostra aumenta. Se você puder obter mais dados, terá uma probabilidade menor de cometer esse tipo de erro.

Testes com um α menor têm uma P(erro do Tipo II) maior

Como a Seção 6.4 discutiu, quanto menor é α = P(erro do Tipo I) de um teste, maior a P(erro do Tipo II). Para ilustrar, suponha que o Exemplo 6.9 usou $\alpha = 0,01$. Portanto, você pode verificar que P(erro do Tipo II) = 0,08, comparado a P(erro do Tipo II) = 0,02 quando $\alpha = 0,05$.

A razão de não se utilizar valores extremamente pequenos para α, como $\alpha =$ 0,0001, é que P(erro do Tipo II) é muito alta. Será improvável rejeitar H_0 mesmo se o parâmetro estiver longe da hipótese nula. Em resumo, para os valores fixos dos demais fatores,

- P(erro do Tipo II) diminui à medida que:
  - o valor do parâmetro está longe de H_0.
  - o tamanho da amostra aumenta.
  - P(erro do Tipo I) aumenta.

O poder de um teste

Quando H_0 é falsa, você quer que a probabilidade de rejeitar H_0 seja alta. A probabilidade de rejeitar H_0 é chamada de **poder** de um teste. Para um valor, em particular, do parâmetro no intervalo de valores da H_a,

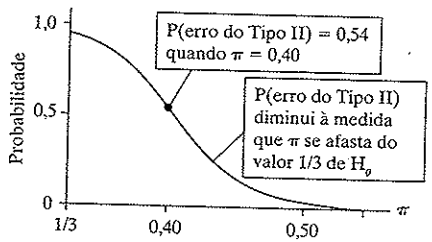
$$\text{Poder} = 1 - \text{P(erro do Tipo II)}$$

No Exemplo 6.9, tem-se que o teste da H_0: $\pi = 1/3$ tem P(erro do tipo II) = 0,02 quando $\pi = 0,50$. Portanto, o poder do teste quando $\pi = 0,50$ é igual a $1 - 0,02 = 0,98$.

O poder aumenta para os valores do parâmetro que estão mais longe do valor de H_0. Assim como a curva para a P(erro do Tipo II), na Figura 6.12, diminui à medida que π se afasta acima de $\pi_0 = 1/3$, a curva para o poder aumenta.

Na prática, os estudos deveriam ter, em condições ideais, alto poder. Antes de conceder apoio financeiro para um estudo planejado, muitas agências de pesquisa esperam que os principais investigadores mostrem que exista um poder razoável (geralmente, pelo menos 0,80) nos valores do parâmetro que são considerados praticamente significativos.

Quando você lê que os resultados de um estudo são significativos, seja cético



☑ **Figura 6.12** Probabilidade do erro do Tipo II para testar H_0: $\pi = 1/3$ contra H_a: $\pi > 1/3$ ao nível $\alpha = 0,05$, representado graficamente para os valores potenciais em H_a.

se nenhuma informação é dada sobre o poder. O poder pode ser baixo, especialmente se n é pequeno. Para mais detalhes sobre o cálculo da P(erro do tipo II) ou poder, veja Cohen (1988).

6.7 TESTE PARA UMA PROPORÇÃO EM PEQUENAS AMOSTRAS – A DISTRIBUIÇÃO BINOMIAL*

Para uma proporção da população π, a Seção 6.3 apresentou um teste de significância que é válido para grandes amostras. A distribuição amostral da proporção $\hat{\pi}$ é, nesse caso, aproximadamente normal, o que justifica o uso da estatística-teste z.

Para um n pequeno, a distribuição amostral de $\hat{\pi}$ apresenta valores somente em alguns pontos. Se $n = 5$, por exemplo, os únicos valores possíveis para a proporção amostral $\hat{\pi}$ são 0, 1/5, 2/5, 3/5, 4/5 e 1. Uma aproximação contínua como a distribuição normal é inapropriada, nessa situação. Além disto, veremos que, quanto mais próximo π está de 0 ou 1 para um tamanho da amostra dado, mais assimétrica se torna a distribuição amostral real.

Esta seção introduz um teste para proporção em pequenas amostras. Ele usa a distribuição de probabilidade mais importante para variáveis discretas, a *binomial*.

A distribuição binomial

Para dados categóricos, as seguintes condições podem ser observadas:

1. Cada observação está em apenas uma de duas categorias.
2. As probabilidades para as duas categorias são as mesmas para cada observação. Representamos as probabilidades por π para a categoria 1 e $(1 - \pi)$ para a categoria 2.
3. Os resultados de observações sucessivas são independentes. Isto é, o resultado para uma observação não depende dos resultados de outras observações.

Lançar uma moeda repetidamente é um protótipo para essas condições. Para cada lançamento, observamos se o resultado é cara (categoria 1) ou coroa (categoria 2). A probabilidade dos resultados são as mesmas para cada lançamento (0,50 para cada lançamento se a moeda for equilibrada). O resultado para um lançamento em particular não depende do resultado de outros lançamentos.

Agora, para n observações, considere x a representação do número que ocorre na categoria 1. Por exemplo, para $n =$ 5 lançamentos da moeda, x = número de caras, poderia ser igual a 0, 1, 2, 3, 4 ou 5. Quando as observações satisfazem as condições acima, a distribuição de probabilidade de x é a **distribuição binomial**.

A variável binomial x é discreta, assumindo um dos valores inteiros 0, 1, 2,..., n. A fórmula para as probabilidades binomiais segue:

☑ **Probabilidades para uma distribuição binomial**

Represente a probabilidade da categoria 1, para cada observação, por π. Para n observações independentes, a probabilidade de x resultados na categoria 1 é:

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x},$$

$$x = 0, 1, 2, \ldots, n.$$

O símbolo $n!$ é chamado de ***n* fatorial**. Ele representa $n! = 1\times2\times3\times\ldots\times n$. Por exemplo, $1! = 1$; $2! = 1\times2 = 2$; $3! = 1\times2\times3 = 6$ e assim por diante. Também, $0!$ é definido como 1.

Para valores em particular para π e n, substituir os valores possíveis para x na fórmula para $P(x)$ fornece as probabilidades dos resultados possíveis. A soma das probabilidades se iguala a 1,0.

EXEMPLO 6.10 Gênero e a seleção de *trainees* para a gerência

O Exemplo 6.1 (página 169) discutiu um caso envolvendo tendenciosidade potencial contra mulheres na seleção de *trainees* para gerência de uma grande cadeia de supermercados. O grupo de empregados é metade feminina e metade masculina. Dez *trainees* devem ser escolhidos aleatoriamente deste grupo. Se eles são verdadeiramente selecionados ao acaso, quantas mulheres seriam esperadas?

A probabilidade de que uma pessoa selecionada seja uma mulher é $\pi = 0{,}50$, a proporção de *trainees* disponíveis que são mulheres. Da mesma forma, a probabilidade de que uma pessoa selecionada seja um homem é $(1 - \pi) = 0{,}50$. Seja x = número de mulheres selecionadas. Essa variável tem uma distribuição binomial com $n = 10$ e $\pi = 0{,}50$. Para cada x entre 0 e 10, a probabilidade de que x, das dez pessoas selecionadas, seja mulher é:

$$P(x) = \frac{10!}{x!(10-x)!}(0{,}50)^x(0{,}50)^{10-x}, \quad x = 0, 1, 2, \ldots, 10.$$

Por exemplo, a probabilidade de que nenhuma mulher seja selecionada ($x = 0$) é igual a:

$$P(0) = \frac{10!}{0!10!}(0{,}50)^0(0{,}50)^{10} = (0{,}50)^{10} = 0{,}001.$$

Lembre que qualquer número elevado a potência 0 é igual a 1. Também que 0! = 1 e que termos 10! no numerador e denominador se cancelam, deixando $P(0) = (0{,}50)^{10}$. A probabilidade de que exatamente uma mulher seja escolhida igual:

$$P(1) = \frac{10!}{1!9!}(0{,}50)^1(0{,}50)^9 = 10(0{,}50)(0{,}50)^9 = 0{,}010.$$

Este cálculo é simplificado consideravelmente observando que 10!/9! = 10, visto que 10! é apenas 9! multiplicado por 10. A Tabela 6.6 lista toda a distribuição binomial para $n = 10$, $\pi = 0{,}50$.

Na Tabela 6.6, a probabilidade é aproximadamente 0,98 de que x esteja entre 2 e 8, inclusive. Os valores menos prováveis para x são 0, 1, 9 e 10, que tem uma probabilidade combinada de somente 0,022. Se as amostras foram selecionadas aleatoriamente, algo em torno de duas e oito mulheres seriam provavelmente selecionadas. É especialmente improvável que nenhuma ou dez fossem selecionadas.

As probabilidades para as mulheres determinam as dos homens. Por exemplo, a probabilidade de que nove das dez pessoas selecionadas sejam homens é igual à probabilidade de que uma das dez selecionadas seja mulher. ■

☑ **Tabela 6.6** A distribuição binomial para $n = 10$, $\pi = 0{,}50$. A variável pode assumir qualquer valor entre 0 e 10

x	P(x)	x	P(x)
0	0,001	6	0,205
1	0,010	7	0,117
2	0,044	8	0,044
3	0,117	9	0,010
4	0,205	10	0,001
5	0,246		

Propriedades da distribuição binomial

A distribuição binomial é perfeitamente simétrica somente quando $\pi = 0{,}50$. No Exemplo 6.10 com $n = 10$, por exemplo, visto que a proporção populacional de mulheres é igual a 0,50, $x = 10$ tem a mesma probabilidade de $x = 0$.

A proporção amostral $\hat{\pi}$ se relaciona a x por:

$$\hat{\pi} = x/n.$$

Por exemplo, para $x = 1$ mulher escolhida de $n = 10$, $\hat{\pi} = 1/10 = 0{,}10$. A distribuição amostral de $\hat{\pi}$ também é simétrica quando $\pi = 0{,}50$. Quando $\pi \neq 0{,}50$, as distribuições são assimétricas, com o grau de assimetria aumentando à medida que π está mais próximo de 0 ou 1. A Figura 6.13 ilustra distribuições amostrais de $\hat{\pi}$. Por exemplo, quando $\pi = 0{,}10$, a proporção amostral $\hat{\pi}$ não pode estar muito abaixo de 0,10 visto que ela deve ser positiva, mas ela poderia estar consideravelmente acima de 0,10.

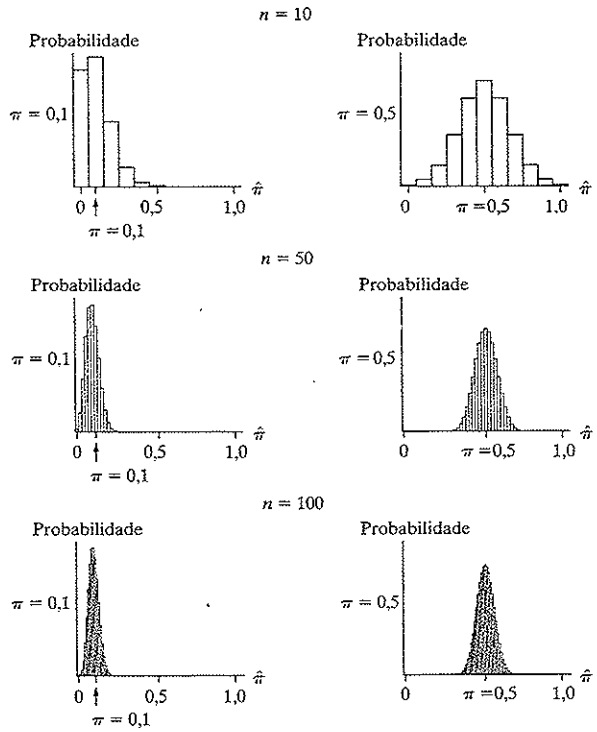
A distribuição binomial tem a média e o desvio padrão dados por:

$$\mu = n\pi, \quad \sigma = \sqrt{n\pi(1-\pi)}.$$

Por exemplo, suponha que a chance de uma mulher em qualquer seleção para o treinamento de gerência é de 0,50, como a cadeia de supermercado afirma. Então, em 10 *trainees*, esperamos ter $\mu = n\pi = 10(0{,}50) = 5$ mulheres.

EXEMPLO 6.11 Quanta variabilidade uma pesquisa de boca de urna pode mostrar?

O Exemplo 4.6 (página 106) discutiu uma pesquisa de boca de urna de 2705 eleitores para a eleição governamental da Califórnia de 2006. Seja x o número, na pesquisa de boca de urna, dos eleitores de Arnold Schwarzenegger. Em uma população de aproximadamente 7 milhões de eleitores,



☑ **Figura 6.13** Distribuição amostral de $\hat{\pi}$ quando $\pi = 0{,}10$ ou 0,50, para $n = 10$, 50, 100.

55,9% votou nele. Se a pesquisa de boca de urna foi aleatoriamente selecionada, então a distribuição binomial para x tem $n = 2705$ e $\pi = 0{,}559$. A distribuição é descrita por:

$$\mu = 2705(0{,}559) = 1512,$$
$$\sigma = \sqrt{2705(0{,}559)(0{,}441)} = 26.$$

É quase certo que x estaria dentro de 3 desvios padrão da média. Esse intervalo varia de 1434 a 1590. Na verdade, naquela pesquisa de boca de urna, 1528 eleitores afirmaram ter votado em Schwarzenegger. ■

Vimos (Seções 5.2, página 134 e 6.3, página 182) que a distribuição amostral da proporção amostral $\hat{\pi}$ tem média e erro padrão $\sigma_{\hat{\pi}} = \sqrt{\pi(1-\pi)/n}$. Para obtê-los, dividimos a média da binomial $\mu = n\pi$ e o desvio padrão $\sigma = \sqrt{n\pi(1-\pi)}$ por n, visto que $\hat{\pi}$ divide x por n.

A distribuição binomial e a distribuição amostral de $\hat{\pi}$ são aproximadamente normais para n grande. Essa aproximação é a base do teste para grandes amostras da Seção 6.3. Quão grande é "grande"? Uma norma é que o número esperado de observações deva ser pelo menos 10 para ambas as categorias. Por exemplo, se $\pi = 0{,}50$, precisamos de pelo menos $n = 20$ porque, então, esperamos $20(0{,}50) = 10$ observações em uma categoria e $20(1 - 0{,}50) = 10$ em outra categoria. Para testar H_0: $\pi = 0{,}90$ ou H_0: $\pi = 0{,}10$, precisamos de $n \geq 100$. O requisito do tamanho da amostra reflete o fato de que uma curva simétrica para a distribuição amostral de $\hat{\pi}$ requer tamanhos da amostra maiores quando π está próximo de 0 ou 1 do que quando π está próximo de 0,50.

O teste binomial

Se o tamanho da amostra não é grande o suficiente para usar o teste normal, podemos usar a distribuição binomial diretamente. Veja o Exemplo 6.10 (página 196) sobre descriminação potencial de gênero. Para a distribuição amostral, a probabilidade π de que uma pessoa selecionada para o treinamento de gerência seja mulher é igual a 0,50. Se existir tendenciosidade contra as mulheres, então $\pi < 0{,}50$. Portanto, podemos testar a afirmação da empresa de amostragem aleatória fazendo:

$$H_0\text{: } \pi = 0{,}50 \quad \textit{versus} \quad H_a\text{: } \pi < 0{,}50.$$

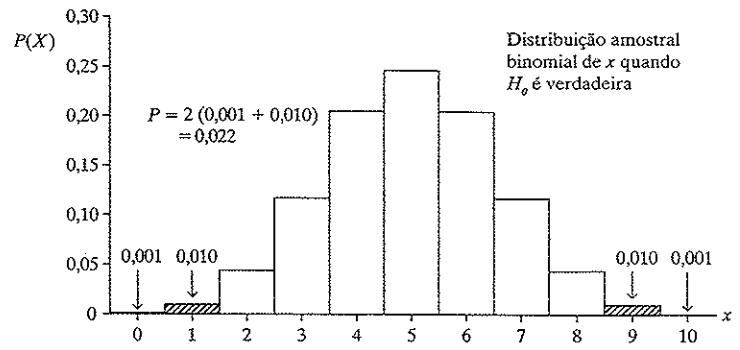
Dos dez empregados selecionados para o treinamento de gerência, seja x o número de mulheres. Sob H_0, a distribuição amostral de x é binomial com $n = 10$ e $\pi = 0{,}50$. A Tabela 6.6 mostrou essa distribuição. Como no Exemplo 6.1 (página 169), suponha que $x = 1$. O valor-p é, então, a probabilidade da cauda esquerda de um resultado, pelo menos, tão extremo; isto é, $x = 1$ ou 0. Da Tabela 6.6, o valor-p é igual a:

$$\text{P} = \text{P}(0) + \text{P}(1) = 0{,}001 + 0{,}010 = 0{,}011.$$

Se a empresa selecionou os *trainees* aleatoriamente, a probabilidade de escolher uma ou menos mulheres é somente de 0,011. Esse resultado fornece evidência contra a hipótese nula de um processo de seleção aleatório. Podemos rejeitar H_0 para $\alpha = 0{,}05$, embora não para $\alpha = 0{,}010$.

Mesmo se tivermos a suspeita de tendenciosidade em uma direção em particular, a forma mais imparcial de executar um teste usa uma alternativa bilateral. Para H_a: $\pi \neq 0{,}50$, o valor-p é $2(0{,}011) = 0{,}022$. Isto é uma probabilidade de duas caudas do resultado de que um ou menos de *ambos os sexos* seja selecionado. A Figura 6.14 mostra a construção desse valor-p.

As suposições para o teste binomial são as três condições vistas anteriormente para ela. Aqui, as condições são satisfeitas. Cada observação tem somente dois resultados possíveis, mulher ou homem. A probabilidade de cada resultado é a mesma para cada seleção, 0,50 para a seleção de uma mulher e 0,50 para a seleção de um homem (sob H_0). Para a amostragem aleatória um resultado não depende de qualquer outro.



☑ **Figura 6.14** Cálculo do valor-p testando H_0: $\pi = 0{,}50$ contra H_a: $\pi \neq 0{,}50$, quando $n = 10$ e $x = 1$.

Em casos raros em que o tamanho da população é pequeno, as condições binomiais não são todas satisfeitas. Para ilustrar, suponha que a população contenha somente quatro pessoas das quais duas são mulheres. Se amostrarmos aleatoriamente dois indivíduos separados, a segunda observação tem probabilidades diferentes da primeira. Por exemplo, se a primeira pessoa selecionada foi mulher, então a probabilidade de que a segunda pessoa selecionada seja mulher é igual a 1/3, visto que 1 mulher permanece entre 3 sujeitos. Portanto, as probabilidades não são as mesmas para cada seleção, que a binomial requer. Para as observações sucessivas terem essencialmente as mesmas probabilidades e serem independentes, o tamanho da população deve ser bem maior do que o tamanho da amostra. O tamanho da amostra não deve ser mais do que 10% do da população de sujeitos nas duas categorias. Isto é, em geral, facilmente satisfeito na prática.

6.8 RESUMO DO CAPÍTULO

A inferência estatística usa dados amostrais para fazer previsões sobre os parâmetros da população. Os Capítulos 5 e 6 introduziram dois métodos de inferência – **estimação** e **testes de significância**. O método de estimação chamado de *intervalos de confiança* fornece um conjunto de valores mais plausíveis para um parâmetro. Um teste de significância julga se um valor em particular para um parâmetro é plausível. Ambos os métodos utilizam a distribuição amostral do estimador do parâmetro.

Os testes de significância têm cinco partes:

***1.* Suposições:**

- Os testes para *médias* se aplicam a variáveis quantitativas, enquanto testes para *proporções* se aplicam a variáveis categóricas.
- Os testes assumem que a amostra utilizada é aleatória.
- Testes para proporções com grandes amostras não requerem suposição sobre a distribuição da população porque o Teorema Central do Limite implica normalidade aproximada da distribuição amostral da proporção. Isto justifica o uso da estatística-teste z. Testes

para proporções com amostras pequenas utilizam a *distribuição binomial*.
- Os testes para as médias usam a distribuição t para a estatística-teste. O teste assume que a distribuição da população seja normal. Na prática, testes bilaterais (como os intervalos de confiança) são *robustos* às violações da suposição de normalidade, especialmente para grandes amostras em virtude do Teorema Central do Limite.

2. **Hipótese nula e alternativa** sobre o parâmetro: a hipótese nula tem a forma H_0: $\mu = \mu_0$ para uma média e H_0: $\pi = \pi_0$ para uma proporção. Aqui, μ_0 e π_0 representam os valores hipotéticos para os parâmetros, como 0,50 em H_0: $\pi = 0{,}50$. A alternativa mais comum é a bilateral tal como Ha: $\pi \neq 0{,}50$. Hipóteses como H_a: $\pi > 0{,}50$ e H_a: $\pi < 0{,}50$ são *unilaterais*, delineadas para detectar desvios da H_0 em uma direção em particular.
3. Uma **estatística-teste** descreve quão longe a estimativa por ponto está do valor H_0. A estatística z para proporções e uma estatística t para médias mensura o número de erros padrão que a estimativa por ponto ($\hat{\pi}$ ou $\bar{y}$) está do valor H_0.
4. O **valor-*p*** descreve o peso da evidência que os dados fornecem sobre H_0.
   - O valor-p é calculado assumindo que H_0 é verdadeira. Ele se iguala à probabilidade de que a estatística-teste é igual ao valor observado ou a um valor ainda mais extremo.
   - Os resultados "mais extremos" são determinados pelo tipo de hipótese alternativa. Para H_a bilateral, o valor-p é uma probabilidade de duas caudas.
   - Os valores-p pequenos resultam quando a estimativa por ponto está longe do valor H_0, de forma que a estatística-teste é grande. Quando o valor-p é pequeno, seria incomum observar tais dados se H_0 fosse verdadeira. Quanto menor o valor-p, mais forte a evidência contra H_0.
5. Uma **conclusão** baseada na evidência da amostra sobre H_0: relatamos e interpretamos o valor-p. Algumas vezes é necessário tomar uma decisão. Se o valor-p é menor ou igual a um nível α fixo (como $\alpha = 0{,}05$), rejeitamos H_0. De outra forma, não podemos rejeitá-la.

Quando tomamos uma decisão, dois tipos de erros podem ocorrer.

☑ **Tabela 6.7** Resumo dos testes de significância para médias e proporções

Parâmetro	Média	Proporção
1. Suposições	Amostra aleatória, variável quantitativa e população normal	Amostra aleatória, variável categórica e contagem esperada da hipótese de pelo menos 10
2. Hipóteses	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$ H_a: $\mu > \mu_0$ H_a: $\mu < \mu_0$	H_0: $\pi = \pi_0$ H_a: $\pi \neq \pi_0$ H_a: $\pi > \pi_0$ H_a: $\pi < \pi_0$
3. Estatística-teste	$t = \dfrac{\bar{y} - \mu_0}{ep}$ com $ep = \dfrac{s}{\sqrt{n}}$, $gl = n - 1$	$z = \dfrac{\hat{\pi} - \pi_0}{ep_0}$ com $ep_0 = \sqrt{\pi_0(1 - \pi_0)/n}$
4. Valor-p	Probabilidade de duas caudas na distribuição amostral para um teste bilateral (H_0: $\mu \neq \mu_0$ ou H_a: $\pi \neq \pi_0$); Probabilidade de uma cauda para um teste unilateral	
5. Conclusão	Rejeitar H_0 se o valor-$p \leq$ nível α como 0,05	

- Quando H_0 é verdadeira, um erro do Tipo I resulta se a rejeitarmos.
- Quando H_0 é falsa, um erro do Tipo II resulta se não conseguimos rejeitá-la.

O valor α, o ponto de corte, para o valor-p na tomada da decisão, é igual a P(erro do Tipo I). Normalmente, escolhemos pequenos valores como $\alpha = 0{,}05$ ou 0,01. Para um α fixo, P(erro do Tipo II) diminui à medida que a distância aumenta entre o parâmetro e o valor H_0 ou à medida que o tamanho da amostra aumenta.

A Tabela 6.7 resume as cinco partes dos testes que este capítulo apresentou.

O tamanho da amostra é um fator crítico em estimação e testes de significância. Com tamanhos das amostras pequenos, os intervalos de confiança são grandes, tornando a estimativa imprecisa. Tamanhos da amostra pequenos também dificultam a rejeição de hipóteses nulas falsas a não ser que o parâmetro verdadeiro esteja longe do valor da hipótese nula. A P(erro do Tipo II) pode ser alta para os valores do parâmetro de interesse.

Para introduzir a estimação e os testes de significância, os Capítulos 5 e 6 apresentaram inferência sobre um único parâmetro para uma única variável. Na prática, geralmente é artificial ter um número fixo em particular para o valor H_0 de um parâmetro. Uma das poucas vezes em que isso acontece é quando os escores da resposta resultam da diferença de dois valores, como a mudança no peso do Exemplo 6.4 (página 177). Naquele caso, $\mu_0 = 0$ é uma linha de base natural. Os testes de significância mais comumente se referem às comparações de médias para duas amostras do que para um valor fixo de um parâmetro para uma única amostra. O próximo capítulo mostra como comparar médias ou proporções para dois grupos.

EXERCÍCIOS

Praticando o básico

6.1 De (a) a (c), verifique se é uma hipótese nula ou uma hipótese alternativa?
- **(a)** No Canadá, a proporção de adultos que são a favor do jogo é igual a 0,50.
- **(b)** A proporção de todos os estudantes universitários canadenses que são fumantes regulares agora é de 0,24 (o valor que era dez anos atrás).
- **(c)** A média dos QIs de todos os estudantes da Escola de Ensino Médio Lake Wobegon é maior do que 100.
- **(d)** Introduzindo uma notação para um parâmetro, declare as hipóteses em (a) a (c) em termos dos valores do parâmetro.

6.2 Você quer saber se os adultos em seu país acham que o número ideal de filhos é igual a 2 ou maior ou menor do que isto.
- **(a)** Defina a notação e declare a hipótese nula e a alternativa para testar isso.
- **(b)** Para respostas de uma PSG recente à pergunta: "Qual é o número de filhos que você acha ideal?", um *software* mostrou os seguintes resultados:

```
Teste de μ = 2,0 versus μ ≠ 2,0
Variável  N     Média  Desvio Padrão  EP da média  T      Valor-p
CHILDREN  1302  2,490  0,850          0,0236       20,80  0,0000
```

Relate o valor da estatística-teste e mostre como ele foi obtido dos outros valores apresentados na tabela.
- **(c)** Explique o que o valor-p representa e interprete o seu valor.

6.3 Para um teste H_0: $\mu = 0$ contra H_a: $\mu \neq 0$ com $n = 1000$, a estatística-teste é igual a 1,04.
- **(a)** Encontre o valor-p e interprete-o. (*Nota*: você pode usar a normal padrão para aproximar a distribuição t).
- **(b)** Suponha que $t = -2{,}50$ em vez de 1,04. Encontre o valor-p. Isto for-

nece evidência mais forte ou mais fraca contra a hipótese nula? Explique.

(c) Quando $t = 1{,}04$, encontre o valor-p para (i) H_a: $\mu > 0$, (ii) H_a: $\mu < 0$.

6.4 O valor-p para um teste sobre a média com $n = 25$ é 0,05.

(a) Encontre o valor da estatística-teste t que tem esse valor-p para (i) H_a: $\mu \neq 0$, (ii) H_a: $\mu > 0$, (iii) H_a: $\mu < 0$.

(b) Este valor-p fornece evidência mais forte ou mais fraca contra a hipótese nula do que 0,01? Explique.

6.5 Encontre e interprete o valor-p para testar H_0: $\mu = 100$ contra H_0: $\mu \neq 100$, se uma amostra tem

(a) $n = 400$, $\bar{y} = 103$, $s = 40$.

(b) $n = 1600$, $\bar{y} = 103$, $s = 40$. Comente sobre o efeito de n nos resultados do teste de significância.

6.6 O Exemplo 6.4 (página 177) descreveu um estudo sobre terapias para meninas adolescentes que sofrem de anorexia. Para as 17 meninas que receberam terapia familiar, as mudanças no peso foram:

11, 11, 6, 9, 14, −3, 0, 7, 22, −5, −4, 13, 13, 9, 4, 6, 11.

Parte da saída do SPSS para os dados mostra o seguinte:

Inferior	Superior	Valor-t	gl	Sig. bilateral
3,60				0,0007

Preencha os resultados que estão faltando.

6.7 Segundo a declaração de um sindicato, a renda média para todos os trabalhadores da categoria superior da linha de montagem de uma grande empresa é igual a $500 por semana. Um representante de um grupo de mulheres decide analisar se a renda média μ para funcionárias se equipara a esse valor. Para uma amostra aleatória de nove mulheres funcionárias, $\bar{y} = \$410$ e $s = \$90$.

(a) Teste se a renda média das funcionárias difere de $500 por semana. Inclua as suposições, as hipóteses, a estatística-teste e o valor-p. Interprete o resultado.

(b) Determine o valor-p para H_a: $\mu < 500$. Interprete.

(c) Determine e interprete o valor-p para H_a: $\mu > 500$. (*Dica*: os valores-p para os dois testes unilaterais devem somar 1.)

6.8 Por lei, uma instalação industrial pode despejar não mais do que 500 galões de água com resíduos por hora, em média, em um lago vizinho. Baseado em outras infrações que eles observaram, um grupo de ação ambiental acredita que este limite está sendo ultrapassado. Monitorar a instalação é caro e uma amostra aleatória de quatro horas é selecionada de um período de uma semana. Um *software* relata:

Variável	N° de casos	Média	DP	EP da Média
RESÍDUO	4	1000,0	400,0	200,0

(a) Teste se o resíduo médio despejado é igual a 500 galões por hora contra a alternativa de que o limite está sendo ultrapassado. Encontre o valor-p e interprete.

(b) Explique por que o teste pode ser altamente aproximado ou até mesmo inválido se a distribuição da população do resíduo desejado está longe de uma distribuição normal.

(c) Explique como a sua análise unilateral implicitamente testa a hipótese nula mais ampla de que $\mu \leq 500$.

6.9 Em resposta à declaração: "É provável que uma criança da pré-escola sofra se sua mãe trabalha", as categorias da resposta (Concorda plenamente, Concorda, Discorda, Discorda plenamente) têm a contagem (91, 385, 421, 99) para respostas em uma PSG. Para tratar esta variável ordinal como quantitativa, designamos escores às categorias. Para os escores (2, 1, −1, −2), os quais tratam a distância entre concorda e discorda como o dobro da distância entre concorda plenamente e concorda, ou entre discorda e discorda plenamente, o *software* apresenta:

N	Média	Desvio Padrão	Erro Padrão
996	−0,052	1,253	0,0397

(a) Estabeleça as hipóteses nula e alternativa para testar se a resposta média da população difere do valor neutro 0.

(b) Encontre a estatística-teste e o valor-p. Interprete e tome uma decisão sobre H_0 usando $\alpha = 0{,}05$.

(c) Baseado em (b), você pode "aceitar" H_0: $\mu = 0$? Por que ou por que não?

(d) Construa um intervalo de 95% de confiança para μ. Mostre a correspondência entre 0 estar no intervalo e a decisão sobre H_0.

6.10 No Exemplo 6.2 da ideologia política (página 175), suponha que sejam utilizados os escores (-3, −2, −1, 0, 1, 2, 3) em vez dos escores (1, 2, 3, 4, 5, 6) usados no exemplo. Teste, então, H_0: $\mu = 0$. Explique o efeito da mudança dos escores em **(a)** a média amostral e o desvio padrão, **(b)** a estatística-teste, **(c)** o valor-p e a sua interpretação.

6.11 Os resultados de um intervalo de 99% de confiança para médias são consistentes com os resultados de testes bilaterais com qual nível α? Explique a conexão.

6.12 Para um teste de H_0: $\pi = 0{,}50$, a estatística-teste z é igual a 1,04.

(a) Encontre o valor-p para H_a: $\pi > 0{,}50$.

(b) Encontre o valor-p para H_a: $\pi \neq 0{,}50$.

(c) Encontre o valor-p para H_a: $\pi < 0{,}50$.

(d) Os valores-p em (a), (b) ou (c) fornecem fortes evidências contra H_0? Explique.

6.13 Para um teste de H_0: $\pi = 0{,}50$, a proporção amostral é 0,35 baseado em uma amostra de tamanho 100.

(a) Mostre que a estatística-teste é $z = -3{,}0$.

(b) Encontre e interprete o valor-p para H_a: $\pi < 0{,}50$.

(c) Para um nível de significância de $\alpha = 0{,}05$, qual a sua decisão?

(d) Se a decisão em (c) foi um erro, que tipo de erro foi? O que você poderia fazer para reduzir a chance deste tipo de erro?

6.14 O casamento entre pessoas do mesmo sexo foi legalizado em todo o Canadá pela Lei do Casamento Civil de 2005. Esta lei é apoiada pela maioria ou minoria da população canadense? Uma pesquisa de opinião pública conduzida para o jornal *Globe and Mail* em julho de 2005 com 1000 canadenses perguntou se esta lei deveria permanecer ou ser derrubada. As respostas foram 55% para *deveria permanecer*, 39% para *deveria ser* e 6% não sabiam. Seja π a representação da proporção populacional de adultos canadenses que acreditam que ela deveria permanecer. Para testar H_0: $\pi = 0{,}50$ contra H_a: $\pi \neq 0{,}50$:

(a) Encontre o erro padrão e interprete.

(b) Encontre a estatística-teste e interprete.

(c) Encontre o valor-p e interprete no contexto.

6.15 Quando uma PSG recente perguntou: "Você estaria disposto a pagar taxas mais altas para proteger o meio ambiente?" (variável "GRNTAXES"), 369 pessoas responderam *sim* e 483 responderam *não*. O *software* mostra os seguintes resultados para analisar se a maioria ou minoria dos norte-americanos iria responder *sim*:

Teste de proporção igual a 0,5 *versus* diferente de 0,5

N	Proporção Amostral	IC 95%	Valor-z	Valor-p
852	0,4331	(0,400; 0,466)	-3,91	0,000

(a) Especifique a hipótese que está sendo testada.

(b) Relate e interprete o valor da estatística-teste.

(c) Relate e interprete o valor-p como uma probabilidade.

(d) Explique uma vantagem do intervalo de confiança mostrado sobre o teste de significância.

6.16 Uma pesquisa de opinião pública da Pew Research Center (14 de maio de

2003) com 1201 adultos perguntou: "De modo geral, você acha que os programas de ações afirmativas delineados para aumentar o número de negros e minorias de estudantes nos *campi* das universidades são uma coisa boa ou ruim?". Sessenta por cento disserem *bom*, 30% disseram *ruim* e 10% não sabiam. Seja π a proporção populacional dos que disseram que era uma coisa boa. Encontre o valor-p para testar H_0: $\pi = 0{,}50$ contra H_a: $\pi \neq 0{,}50$. Interprete.

6.17 No teste científico de astrologia discutido no Exemplo 6.9 (página 192), os astrólogos acertaram 40 das suas 116 previsões. Teste H_0: $\pi = 1/3$ contra H_a: $\pi > 1/3$. Encontre o valor-p, tome uma decisão usando $\alpha = 0{,}05$ e interprete.

6.18 O exercício anterior analisou se os astrólogos poderiam prever o mapa correto da personalidade, para um horóscopo dado, fazendo melhor do que estimativa aleatória. No contexto do estudo o que seria um:

(a) Erro do Tipo I?

(b) Erro do Tipo II?

6.19 Uma eleição para a prefeitura de Madison, Wisconsin, tem dois candidatos. Exatamente metade dos residentes prefere cada candidato.

(a) Para uma amostra aleatória de 400 eleitores, 230 votaram em um candidato em particular. Você gostaria de prever o vencedor? Por quê?

(b) Para uma amostra aleatória de 40 eleitores, 23 votaram em um candidato em particular. Você gostaria de prever o vencedor? Por quê? (A proporção amostral é a mesma em (a) e (b), mas o tamanho da amostra difere.)

6.20 A autoria de um documento antigo está em dúvida. Um historiador teoriza que a autora era uma jornalista chamada Jacalyn Levine. Sob uma investigação meticulosa dos trabalhos conhecidos de Levine, é observado que uma característica incomum do seu trabalho era que ela consistentemente começava 6% das suas frases com a palavra "enquanto". Para testar a hipótese do historiador, decidiu-se contar o número de frases no documento em disputa que iniciam com "enquanto". De 300 frases, nenhuma começava. Seja π a representação da probabilidade de que cada frase escrita pelo autor desconhecido inicia com "enquanto". Teste H_0: $\pi = 0{,}06$ contra H_a: $\pi \neq 0{,}06$. Quais suposições são necessárias para a sua conclusão ser válida? (F. Mosteller e D. L. Wallace conduziram esse tipo de investigação para determinar se Alexander Hamilton ou James Madison eram os autores dos 12 *Federalist Papers*. Veja *Inference and Disputed Autorship: The Federalist*, Addison-Wesley, 1964).

6.21 Um teste de escolha múltipla tem quatro respostas possíveis. A pergunta é difícil, com nenhuma das quatro respostas sendo obviamente erradas, todavia com somente uma resposta correta. Ele ocorreu, inicialmente, em um exame prestado por 400 estudantes. Teste se mais pessoas responderam à pergunta corretamente do que seria esperado devido ao caso (isto é, todos acertaram a resposta correta).

(a) Estabeleça as hipóteses para o teste.

(b) Dos 400 estudantes, 125 responderam corretamente a questão. Encontre o valor-p e interprete.

6.22 O Exemplo 6.4 (página 177) testou a terapia para a anorexia usando H_0: $\mu = 0$ e H_a: $\mu > 0$ sobre a alteração média no peso da população.

(a) Nas palavras daquele exemplo, o que seria um (i) erro do Tipo I, (ii) erro do Tipo II?

(b) O valor-p era de 0,017. Se a decisão para $\alpha = 0{,}05$ estava errada, que tipo de erro é esse?

(c) Suponha que, ao contrário, $\alpha = 0{,}01$. Qual a decisão que você iria tomar? Se estiver errado, que tipo de erro é esse?

6.23 Jones e Smith conduziram, separadamente, estudos para testar H_0: $\mu = 500$ contra H_a: $\mu \neq 500$, cada um com $n = 1000$. Jones obtém $\bar{y} = 519{,}5$, com um $ep = 10{,}0$. Smith obtém $\bar{y} = 519{,}7$, com um $ep = 10{,}0$.

(a) Mostre que $t = 1{,}95$ e o valor-$p = 0{,}051$ para Jones. Mostre que $t = 1{,}97$ e o valor-$p = 0{,}049$ para Smith.

(b) Usando $\alpha = 0{,}050$, para cada estudo indique se o resultado é "estatisticamente significativo".

(c) Usando esse exemplo, explique os aspectos equivocados de relatar o resultado de um teste como "$p \leq 0{,}05$" *versus* "$p > 0{,}05$", ou como "rejeite H_0" *versus* "não rejeite H_0" sem relatar o valor-p real.

6.24 Jones e Smith conduziram, separadamente, estudos para testar H_0: $\pi = 0{,}50$ contra H_a: $\pi \neq 0{,}50$, cada uma com $n = 400$. Jones obtém $\hat{\pi} = 220/400 = 0{,}550$. Smith obtém $\hat{\pi} = 219/400 = 0{,}5475$.

(a) Mostre que $z = 2{,}00$ e o valor-$p = 0{,}046$ para Jones. Mostre que $z = 1{,}90$ e o valor-$p = 0{,}057$ para Smith.

(b) Usando $\alpha = 0{,}05$, indique em cada caso se o resultado é "estatisticamente significativo". Interprete.

(c) Use este exemplo para explicar por que uma informação importante é perdida relatando o resultado do teste como "valor-$p \leq 0{,}05$" *versus* "valor-$p > 0{,}05$" ou como "rejeite H_0" *versus* "não rejeite H_0", sem relatar o valor-p.

(d) O intervalo de 95% de confiança para π é (0,501; 0,599) para Jones e (0,499; 0,596) para Smith. Explique por que esse método mostra que, em termos práticos, os dois estudos tinham resultados muito similares.

6.25 Um estudo quer verificar se o escore médio μ de um vestibular de uma faculdade em 2007 não é diferente da média 500 obtida em 1957. Teste H_0: $\mu = 500$ contra H_a: $\mu \neq 500$, se, para uma amostra aleatória, de 10000 estudantes de todo país que fizeram o exame em 2007, $\bar{y} = 497$ e $s = 100$. Mostre que o resultado é estatisticamente altamente significativo, mas não praticamente significativo.

6.26 Um relatório publicado em 25 de setembro de 2006 pela Collaborative on Academic Careers in Higher Education (Colaboradores das Carreiras Acadêmicas na Educação Superior) indicou que existe uma lacuna notável entre os acadêmicos do sexo feminino e masculino na sua confiança de que as regras de estabilidade sejam claras, com os homens sendo mais confiantes. Foi solicitado aos 4500 membros do corpo docente, em um levantamento de dados, que avaliassem essas políticas em uma escala de 1 a 5 (não muito clara a muito clara). A resposta média sobre os critérios de estabilidade foi de 3,51 para mulheres e 3,55 para homens, que foi relatado como estatisticamente significativo, com a média das mulheres sendo menor do que a média dos homens. Use este estudo para explicar a distinção entre significância estatística e significância prática.

6.27 Considere o Exemplo 6.8 sobre as "descobertas médicas" (página 191). Usando um diagrama de árvore, aproxime P(erro do Tipo I) sob a suposição de que um efeito verdadeiro exista 20% das vezes e que P(erro do Tipo II) = 0,30.

6.28 Uma decisão é planejada em um teste de H_0: $\mu = 0$ contra H_a: $\mu > 0$, usando $\alpha = 0{,}05$. Se $\mu = 5$, P(erro do Tipo II) = 0,17.

(a) Explique o significado da última frase.

(b) Se o teste usou $\alpha = 0{,}01$, a P(erro do Tipo II) seria menor, igual ou maior do que 0,17? Explique.

(c) Se $\mu = 10$, a P(erro do Tipo II) seria menor, igual ou maior do que 0,17? Explique.

6.29 Seja π a proporção de esquizofrênicos que respondem positivamente a um tratamento. Um teste é conduzido com H_0: $\pi = 0{,}50$ contra H_a: $\pi > 0{,}50$, para uma amostra de tamanho 25, utilizando $\alpha = 0{,}05$.

(a) Encontre a região dos valores da proporção amostral para a qual H_0 é rejeitada.

(b) Suponha que $\pi = 0{,}60$. Encontre P(erro do Tipo II).

6.30 Estudos têm levado em conta se diferenças existem, quanto ao sexo do recém-nascido, em relação às reações comportamentais e psicológicas ao estresse. Um estudo[6] avaliou mudanças no batimento cardíaco para uma amostra de bebês colocados em uma situação de

estresse. A média amostral da mudança no batimento cardíaco foi pequena para bebês do sexo feminino quando comparada aos bebês do sexo masculino: −1,2 contra 10,7, cada uma com desvios padrão de aproximadamente 18. Suponha que estejamos céticos sobre o resultado dos bebês do sexo masculino e planejamos um experimento maior para testar se a média do batimento cardíaco aumenta quando os bebês do sexo masculino são submetidos a uma experiência estressante. Seja μ a média populacional da diferença dos batimentos cardíacos depois *versus* antes da situação estressante. Testaremos H_0: $\mu = 0$ contra H_a: $\mu > 0$, no nível $\alpha = 0{,}05$ usando $n = 30$ bebês do sexo masculino. Suponha que o desvio padrão é 18. Encontre a P(erro do Tipo II) se $\mu = 10$ mostrando que **(a)** uma estatística-teste de $t = 1{,}699$ tem um valor-p de 0,05, **(b)** que não conseguimos rejeitar H_0 se $\bar{y} < 5{,}6$, **(c)** isso acontece se $\bar{y}$ está a mais do que 1,33 erros padrão abaixo de 10, **(d)** isso acontece com probabilidade aproximada de 0,10.

6.31 Considere o exercício anterior.
- **(a)** Encontre P(erro do Tipo II) se $\mu = 5$. Como P(erro do Tipo II) depende do valor de μ?
- **(b)** Encontre P(erro do Tipo II) se $\mu = 10$ e $\alpha = 0{,}01$. Como P(erro do Tipo II) depende de α?
- **(c)** Como P(erro do Tipo II) depende de n?

6.32 Uma lista contém nomes de todos os indivíduos que possam ser chamados para atuar como jurados. A proporção de mulheres na lista é de 0,53. Um júri de 12 pessoas é selecionado ao acaso da lista. Nenhum dos selecionados é mulher.
- **(a)** Encontre a probabilidade da seleção de zero mulheres na amostra obtida.
- **(b)** Teste a hipótese de que a seleção foi aleatória contra a alternativa de tendenciosidade contra as mulheres. Relate o valor-p e interprete.

6.33 Uma pessoa afirmando possuir percepção extrassensorial (PES) diz que pode estimar em um maior número de vezes do que não estimar o resultado de uma moeda honesta lançada em outra sala, não visível a ela.
- **(a)** Introduza uma notação apropriada e declare as hipóteses para testar essa afirmação.
- **(b)** De 5 lançamentos de moedas, ela estima o resultado correto 4 vezes. Encontre o valor-p e interprete.

6.34 Em uma pesquisa de boca de urna da CNN de 1336 eleitores na eleição para o Senado de 2006 no estado de Nova Iorque, seja x = o número de eleitores de boca de urna da candidata Democrata, Hillary Clinton.
- **(a)** Explique por que este cenário iria satisfazer as três condições necessárias para usar a distribuição binomial.
- **(b)** Se a proporção da população que votou em Clinton tivesse sido 0,50, encontre a média e o desvio padrão da distribuição da probabilidade de x.
- **(c)** Para (b), usando a aproximação da distribuição normal, dê um intervalo no qual x certamente iria estar.
- **(d)** Na verdade, a pesquisa de boca de urna forneceu $x = 895$. Explique como você poderia fazer uma inferência sobre se π está acima ou abaixo de 0,50.

6.35 Em um dado ano, a probabilidade de que uma mulher norte-americana morra em um acidente de carro é igual a 0,0001 (*Statistical Abstract for the United States* – Resumo estatístico para os Estados Unidos).
- **(a)** Em uma cidade com 1 milhão de mulheres, encontre a média e o desvio padrão de x = número de mortes de mulheres em acidentes de carro. Determine as suposições para esse estudo ser válido. (*Dica:* encontre μ e σ para a distribuição binomial.)
- **(b)** Seria surpreendente se $x = 0$? Explique. (*Dica:* quantos desvios padrão 0 está do valor esperado?)
- **(c)** Baseado na aproximação normal para a binomial encontre um intervalo dentro do qual x tenha a probabilidade de 0,95 de ocorrer.
- **(d)** A probabilidade para homens norte-americanos é de 0,0002. Repita (a) para homens e compare os resultados aos das mulheres.

Conceitos e aplicações

6.36 Você pode usar um *applet* para gerar repetidamente amostras aleatórias e conduzir testes de significância, para ilustrar seu comportamento quando usada para muitas amostras. Para tentar isto, vá para o *applet* de *testes de significância para uma proporção* em www.grupoa.com.br. Determine a hipótese nula para H_0: $\pi = 1/3$ para um teste unilateral ($\pi > 1/3$) com tamanho da amostra de 116, caso do Exemplo 6.9 (página 192) do experimento em astrologia. No menu, determine o valor da proporção verdadeira para 0,33.
- **(a)** Clique em [*Simulate*] e 100 amostras deste tamanho serão geradas, com o valor-p encontrado para cada amostra. Que percentual de testes foram significantes no nível de significância 0,05?
- **(b)** Para ter uma ideia do que acontece "a longo prazo" repita essa simulação 50 vezes, assim você terá um total de 5000 amostras do tamanho 116. Que percentual das amostras resultou em um erro do Tipo I? Que percentual você esperaria que resultasse em um erro do Tipo I?
- **(c)** A seguir, mude a proporção para 0,50, assim H_0 é, na verdade, falsa. Simule 5000 amostras. Qual o percentual de vezes que você cometeu o erro do Tipo II? Pelo Exemplo 6.9 isto deveria acontecer aproximadamente 2% das vezes.

6.37 Considere o arquivo de dados *Student Survey* (Exercício 1.11 da página 25).
- **(a)** Teste se a média populacional da ideologia política difere de 4,0. Relate o valor-p e interprete.
- **(b)** Teste se a proporção em favor da legalização do aborto é igual ou difere de 0,50. Relate o valor-p e interprete.

6.38 Considere o arquivo de dados que a sua turma criou no Exercício 1.12 (página 26). Para as variáveis escolhidas pelo seu professor, faça uma pergunta de pesquisa e conduza uma análise de estatística inferencial. Use, também, métodos gráficos e numéricos apresentados anteriormente, neste livro, para descrever os dados e, se necessário, verifique as suposições para a sua análise. Prepare um relatório resumindo e interpretando suas descobertas.

6.39 Um estudo considerou os efeitos de uma classe especial designada para melhorar as habilidades verbais das crianças. Cada criança fez um teste verbal antes e depois de frequentar a classe por três semanas. Seja y = o escore do segundo exame – o escore do primeiro exame. Os escores em y para uma amostra aleatória de quatro crianças tendo problemas de aprendizagem foram 3, 7, 3, 3. Conduza métodos estatísticos inferenciais para determinar se a turma tem um efeito positivo. Resuma suas análises e interpretações em um breve relatório. (*Nota:* os escores podem melhorar simplesmente pelo fato de os estudantes sentirem-se mais confortáveis com o processo do teste. Um delineamento mais apropriado iria, também, administrar o exame duas vezes a um grupo de controle que não frequentava a classe especial, comparando as mudanças para o grupo experimental e de controle usando métodos do Capítulo 7.)

6.40 Os 49 estudantes de uma turma da Universidade da Flórida fizeram avaliações cegas de pares de refrigerantes do tipo cola. Para as 49 comparações entre a Coca-Cola e a Pepsi-Cola, a Coca-Cola foi a preferida 29 vezes. Na população que esta amostra representa, isto é uma forte evidência de que a maioria prefere um dos dois refrigerantes? Considere a seguinte saída do computador:

```
Teste do parâmetro igual a 0,5 versus
diferente de 0,5

N   Proporção   IC de      Valor-z   Valor-p
    Amostral   95%
49  0,5918     (0,454;    1,286     0,1985
                0,729)
```

Explique como cada resultado nesta saída do computador foi obtido. Resuma os resultados em uma forma que seria claro para alguém que não está familiarizado com a inferência estatística.

6.41 Nos anos de 1990, o U.S. Justice Department (Departamento de Justiça dos Estados Unidos) e outros grupos estudaram possíveis abusos de policiais da Filadélfia no seu relacionamento com as minorias. Um estudo conduzido pela American Civil Liberties Union (União Americana de Liberdades Civis) analisou se motoristas afro-americanos tinham uma probabilidade maior do que outros na população de serem parados pela polícia enquanto dirigiam. Os pesquisadores estudaram os resultados de 262 abordagens de motoristas por policiais durante uma semana em 1997. Destas, 207 dos motoristas eram afro-americanos, ou 79% do total. Naquele período, a população da Filadélfia consistia em 42,2% de afro-americanos. O número de afro-americanos parados dá forte evidência de uma possível tendenciosidade, sendo mais alto do que você esperava se levarmos em consideração uma variação apenas aleatória? Explique seu raciocínio em um relatório com, no máximo, 250 palavras.

6.42 Um experimento com 26 estudantes em uma sala de aula de Israel consistia em dar a todos cartões da loteria e, então, perguntar se eles gostariam de trocar seu cartão por outro mais um pequeno incentivo monetário. Somente 7 estudantes concordaram com a troca. Em um experimento separado, 31 estudantes receberam uma caneta nova e, então mais tarde, foi perguntados se queriam trocá-la por outra caneta e um pequeno incentivo monetário. Todos os 31 concordaram.[7] Conduza métodos de estatística inferencial para analisar os dados. Resuma sua análise e interpretação em um pequeno relatório.

6.43 Em condições ideais, os resultados de uma análise estatística não deveriam depender grandemente de uma única observação. Para verificar isto, é uma boa ideia realizar uma **análise de sensibilidade**: refazer a análise após eliminar um valor atípico do conjunto de dados ou mudar o seu valor para um mais típico e verificar se os resultados mudam muito. Para os dados da anorexia mostrados no Exemplo 5.5 (página 144), a mudança de peso de 20,9 libras foi um valor atípico severo. Suponha que esta observação foi, na verdade, 2,9 libras, mas foi registrada incorretamente. Refaça o teste unilateral do Exemplo 6.4 (página 177) e analise a influência daquela observação.

6.44 Tomando uma decisão em um teste, um pesquisador se preocupa sobre a possibilidade de rejeitar H_0 quando, na verdade, ela é verdadeira. Explique como controlar a probabilidade deste tipo de erro.

6.45 Considere a analogia entre tomar uma decisão em um teste e tomar uma decisão sobre a inocência ou a culpa de um acusado num processo criminal.

(a) Explique que erro do Tipo I e erro do Tipo II estão no julgamento.

(b) Explique intuitivamente por que, diminuindo P(erro do Tipo I), aumenta P(erro do Tipo II).

(c) Os acusados são condenados se o júri os considera culpados "além de uma dúvida razoável". Um júri interpreta isto significando que, se o acusado for inocente, a probabilidade de ser culpado deveria ser somente de uma em um bilhão. Descreva todos os problemas que esta estratégia tem.

6.46 Os testes médicos para o diagnóstico de condições como câncer de mama são falíveis, como as decisões em testes de significância. Identifique (H_0 verdadeira, H_0 falsa) com doença (ausente, presente) e (Rejeite H_0, Não rejeite H_0) com teste de diagnóstico (positivo, negativo), onde um diagnóstico positivo significa que o teste prevê que a doença está presente. Explique a diferença entre os erros do Tipo I e II nesse contexto. Explique por que, diminuindo P(erro do Tipo I), aumenta a P(erro do Tipo II) nesse contexto.

6.47 Um artigo em um periódico da sociologia que trata das mudanças nas crenças religiosas ao longo do tempo declara: "Para esses sujeitos, a diferença entre as suas respostas médias na escala da religiosidade entre a idade de 16 anos e o levantamento de dados atual era significativa ($p < 0,05$)".

(a) Explique o que significa o resultado ser "significativo".

(b) Explique por que seria mais informativo se os autores fornecessem o valor-p real do que meramente indicar que ele está abaixo de 0,05. Que outra informação eles poderiam ter fornecido?

6.48 Um artigo em um periódico de ciência política declara que "nenhuma diferença significativa foi encontrada entre homens e mulheres nas suas taxas de voto ($p = 0,63$)". Podemos concluir que as taxas de voto da população são idênticas para homens e mulheres? Explique.

6.49 Você conduz um teste de significância usando um *software*. A saída apresenta um valor-p de 0,4173545. Resumindo sua análise em um artigo de pesquisa, explique por que faz mais sentido relatar o valor 0,42 em vez de 0,4173545.

6.50 Uma pesquisa conduz 60 testes de significância. Destes, 3 foram significativos no nível 0,05. Os autores escrevem um relatório ressaltando somente os três resultados "significativos", não mencionando os outros 57 testes que foram "não significativos". Explique o que está equivocado nesse relatório.

6.51 Alguns periódicos têm a política de publicar resultados de pesquisa somente se eles obtêm significância estatística no nível $\alpha = 0,05$.

(a) Explique o perigo disto.

(b) Quando estórias médicas nos meios de comunicação relatam supostos perigos ou benefícios de certos agentes (por exemplo, consumo do café, fibra no cereal), pesquisas posteriores geralmente sugerem que os efeitos são menores do que primeiro se acreditava ou podem, até mesmo, não existir. Explique por que.

Selecione a(s) resposta(s) correta(s) nos Exercícios 6.52 a 6.56. (Mais de uma resposta pode estar correta.)

6.52 Analisamos se a média verdadeira da liberação de água poluída por hora de uma fábrica industrial excede a afirmação da empresa de 1000 galões. Para a decisão em um teste unilateral utilizando $\alpha = 0,05$:

(a) Se a fábrica não estiver excedendo o limite, mas na verdade $\mu = 1000$, existe somente 5% de chance de que iremos concluir que eles estão excedendo o limite.

(b) Se a fábrica estiver excedendo o limite, existe somente uma chance de 5% de que iremos concluir que eles não estão excedendo o limite.

(c) A probabilidade de que a média amostral é igual ao valor observado seria igual a 0,05 se H_0 fosse verdadeira.

(d) Se rejeitarmos H_0, a probabilidade de que ela seja realmente verdadeira é 0,05.

(e) Todas as respostas acima estão corretas.

6.53 O valor-p para testar H_0: $\mu = 100$ contra H_a: $\mu \neq 100$ é 0,001. Isso indica que:

(a) Existe uma forte evidência de que $\mu = 100$.

(b) Existe uma forte evidência de que $\mu \neq 100$.

(c) Existe uma forte evidência de que $\mu > 100$.

(d) Existe uma forte evidência de que $\mu < 100$.

(e) Se μ fosse igual a 100, seria incomum obter dados como os observados.

6.54 No exercício anterior, suponha que a estatística-teste $t = 3,29$.

(a) Existe uma forte evidência de que $\mu = 100$.

(b) Existe uma forte evidência de que $\mu > 100$.

(c) Existe uma forte evidência de que $\mu < 100$.

6.55 Um intervalo de 95% de confiança para μ é (96; 110). Quais das duas afirmações

sobre testes de significância para os mesmos dados estão corretas?

(a) Testando H_0: $\mu = 100$ contra H_a: $\mu \neq 100$, $p > 0{,}05$.

(b) Testando H_0: $\mu = 100$ contra H_a: $\mu \neq 100$, $p < 0{,}05$.

(c) Testando H_0: $\mu = \mu_0$ contra H_a: $\mu \neq \mu_0$, $p > 0{,}05$ se μ_0 é qualquer um dos números dentro do intervalo de confiança.

(d) Testando H_0: $\mu = \mu_0$ contra H_a: $\mu \neq \mu_0$, $p > 0{,}05$ se μ_0 é qualquer um dos números fora do intervalo de confiança.

6.56 Seja β a representação de P(erro do Tipo II). Para um teste com nível $\alpha = 0{,}05$ de H_0: $\mu = 0$ contra H_a: $\mu > 0$ com $n = 30$ observações, $\beta = 0{,}36$ em $\mu = 4$. Então:

(a) Em $\mu = 5$, $\beta > 0{,}36$.

(b) Se $\alpha = 0{,}01$, então em $\mu = 4$, $\beta > 0{,}36$.

(c) Se $n = 50$, então em $\mu = 4$, $\beta > 0{,}36$.

(d) O poder do teste é 0,64 em $\mu = 4$.

(e) Isso deve ser falso porque necessariamente $\alpha + \beta = 1$.

6.57 Responda verdadeiro ou falso para cada uma das seguintes afirmações e explique a sua resposta:

(a) P(erro do Tipo II) = 1 − P(erro do Tipo I).

(b) Se rejeitarmos H_0 usando $\alpha = 0{,}01$, então também a rejeitamos usando $\alpha = 0{,}05$.

(c) O valor-p é a probabilidade de que H_0 é verdadeira (*Dica:* encontramos as probabilidades sobre as variáveis e suas estatísticas ou sobre os parâmetros?)

(d) Um artigo num periódico sobre antropologia relata $p = 0{,}063$ para testar H_0: $\mu = 0$ contra H_a: $\mu \neq 0$. Se os autores tivessem, ao contrário, relatado um intervalo de 95% de confiança para μ, então o intervalo conteria 0 e os leitores poderiam ter julgado melhor quais valores são plausíveis para μ.

6.58 Explique a diferença entre as hipóteses alternativas unilaterais e bilaterais e explique como isto afeta o cálculo do valor-p.

6.59 Explique por que a terminologia "não rejeitar H_0" é preferível a "aceitar H_0".

6.60 Seu amigo planeja fazer um levantamento de dados com os estudantes da sua faculdade para verificar se a maioria acha que a idade legal para o consumo de álcool deveria ser reduzida. Ele nunca estudou estatística. Como você iria explicar para ele os conceitos de:

(a) hipótese nula e alternativa,

(b) valor-p,

(c) nível α,

(d) erro do Tipo II?

6.61 Uma amostra aleatória de tamanho 40 tem $\bar{y} = 120$. O valor-p para testar H_0: $\mu = 100$ contra H_a: $\mu \neq 100$ é 0,057. Explique o que está incorreto em cada uma das seguintes interpretações deste valor-p e forneça uma interpretação apropriada.

(a) A probabilidade de que a hipótese nula esteja correta é de 0,057.

(b) A probabilidade de que $\bar{y} = 120$ se H_0 é verdadeira é igual a 0,057.

(c) Se, na verdade, $\mu \neq 100$, a probabilidade é igual a 0,057 de que os dados seriam, pelo menos, contraditórios para H_0 como os dados observados.

(d) A probabilidade de um erro do Tipo I é igual a 0,057.

(e) Aceitamos H_0 no nível $\alpha = 0{,}05$.

(f) Podemos rejeitar H_0 no nível $\alpha = 0{,}05$.

***6.62** Considere o exercício anterior e o valor-p de 0,057.

(a) Explique por que o valor-p é o menor nível α no qual H_0 pode ser rejeitada; isto é, ele é igual ao menor nível no qual os dados são significativos.

(b) Considere a correspondência entre os resultados dos intervalos de confiança e os testes bilaterais. Quando o valor-p é 0,057, explique por que o intervalo de 94,3% de confiança é o intervalo de confiança mais estreito para μ que contém $\mu_0 = 100$.

***6.63** Uma pesquisadora conduz um teste de significância cada vez que ela analisa um novo conjunto de dados. Ao longo do tempo, ela conduz 100 testes.

(a) Suponha que H_0 é verdadeira em cada caso. Qual é a distribuição do número de vezes que ela rejeita H_0 no nível 0,05?

(b) Suponha que ela rejeite H_0 em cinco testes. É plausível que H_0 esteja correta em todos os casos? Explique.

***6.64** A cada ano em Liverpool, Nova Iorque, uma bibliotecária de uma instituição pública estima o número médio de vezes em que os livros daquela biblioteca foram retirados no ano anterior. Para fazer isto, a bibliotecária amostra aleatoriamente registros do computador de 100 livros e forma um intervalo de 95% de confiança para a média. Isto tem sido feito por 20 anos.

(a) Encontre a probabilidade de que todos os intervalos de confiança contenham as médias verdadeiras. (*Dica:* use a distribuição binomial.)

(b) Encontre a probabilidade de que pelo menos um intervalo de confiança não contenha a média verdadeira.

***6.65** Suponha que você queira testar H_0: $\pi = 0{,}50$, mas das $n = 30$ observações, 0 estavam na categoria de interesse. Se você encontrou a estatística-teste z utilizando o $ep = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ para intervalos de confiança, mostre o que acontece à estatística-teste. Explique por que o $ep_0 = \sqrt{\pi_0(1 - \pi_0)/n}$ é mais apropriado para o teste.

***6.66** Você testa H_0: $\pi = 0{,}50$ contra H_a: $\pi > 0{,}50$ usando $\alpha = 0{,}05$. Na verdade, H_a é verdadeira. Explique por que a P(erro do Tipo II) aumenta em direção a 0,95 à medida que π se move na direção de 0,50. (Assuma que n e α permaneçam fixos.)

***6.67** Considere o experimento da PES do Exercício 6.33 com $n = 5$.

(a) Para qual(is) valor(es) de x = número de estimativas corretas, você pode rejeitar H_0: $\pi = 0{,}50$ a favor de H_a: $\pi > 0{,}50$, usando $\alpha = 0{,}05$?

(b) Para qual(is) valor(es) de x você pode rejeitar H_0 usando $\alpha = 0{,}01$? (*Nota*: para amostras pequenas pode não ser possível alcançar valores-p muito pequenos.)

(c) Suponha que você teste H_0 usando $\alpha = 0{,}05$. Se $\pi = 0{,}50$, qual é a P(erro do Tipo I)? (*Nota:* para distribuições discretas P(erro do Tipo I) pode ser menor do que o pretendido. É melhor relatar o valor-p.)

☑ NOTAS

[1] *Tampa Tribune*, 6 de abril de 1996.

[2] *New York Times*, 7 de fevereiro de 2007.

[3] www.fiu.edu/orgs/ipor/ffp.

[4] STERNE, J., SMITH, G., COX, D.R. *British Medical Journal*, v. 322, p. 226-31, 2001.

[5] CARLSON, S. *Nature*, v. 318, p. 419-25, 1985.

[6] DAVIS, M, EMORY, E. *Child Development*, v. 66, p. 14-27, 1995.

[7] BAR-HILLEL, M., NETER, E. J. *Personality and Social Psychology*, v. 70, p. 17-27, 1996.