anos ou mais. Ele coleta uma amostra aleatória de 500 doadores a partir dos registros. Dos dados fornecidos, construa um intervalo de confiança de 95% para a proporção de doadores que têm 50 anos ou mais.

53. **Seguro saúde.** Com base em um levantamento de dados feito em 2007 em lares norte-americanos (veja www.census.gov), 87% (de 3060) dos homens em Massachusetts (MA) tinham seguro saúde.

a) Examine as condições para a construção de um intervalo de confiança para a proporção de homens de MA que tinha seguro saúde.
b) Encontre o intervalo de confiança de 95% para o percentual de homens com seguro saúde.
c) Interprete seu intervalo de confiança.

54. **Seguro saúde, parte 2.** Usando o mesmo levantamento e dados do Exercício 53, encontramos que 84% dos respondentes de Massachusetts que se definiam negros/afro-americanos (dentre 440) tinham seguro saúde.

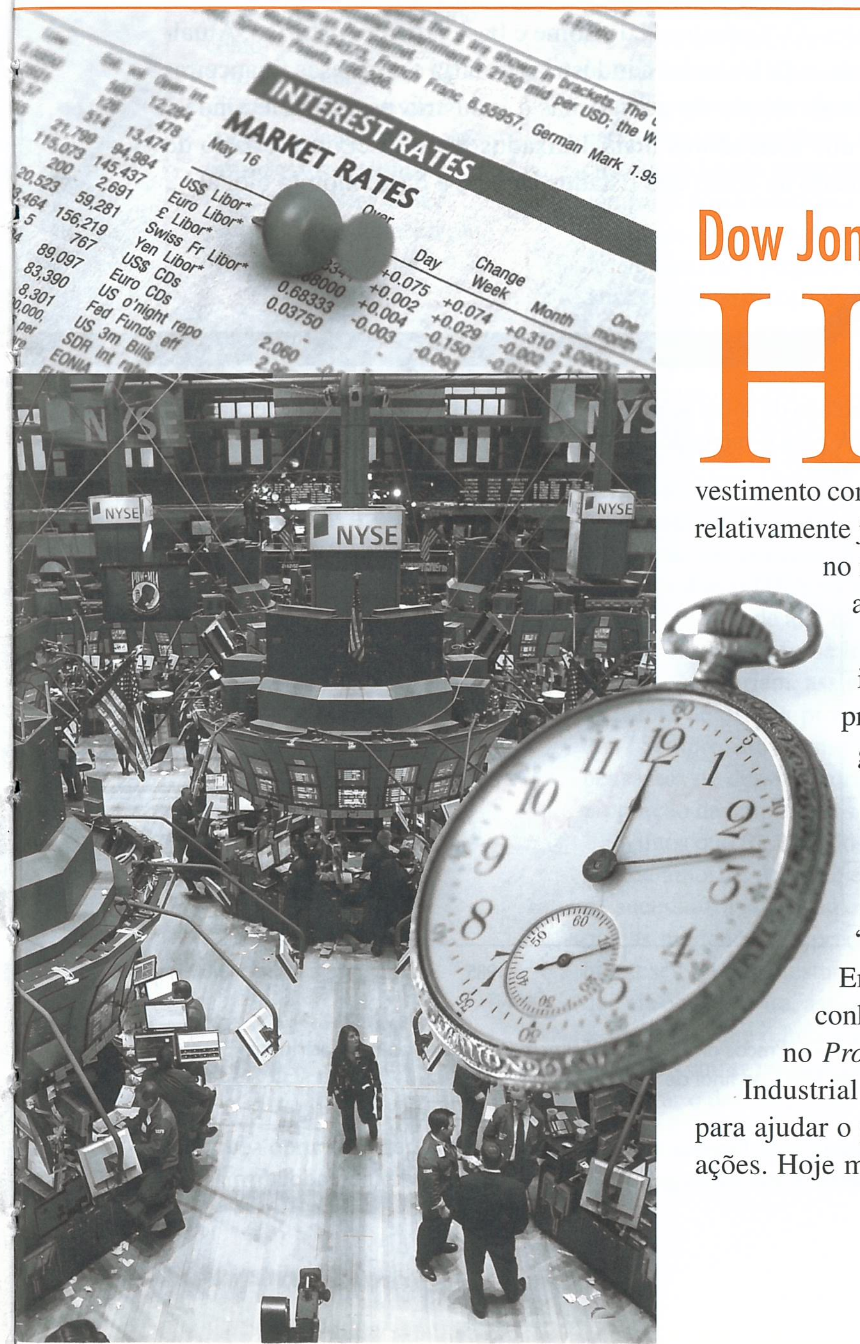
a) Examine as condições para a construção de um intervalo de confiança para a proporção de negros/afro-americanos de MA que tinham seguro saúde.
b) Encontre o intervalo de confiança de 95%.
c) Interprete seu intervalo de confiança.

RESPOSTAS DO TESTE RÁPIDO

1 Maior.
2 Mais baixo.
3 Menor.

Capítulo 11

Testando Hipóteses sobre Proporções



Dow Jones Industrial Average

Há mais de cem anos, Charles Dow mudou a maneira como as pessoas observam o mercado de ações. Surpreendentemente, ele não era um perito em investimentos ou um financista de alto risco. Era um jornalista que queria tornar o investimento compreensível para os leigos. Embora tenha falecido relativamente jovem, com 51 anos, em 1902, o impacto de Dow no modo como nos informamos sobre o mercado de ações tem sido duradouro e de longo alcance.

No final do século XIX, quando Charles Dow informava sobre a Wall Street, os investidores preferiam títulos, em vez de ações. Os títulos eram garantidos pelas máquinas e por outros ativos que a empresa tinha. Também eram confiáveis: o proprietário sabia quando o título atingiria seu máximo, assim, sabia quando e quanto o título pagaria. As ações simplesmente representavam "cotas" de posse, que eram arriscadas e instáveis. Em maio de 1896, Dow e Edward Jones, o qual ele conhecia desde quando trabalhavam como repórteres no *Providence Evening Press*, lançaram o Dow Jones Industrial Average (DJIA) (Média Indistrial Dow Jones) para ajudar o público a entender as tendências do mercado de ações. Hoje mundialmente conhecido, o DJIA original calculava a média de 11 preços de ações. Dessas ações industriais, somente a General Electric ainda está no DJIA atualmente.

Desde então, o DJIA se tornou sinônimo de desempenho geral do mercado e é geralmente chamado de Dow. O índice foi ampliado para 20 ações em 1916 e para 30 em 1928, no auge da euforia do mercado dos anos 1920. A bolsa em alta teve o seu pico no dia 3 de setembro de 1929, quando o índice Dow atingiu 381,17. Em 28 e 29 de outubro de 1929, o índice perdeu aproximadamente 25% do seu valor. A situação piorou. Em quatro anos, em 8 de julho de 1932, as 30 ações industriais atingiram 40,65, a maior baixa de todos os tempos. A elevação de setembro de 1929 só foi atingida novamente em 1954.

Hoje, o índice Dow é uma média ponderada de 30 ações industriais, com medidas usadas para explicar cisões e outros ajustes. O "Industrial" do nome é basicamente histórico. Atualmente, o DJIA inclui a indústria terciária e empresas financeiras e é mais amplo do que apenas a indústria pesada. Ele ainda é um dos indicadores mais utilizados para observar o estado do mercado de ações dos Estados Unidos e da economia global.

Quem	Dias em que o mercado de ações estava aberto ("dias de negócios da bolsa")
O quê	Preço de fechamento da Dow Jones Industrial (*Fechamento*)
Unidades	Pontos
Quando	Agosto de 1982 a dezembro de 1986
Por quê	Para testar a teoria do comportamento do mercado de ações

Como o mercado de ações se movimenta? Aqui estão os preços de fechamento do DJIA para o mercado em alta que ocorreu da metade de 1982 até o final de 1986.

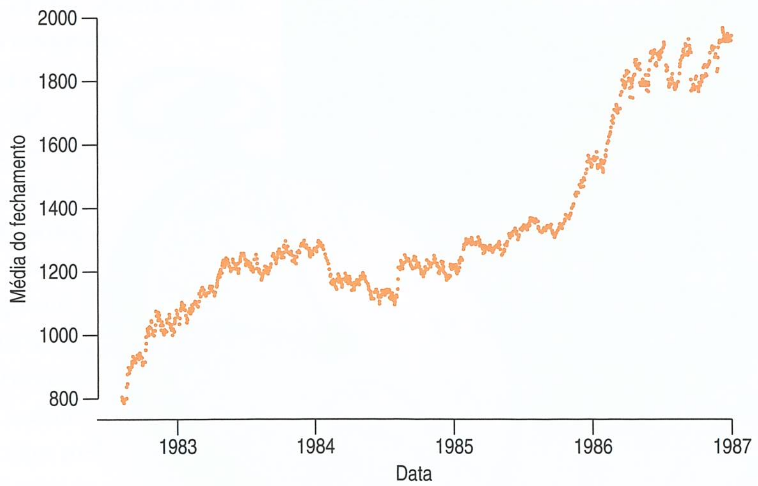


Figura 11.1 Preços diários do fechamento do Dow Jones Industrial da metade de 1982 ao final de 1986.

O DJIA cresceu durante essa famosa euforia da bolsa, mais que dobrando seu valor em menos de cinco anos. Uma teoria geral do comportamento do mercado afirma que, em um dado dia, é provável que o mercado se mova tanto para cima quanto para baixo. Em outras palavras, o comportamento diário do mercado é aleatório. Isso pode ser verdade durante os períodos de crescimento óbvio? Vamos investigar qual é a chance de o índice Dow subir ou descer em um determinado dia. Dentre os 1112 dias de negócios na bolsa naquele período, a média aumentou em 573 dias, uma proporção amostral de 0,5153 ou 51,23%. *São* mais dias "acima" que dias "abaixo", mas esse valor está longe o suficiente dos 50% para lançar dúvidas sobre a hipótese de um movimento igualmente provável para cima ou para baixo?

11.1 Hipóteses

Hipótese (s):
pl. (Hipóteses)
Uma suposição; uma proposição ou princípio que é suposto ou assumido como verdade, de modo a tirar uma conclusão ou inferência para provar o ponto em questão; algo não provado, mas assumido como verdadeiro com o propósito de argumentação.
— Dicionário Webster, 1913.

ALERTA DE NOTAÇÃO:
A letra maiúscula H é a letra padrão para as hipóteses. H_0 representa a hipótese nula e H_A representa a hipótese alternativa.

Como podemos enunciar e testar uma hipótese sobre as mudanças diárias no DJIA? As hipóteses são modelos de trabalho que adotamos temporariamente. Para testar se é igualmente provável que as flutuações diárias subam ou desçam, assumimos que isso ocorre e que qualquer diferença aparente com relação ao valor de 50% é apenas uma flutuação aleatória. Portanto, nossa hipótese inicial, chamada de hipótese nula, é que a proporção dos dias nos quais o DJIA aumenta é igual a 50%. A **hipótese nula**, que denotamos H_0, especifica um parâmetro do modelo populacional e propõe um valor para ele. Geralmente, escrevemos a hipótese nula sobre uma proporção na forma H_0: $p = p_0$. Essa é uma maneira concisa de especificar os dois fatores de que mais precisamos: a identidade do parâmetro que esperamos conhecer (a verdadeira proporção) e o valor hipotético específico para aquele parâmetro (nesse caso, 50%). Precisamos de um valor hipotético para comparar nossa estatística observada com ele. Qual valor será utilizado para a hipótese não é uma questão estatística. Pode ser óbvio a partir do contexto dos dados, mas, às vezes, interpretar a pergunta que esperamos responder em uma hipótese sobre o parâmetro demanda algum tempo. Para a nossa hipótese sobre se o DJIA se move para cima ou para baixo com a mesma probabilidade, está bem claro que precisamos testar H_0: $p = 0{,}5$.

A **hipótese alternativa,** que denotamos H_A, contém os valores do parâmetro que consideramos plausíveis se rejeitarmos a hipótese nula. No nosso exemplo, a hipótese nula é que a proporção *p*, de dias "acima", é 0,5. Qual é a alternativa? Durante um período de alta da bolsa, você deve esperar mais dias em alta do que em baixa, mas iremos assumir que estamos interessados num desvio em qualquer direção, portanto, a alternativa é H_A: $p \neq 0{,}5$.

O que nos convenceria de que a proporção de dias acima não é de 50%? Se em 95% dos dias o DJIA fechou acima, a maioria das pessoas estaria convencida de que a probabilidade de alta não é a mesma da de baixa. No entanto, se a proporção amostral de dias acima fosse levemente mais alta que 50%, você ficaria cético. Afinal, as observações variam, assim, não ficaríamos surpresos em ver alguma diferença. Quão diferente de 50% a proporção deve ser para *ficarmos* convencidos de que ela mudou? Sempre que questionamos o tamanho de uma diferença estatística, naturalmente pensamos no desvio padrão. Assim, vamos começar encontrando o desvio padrão da proporção da amostra dos dias em que o DJIA aumentou.

Vimos que 51,53% dos dias, dentre os 1112, apresentaram altas da bolsa. O tamanho da amostra de 1112 é certamente grande o suficiente para satisfazer a condição de sucesso/fracasso (esperamos $0{,}50 \times 1112 = 556$ aumentos diários). Suspeitamos que as mudanças do preço diário sejam aleatórias e independentes. Sabemos quais hipóteses estamos testando. Para testar uma hipótese, *assumimos* (temporariamente) que ela é verdadeira, assim, podemos ver se aquela descrição do mundo é plausível. Se assumirmos que o índice Dow aumenta ou diminui com a mesma probabilidade, precisamos centrar nosso modelo amostral Normal numa média de 0,5. Então, podemos encontrar o desvio padrão do modelo amostral como:

$$DP(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0{,}5)(1-0{,}5)}{1112}} = 0{,}015$$

◆ **Por que esse valor é um desvio padrão e não um erro padrão?** Porque não estimamos nada. Quando assumimos que a hipótese nula é verdadeira, ela fornece um valor para o parâmetro do modelo, p. Com proporções, se conhecemos p, então automaticamente conhecemos seu desvio padrão. Como encontramos o desvio padrão a partir do modelo do parâmetro e não da sua estimativa, ele é um desvio padrão e não um erro padrão. Quando encontramos um intervalo de confiança para p, não podíamos assumir que sabíamos seu valor, assim, estimamos o valor de p e, consequentemente, do desvio padrão do valor amostral, $\hat{p}$.

Para lembrarmos que o valor do parâmetro provém da hipótese nula, às vezes ele é escrito como p_0 e o desvio padrão como:

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$$

Agora conhecemos ambos os parâmetros do modelo Normal para a nossa hipótese nula. Para a média, μ, usamos $p = 0{,}50$, e para σ, usamos o desvio padrão das proporções amostrais $DP(\hat{p}) = 0{,}015$. Queremos descobrir quão provável seria encontramos o valor observado $\hat{p}$ tão longe de 50% quanto o valor obtido de 51,53% que observamos. Observando em primeiro lugar a Figura 11.2, podemos ver que 51,53% não é um valor surpreendente. Uma resposta mais próxima (obtida por uma calculadora, programa de computador ou pela tabela Normal) é que esse valor é 0,308. Essa é a probabilidade de observarmos mais de 51,53% dias de alta (ou mais de 51,53% dias de baixa) se o modelo nulo fosse verdadeiro. Em outras palavras, se a probabilidade de um dia em alta (baixa) para o índice Dow é 50%, esperamos ver intervalos de 1112 dias de negócios na bolsa com até 51,53% dias de alta aproximadamente 15,4% do tempo e com até 51,53% dias de baixa aproximadamente 15,4% do tempo também. Como isso não é incomum, não há realmente evidências convincentes de que o mercado não tenha se comportado aleatoriamente.

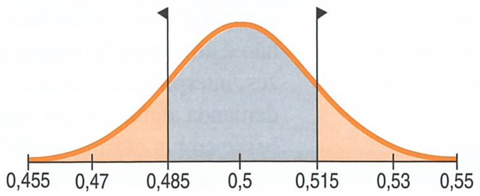
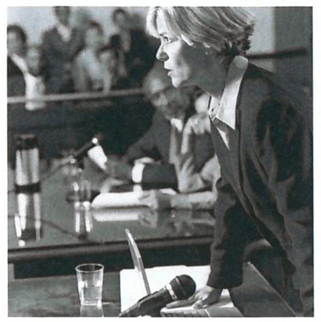


Figura 11.2 Quão provável é uma proporção de mais de 51,5% ou menos de 48,5% quando a média verdadeira é 50%? Veja como ela se parece. Cada área laranja representa 0,154 da área total sob a curva.

Talvez seja surpreendente saber que, mesmo durante a alta, a direção dos movimentos diários da bolsa é aleatória. Porém, a probabilidade de que um dia qualquer termine em alta ou em baixa parece ser aproximadamente 0,5, independentemente das tendências de longo prazo. Talvez o mercado de ações tenha uma alta de longo prazo (ou baixa, embora não tenhamos verificado esse fato), não por ter mais dias de valores altos ou baixos, mas porque as quantidades reais de altas e baixas sejam desiguais.

11.2 Um julgamento como um teste de hipóteses



Começamos assumindo que a probabilidade de um dia de alta fosse de 50%. Depois, analisamos os dados e concluímos que estávamos corretos, já que a proporção que observamos não estava longe o suficiente dos 50%. Esse raciocínio dos testes da hipótese parece inverso? Talvez tenhamos essa impressão porque geralmente preferimos pensar em como conseguir fazer as coisas certas, em vez de erradas. Mas você já viu esse raciocínio em um contexto diferente. Essa é a lógica de um processo judicial.

Vamos supor que um réu tenha sido acusado de roubo. Na lei britânica comum e nos sistemas derivados dela (incluindo a lei norte-americana), a hipótese nula é que o réu é inocente. As instruções dadas aos jurados são bastante explícitas quanto a isso.

As evidências tomam a forma dos fatos que parecem contradizer a suposição de inocência. Para nós, isso significa coletar dados. No julgamento, o promotor público apresenta as provas ("Se o réu fosse inocente, não seria incrível que a polícia o tenha encontrado na cena do crime com uma sacola cheia de dinheiro na mão, uma máscara no rosto e um carro para a fuga estacionado no lado de fora?"). O próximo passo é julgar as provas. Avaliá-las é de responsabilidade do júri num julgamento, mas é de nossa responsabilidade testar a hipótese. O júri considera as provas à luz da *suposição* de inocência e julga se as provas contra o réu seriam plausíveis *se o réu fosse de fato inocente*.

Como o júri, perguntamos: "Esses dados poderiam ter acontecido ao acaso se a hipótese nula fosse verdadeira?". Se fosse improvável que eles tivessem ocorrido, a evidência levanta uma dúvida razoável sobre a hipótese nula. Finalmente, *você* deve tomar uma decisão. O padrão de "além de uma dúvida razoável" é intencionalmente ambíguo, porque leva o júri a decidir até que ponto as provas contradizem a hipótese de inocência. Os júris não usam explicitamente a probabilidade para ajudá-los a decidir se rejeitam aquela hipótese. No entanto, quando você faz a mesma pergunta com relação à sua hipótese nula, tem a vantagem de ser capaz de quantificar exatamente quão surpreendente a evidência seria caso a hipótese nula fosse verdadeira.

Quão improvável é improvável? Algumas pessoas estabelecem padrões rígidos. Níveis como 1 em 20 (0,05) ou 1 de 100 (0,01) são comuns. No entanto, se *você* tiver de tomar a decisão, deve julgar cada situação e analisar se a probabilidade da ocorrência daqueles dados é pequena o suficiente para constituir uma "dúvida razoável".

11.3 Valores-P

O passo fundamental do nosso raciocínio é a pergunta: "Os dados são surpreendentes, dada a hipótese nula?". E o cálculo-chave é determinar exatamente quão prováveis seriam os dados observados caso a hipótese nula fosse o modelo verdadeiro do mundo. Portanto, precisamos de uma *probabilidade*. Especificamente, queremos encontrar a probabilidade de ver dados como esses (ou algo ainda menos prováveis) *dada* a hipótese nula. Essa probabilidade é o valor no qual baseamos nossa decisão; portanto, ela recebe um nome especial: **valor-P**.

Além da dúvida razoável

Perguntamos se os dados eram improváveis além da dúvida razoável. Já calculamos aquela probabilidade. A probabilidade de que o valor da estatística observada (ou até mesmo um valor mais extremo) poderia acontecer se o modelo nulo fosse verdadeiro – neste caso, 0,308 – é o valor-P.

Um valor-P baixo o suficiente indica que os dados observados seriam muito improváveis caso nossa hipótese nula fosse verdadeira. Começamos com um modelo e agora esse modelo nos diz que os dados que temos são improváveis de terem acontecido. É surpreendente. Nesse caso, o modelo e os dados estão em contradição uns com os outros, por isso temos de fazer uma escolha. Ou a hipótese nula está correta, e acabamos de ver algo fora do comum, ou a hipótese nula está errada (e, de fato, estávamos errados em usá-la como a base para calcular o nosso valor-P). Se você acredita mais em dados do que em suposições, então quando encontrar um valor-P baixo, deve rejeitar a hipótese nula.

Quando o valor-P é *alto* (ou apenas não baixo o *suficiente*), o que concluímos? Naquele caso, nada vimos de improvável ou surpreendente. Os dados são consistentes com o modelo da hipótese nula e não temos motivo para rejeitar a hipótese nula. Os eventos com uma probabilidade alta de ocorrência acontecem todo o tempo. Portanto, quando o valor-P for alto, significa que provamos que a hipótese nula é verdadeira? Não! Percebemos que muitas outras hipóteses similares também podem ser responsáveis pelos dados que vimos. O máximo que podemos afirmar é que ela não parece falsa. Formalmente, podemos dizer que "fomos incapazes em rejeitar" a hipótese nula. Parece ser uma conclusão fraca, mas é tudo o que podemos dizer quando o valor-P não for baixo o suficiente. Tudo isso significa que os dados são consistentes com o modelo com o qual começamos.

O que fazer com um réu "inocente"

"Se a justiça não conseguir satisfazer o ônus da prova, vocês devem declarar o réu não culpado."

—Júri do Estado de Nova York Instruções

Vejamos o que essa última afirmação significa em um julgamento. Se as provas não forem fortes o suficiente para rejeitar a suposição de inocência do réu, a qual veredicto o júri recorre? Eles não dizem que o réu é inocente. Eles dizem "não culpado". Ou seja, isso significa que eles não viram evidências suficientes para rejeitar a inocência e condenar o réu. O réu pode ser inocente, mas o júri não tem certeza.

Estatisticamente, a hipótese nula do júri é: réu inocente. Se as evidências forem improváveis demais (o valor-P é baixo), então, dada a suposição de inocente, o júri rejeita a hipótese nula e declara o réu culpado. No entanto – e essa é uma distinção importante – se há *evidência insuficiente* para condenar o réu (se o valor-P *não* é baixo), o júri não conclui que a hipótese nula seja verdadeira e declara o réu inocente. Os júris apenas *não rejeitam* a hipótese nula e declaram o réu "não culpado".

Não queremos rejeitar a hipótese nula?

Geralmente, quem coleta os dados ou executa o experimento espera rejeitar a hipótese nula. Eles esperam que o remédio novo seja melhor que o placebo; que a nova campanha publicitária seja melhor que a antiga; ou que seu candidato esteja na frente do seu concorrente. No entanto, quando praticamos a estatística, não podemos permitir que a esperança afete nossa decisão. A atitude essencial para quem testa uma hipótese é o ceticismo. Até nos convencermos do contrário, acreditamos na afirmação da hipótese nula de que nada há de incomum, inesperado, diferente, etc. Como num julgamento, o ônus da prova permanece com a hipótese alternativa – inocente até prova contrária. Quando você testa uma hipótese, deve agir como um juiz e um júri, mas você não é o promotor público.

Da mesma forma, se os dados não são particularmente improváveis sob a suposição de que a hipótese nula seja verdadeira, o máximo que podemos fazer é "não rejeitar" nossa hipótese nula. Nunca declaramos a hipótese nula verdadeira. Na verdade, simplesmente não sabemos se ela é verdadeira ou não. (Afinal, mais provas podem aparecer mais tarde.)

Imagine um teste para ver se o novo projeto de um *site* de uma empresa estimula um alto percentual de visitantes a fazer compras (comparado ao *site* que eles têm usado por anos). A hipótese nula é de que o novo *site* não estimula mais compras do que o antigo. O teste envia os visitantes aleatoriamente a uma versão ou outra do *site*. É claro, alguns irão fazer compras e outros não. Se compararmos os dois *sites* com somente 10 clientes cada, os resultados provavelmente *não serão claros* e seremos incapazes de rejeitar a hipótese. Isso significa que o novo *design* é um fracasso? Não necessariamente. Apenas significa que não temos evidência o suficiente para rejeitar a hipótese nula. Por isso, não começamos assumindo que o novo *design* é *mais* eficaz. Se fizéssemos isso, poderíamos testar apenas alguns clientes, descobrir que os resultados não estavam claros e declarar que, visto que fomos incapazes de rejeitar a nossa suposição original, o *design* deve ser eficaz. É improvável que o conselho diretor fique impressionado com esse argumento.

Conclusão

Se o valor-P é "baixo", rejeite H_0 e aceite H_A.

Se o valor-P não for "baixo o suficiente", não é possível rejeitar H_0 e o teste é inconclusivo.

TESTE RÁPIDO

1 Uma empresa farmacêutica quer saber se a aspirina ajuda a afinar o sangue. A hipótese nula diz que não afina. Os pesquisadores da empresa testam 12 pacientes, observam a proporção com o sangue mais fino e obtêm um valor-P de 0,32. Eles afirmam que a aspirina não funciona. O que você diria?

2 Um medicamento para a alergia foi testado e constatou-se que ele ofereceu alívio a 75% dos pacientes num experimento clínico de larga escala. Agora, os cientistas querem saber se uma versão nova e "aprimorada" funciona ainda melhor. Qual seria a hipótese nula?

3 O novo medicamento para alergia é testado e o valor-P é 0,0001. O que você concluiria sobre o novo medicamento?

11.4 A lógica do teste de hipóteses

O teste de hipóteses segue uma linha de pensamento cuidadosamente estruturada. Para não nos perdermos no caminho, dividimos essa linha de pensamento em quatro seções distintas: hipótese, modelo, mecânica e conclusão.

Hipóteses

"A hipótese nula nunca é provada ou estabelecida, mas é possivelmente refutada, no curso de uma experimentação. Cada experimento existe somente para oferecer aos fatos uma chance de refutar a hipótese nula."

—Sir Ronald Fisher, *The Design of Experiments*, 1931.

Em primeiro lugar, declare a hipótese nula. É geralmente a declaração cética de que nada é diferente. A hipótese nula assume que a falta (geralmente o *status quo*) é verdadeira (o réu é inocente, o novo método não é melhor que o antigo, as preferências dos clientes não mudaram desde o ano passado, etc.).

Nos testes de hipóteses estatísticos, as hipóteses são quase sempre sobre os parâmetros do modelo. Para avaliar quão improváveis nossos dados são, precisamos de um modelo nulo. A hipótese nula especifica um parâmetro determinado para usar no nosso modelo. Na notação usual, escrevemos H_0 = *parâmetro* = *valor hipotético*. A hipótese alternativa, H_A, contém os valores do parâmetro que consideramos plausíveis quando rejeitamos a hipótese nula.

Modelo

Para planejar um teste de hipóteses, especifique o *modelo* para a distribuição amostral da estatística que você irá usar para testar a hipótese nula e o parâmetro de interesse. Para proporções, usamos o modelo Normal para a distribuição amostral. É claro, todos os modelos requerem suposições, por isso você precisará declará-las e verificar as condições correspondentes. Para um teste de uma proporção, as suposições e condições são as mesmas do intervalo-z de uma proporção.

Nossa etapa do modelo deveria terminar com uma declaração como: *visto que as condições foram satisfeitas, podemos modelar a distribuição amostral da proporção com a distribuição Normal.* Cuidado. Sua etapa do Modelo poderia terminar com: *visto que as condições não foram satisfeitas, não podemos continuar com o teste* (se esse for o caso, pare e reconsidere).

Quando a condição falha ...

Você pode proceder com cautela, declarando claramente suas preocupações. Ou talvez você precise fazer a análise com ou sem um valor atípico, em grupos diferentes ou após declarar novamente a variável resposta. Ou, ainda, talvez você não seja capaz de continuar.

Cada teste que discutimos neste livro tem um nome que você deve incluir no seu relatório. Veremos vários testes nos próximos capítulos. Alguns serão sobre mais do que uma amostra, alguns envolverão outras estatísticas além de proporções e outros usarão modelos diferentes da distribuição Normal (e assim não utilizarão os escores-z). O teste sobre proporções é chamado de **teste-z de uma proporção.**[1]

Teste-z de uma proporção

As condições para o teste-z de uma proporção são as mesmas para o intervalo-z de uma proporção. Testamos a hipótese $H_0 : p = p_0$ usando a estatística

$$z = \frac{(\hat{p} - p_0)}{DP(\hat{p})}.$$

Usamos a proporção hipotética para encontrar o desvio padrão: $DP(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$. Quando as condições são satisfeitas e a hipótese nula é verdadeira, essa estatística segue o modelo Normal Padrão, então podemos usar esse modelo para obter o valor-P.

[1] Também chamado de "teste de uma amostra para uma proporção".

Probabilidade condicional

Você notou que um valor-P resulta do que nos referimos como uma "distribuição condicional" no Capítulo 5? Um valor-P é uma "probabilidade condicional" porque é baseado em – ou na condição de – outro evento ser verdadeiro: é a probabilidade de que os resultados observados tenham acontecido se a hipótese nula for verdadeira.

Mecânica

Sob "Mecânica", executamos os cálculos do nosso teste estatístico a partir dos dados. Os testes diferentes que encontraremos terão fórmulas diferentes e testes estatísticos distintos. Em geral, essa parte é realizada por um programa estatístico ou por uma calculadora. O objetivo principal dos cálculos é obter um valor-P – a probabilidade de que o valor estatístico observado (ou um valor mais extremo) possa ocorrer se a hipótese nula estiver correta. Se o valor-P for pequeno o suficiente, rejeitaremos a hipótese nula.

Conclusões e decisões

A conclusão básica num teste de hipótese formal é somente uma declaração sobre a hipótese nula. Ela simplesmente declara se rejeitamos ou falhamos em rejeitar aquela hipótese. Como sempre, a conclusão dever ser declarada em um contexto, mas a sua conclusão sobre a hipótese nula nunca deve ser o final do processo. Você não pode basear uma decisão somente em um valor-P. As decisões de negócios têm consequências, com tomada de ações ou mudanças de políticas. As conclusões de um teste de hipótese podem ajudar a *embasar* a sua decisão, mas elas não devem ser a única fundamentação para a decisão.

As decisões de negócios devem sempre levar em conta três aspectos: a significância estatística do teste, o *custo* da ação proposta e o *tamanho do efeito* da estatística observada. Por exemplo, um fornecedor de telefones celulares percebe que 30% dos seus clientes trocaram de fornecedores (ou fazem rotatividade) quando seu contrato de dois anos expira. Eles tentam um pequeno experimento e oferecem a uma amostra aleatória de clientes um telefone topo de linha de $350 caso eles renovem seus contratos por mais dois anos. Sem surpresa, eles perceberam que a nova taxa de troca é mais baixa por um valor estatisticamente significativo. Eles deveriam oferecer esses telefones grátis a todos os clientes? Obviamente, a resposta depende de outros fatores além do valor-P do teste. Mesmo se o valor-P for estatisticamente significativo, a decisão de negócios correta também depende do custo dos telefones grátis e por quanto a taxa de rotatividade é rebaixada (o tamanho do efeito). É raro que somente o teste de hipótese seja o suficiente para tomar uma decisão segura de negócios.

"Simples tudo parece,
Mas uma pergunta permanece:
Se uma hipótese você perder,
No seu lugar, outra você vai
escolher ..."

—James Russel Lowell, *Credidimus Jovem Regnare*

11.5 Hipótese alternativa

No nosso exemplo sobre o DJIA, também estávamos interessados na proporção que se desvia de 50% em *cada* direção. Assim, escrevemos nossa hipótese alternativa como H_A: $p \neq 0{,}5$. Tal hipótese alternativa é conhecida como uma **alternativa bilateral,** pois estamos igualmente interessados em desvios nos dois lados do valor da hipótese nula. Para alternativas bilaterais o valor-P é a probabilidade de desvios em ambas as direções do valor da hipótese nula.

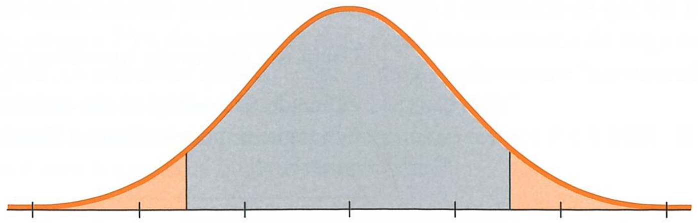


Figura 11.3 O valor-P para uma alternativa bilateral acrescenta as possibilidades nas duas caudas do modelo da distribuição amostral fora do valor que corresponde ao teste estatístico.

Hipótese alternativa

Bilateral

H_0: $p = p_0$

H_A: $p \neq p_0$

Unilateral

H_0: $p = p_0$

H_A: $p < p_0$ ou $p > p_0$

Suponha que queremos testar se a proporção dos clientes que retornam a mercadoria diminuiu sob nosso novo programa de monitoramento da qualidade. Sabemos que a qualidade melhorou, então podemos estar certos de que as coisas não estão piores. Mas os clientes notaram? Estaríamos interessados somente numa amostra da proporção *menor* que o valor da hipótese nula. Escrevemos nossa hipótese alternativa como H_A: $p < p_0$. Uma hipótese alternativa que tem como foco os desvios do valor da hipótese nula em somente uma direção é chamada de **alternativa unilateral.**

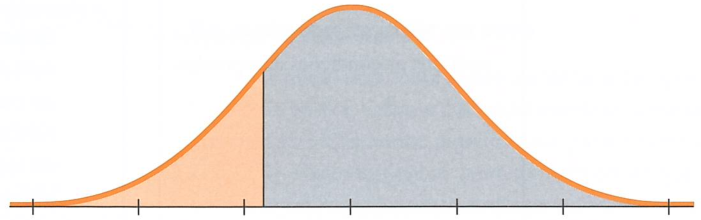


Figura 11.4 O valor-P para uma alternativa unilateral considera somente a probabilidade dos valores além do valor do teste estatístico na direção especificada.

Para um teste de hipótese com uma alternativa unilateral, o valor-P é a probabilidade de desviar da hipótese nula *somente na direção da alternativa*.

EXEMPLO ORIENTADO *Vantagem de jogar em casa*

Os esportes profissionais são grandes negócios. Há maior probabilidade que os torcedores apoiem o time se a equipe da casa tem uma boa chance de vencer. Qualquer um que acompanha ou joga algum esporte já ouviu falar sobre a "vantagem de jogar em casa". Diz-se que os times têm maior probabilidade de vencer quando jogam em casa. Esse fato *incentivaria* os torcedores a irem aos jogos. Será verdade?

Na temporada de 2006 da Major League Baseball (MLB – Liga Nacional do Beisebol), houve 2429 jogos regulares na temporada (um jogo com chuva não foi finalizado). Verificou-se que o time local venceu 1327 dos 2429 jogos, ou em 54,63% das vezes. Se não existisse a vantagem de jogar em casa, os times locais venceriam a metade de todos os jogos. Esse desvio de 50% poderia ser explicado apenas pela variação amostral natural ou a evidência sugere que realmente existe a vantagem de jogar em casa, pelo menos no beisebol profissional?

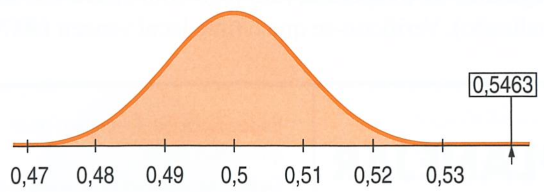
Para testar a hipótese, iremos verificar se a taxa observada de vitórias do time local, 54,63%, é muito maior que 50% e se não podemos explicá-la apenas pela variação casual.

Lembra das quatro etapas principais para executar o teste de hipótese – hipótese, modelo, mecânica e conclusão? Vamos empregá-las a fim de analisar o que descobriremos sobre as chances do time da casa vencer um jogo de beisebol.

PLANEJAR	**Especificação** Declare o que queremos saber. Defina as variáveis e discuta o seu contexto.	*Queremos saber se o time local de beisebol profissional tem mais chances de vencer. Os dados são de todos os 2429 jogos da temporada de 2006 da Major League Baseball. A variável é se o time local venceu ou não. O parâmetro de interesse é a proporção de vitórias do time da casa. Se ele tiver vantagem, esperamos que a proporção seja maior que 0,50. O valor observado de* $\hat{p} = 0{,}5463$.

continua...

continuação

	Hipótese A hipótese nula afirma que não existe vantagem em jogar em casa. Estamos interessados apenas na vantagem de jogar em casa, por isso a hipótese alternativa é unilateral.	H_0: $p = 0{,}50$ H_A: $p > 0{,}50$
	Modelo Pense sobre as suposições e verifique as condições apropriadas.	✓ **Suposição de independência:** Geralmente, o resultado de um jogo não tem efeito no resultado de outro jogo. No entanto, nem sempre isso é verdadeiro. Por exemplo, se um jogador importante se lesiona, a probabilidade de que o time vença nos próximos dois jogos pode diminuir levemente, mas ainda é, grosso modo, verdadeiro.
	Considere o intervalo de tempo cuidadosamente.	✓ **Condição da aleatoriedade:** Temos os resultados de todos os 2429 jogos da temporada de 2006. Mas não estamos interessados apenas em 2006. Embora esses jogos não tenham sido selecionados aleatoriamente, eles podem ser representativos de todos os jogos recentes do beisebol profissional. ✓ **Condição dos 10%:** Essa não é uma amostra aleatória, mas esses 2429 jogos representam menos de 10% de todas as partidas jogadas ao longo dos anos. ✓ **Condição de sucesso/fracasso:** tanto o $np_0 = 2429 \times 0{,}50 = 1214{,}5$ quanto o $nq_0 = 2429 \times 0{,}50 = 1214{,}5$ são pelo menos 10.
	Especifique o modelo da distribuição amostral. Diga qual teste você planeja usar.	Visto que as condições são satisfeitas, usaremos um modelo Normal para a distribuição amostral da proporção e faremos um teste-z para uma proporção.
FAZER	**Mecânica:** O modelo nulo fornece a média e, visto que estamos trabalhando com proporções, a média fornece o desvio padrão.	O modelo nulo é uma distribuição Normal com uma média de 0,50 e um desvio padrão de $DP(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0{,}5)(1-0{,}5)}{2429}} = 0{,}01015$ A proporção observada $\hat{p}$ é 0,5463. 
	Com o uso da tecnologia, podemos encontrar o valor-P que fornece a probabilidade de observar um valor tão extremo (ou maior) do que o observado.	

	A probabilidade de observar um $\hat{p}$ de 0,5463 ou mais no nosso modelo Normal pode ser encontrada usando-se um computador, uma calculadora ou uma tabela.	O valor-P correspondente é < 0,001.
RELATAR	**Conclusão** Escreva sua conclusão sobre o parâmetro, considerando o contexto.	**Memorando:** **Re: vantagem de jogar em casa** Nossa análise dos resultados durante a temporada 2006 da Major League Baseball mostrou uma vantagem estatisticamente significativa para o time local (P < 0,001). Podemos estar bem confiantes de que jogar em casa representa uma vantagem ao time de beisebol.

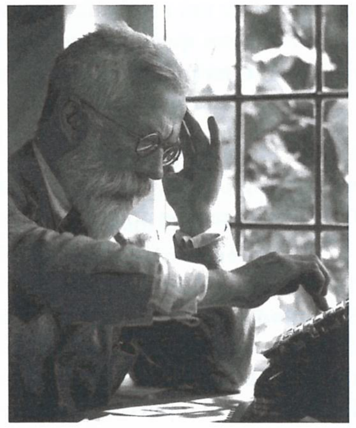
11.6 Níveis alfa e significância

Às vezes, precisamos tomar uma decisão firme se rejeitamos ou não a hipótese nula. Um júri deve *decidir* se as provas alcançam o nível de "além da dúvida razoável". Um administrador deve *selecionar* um *design* para o *site*. Você precisa decidir em qual seção do curso de estatística deve se matricular.

Quando o valor-P é pequeno, ele indica que nossos dados são raros *dado a hipótese nula*. Como humanos, desconfiamos dos eventos raros. Se os dados forem "raros o suficiente", não pensamos que poderiam ter acontecido devido ao acaso. Visto que os dados *realmente* aconteceram, algo deve estar errado. Tudo o que podemos fazer é rejeitar a hipótese nula.

Mas quão raro é "raro"? Quão baixo deve ser o valor-P?

Podemos definir um "evento raro" arbitrariamente determinando um limite para o nosso valor-P. Se o nosso valor-P cair abaixo desse limite, rejeitaremos a hipótese nula. Chamamos tais resultados de *estatisticamente significativos*. A probabilidade limite é denominada **nível alfa**, representado pela letra grega α. Os níveis α comuns são 0,10, 0,05 e 0,01. Você tem a opção – quase uma *obrigação* – de escolher cuidadosamente o nível alfa adequado para a situação. Se você estiver avaliando a segurança dos *air bags*, precisa de um nível alfa baixo; mesmo 0,01 pode não ser baixo o suficiente. Se você quer saber se as pessoas preferem suas *pizzas* com ou sem linguiça calabresa, pode ficar satisfeito com $\alpha = 0{,}10$. Pode ser difícil justificar a sua escolha de α, assim, em geral, arbitrariamente escolhemos 0,05.



Sir Ronald Fisher (1890-1962) foi um dos fundadores da estatística moderna.

◆ **De onde vem o valor 0,05**? Em 1931, numa famosa obra intitulada *The Design of Experiments*, Sir Ronald Fisher discutiu a quantidade de indícios necessários para rejeitar a hipótese nula. Ele afirmou que *dependia da situação*, mas observou que para muitas aplicações científicas, 1 em 20 *seria* um valor razoável, especialmente num *primeiro* experimento – um a ser seguido de confirmação. Desde então, algumas pessoas – na verdade algumas disciplinas – têm agido como se o número 0,05 fosse sagrado.

ALERTA DE NOTAÇÃO:

A primeira letra grega, α, é usada em estatística para a probabilidade limite de um teste de hipóteses. Ela é chamada de nível alfa. Os valores comuns são 0,10, 0,05, 0,01 e 0,001.

O nível alfa também é chamado de **nível de significância**. Quando rejeitamos a hipótese nula, dizemos que o teste é "significativo naquele nível." Por exemplo, podemos dizer que rejeitamos a hipótese nula "num nível de 5% de significância". Você deve selecionar o nível alfa *antes* de analisar os dados. Caso contrário, pode ser acusado de fraudar as conclusões direcionando o nível alfa para os resultados depois de ver os dados.

O que você pode dizer se o valor-P não for menor que α? Quando você não encontrar evidências suficientes para rejeitar a hipótese nula de acordo com o padrão estabelecido, deve dizer: "Os dados falharam no fornecimento de evidências suficientes para

Pode acontecer com você!

É claro, se a hipótese nula *for* verdadeira, não importa qual o nível de alfa escolhido, você ainda tem uma probabilidade α de rejeitar a hipótese nula por engano. Quando realmente rejeitamos a hipótese nula, ninguém jamais pensa que *esta* é uma daquelas raras vezes. Como o estatístico Stu Hunter observa: "Os estatísticos afirmam 'eventos raros acontecem – mas não comigo!'".

Conclusão

Se o valor-P < α, rejeite a H_0.

Se o valor-P $\geq \alpha$, não é possível rejeitar H_0.

a rejeição da hipótese nula". Não diga: "Aceitamos a hipótese nula". Você certamente não provou ou estabeleceu a hipótese nula; ela foi assumida, originalmente. Você *poderia* dizer que *reteve* a hipótese nula, mas é melhor relatar que não pode rejeitá-la.

Olhe novamente para o exemplo sobre a vantagem de jogar em casa. O valor-P era < 0,001. Ele é muito menor que qualquer nível alfa que permite rejeitar H_0. Concluímos: "rejeitamos a hipótese nula. Existem evidências suficientes para concluir que há vantagem em jogar em casa além do que se esperaria de uma variação casual".

A natureza automática da decisão de rejeitar/não rejeitar quando usamos um nível alfa pode deixá-lo desconfortável. Se o seu valor-P estiver levemente acima do seu nível alfa, você não está autorizado a rejeitar a hipótese nula. No entanto, um valor-P pouco abaixo do nível alfa leva à rejeição. É natural que isso seja incômodo. Muitos estatísticos acham que é melhor relatar o valor-P do que escolher um nível alfa e levar a decisão para um veredicto final de rejeitar/não rejeitar. Portanto, quando você toma a sua decisão, é recomendável enunciar o seu valor-P como uma indicação da força das evidências.

- **Está nas estrelas.** Algumas disciplinas vão ainda mais longe e codificam os valores-P pelo seu tamanho. Nesse método, um valor-P entre 0,05 e 0,01 é destacado por um simples asterisco (*). Um valor-P entre 0,01 e 0,001 recebe dois asteriscos (**) e um valor-P menor que 0,001 recebe três (***). Isso pode ser um resumo conveniente do peso das evidências contra a hipótese nula, mas não é prudente levar as distinções muito a sério e tomar decisões rígidas próximas aos limites. Os limites são uma questão de tradição, não ciência; nada existe de especial em 0,05. Um valor-P de 0,051 deve ser observado com seriedade e não casualmente jogado fora porque é maior que 0,05, e um valor-P igual a 0,009 não é muito diferente de um 0,011.

Às vezes, é melhor relatar que a conclusão ainda não é clara e sugerir que sejam coletados mais dados (num julgamento, um júri pode "estar em dúvida" e ser incapaz de chegar a um veredicto). Em tais casos, é recomendável relatar o valor-P, já que ele é o melhor resumo que temos do que os dados dizem ou falham em dizer sobre a hipótese nula.

O que significa quando afirmamos que um teste é estatisticamente significativo? Queremos dizer que o teste estatístico tem um valor-P mais baixo que nosso valor alfa. Não se iluda pensando que aquela "significância estatística" necessariamente traz consigo qualquer importância ou impacto.

Para amostras grandes, mesmo pequenos desvios (não importantes) da hipótese nula podem ser estatisticamente significativos. Por outro lado, se a amostra não for grande o suficiente, mesmo grandes diferenças financeiras ou cientificamente importantes podem não ser estatisticamente significativas.

Significância prática *versus* significância estatística

Uma grande empresa de seguros extraiu seus dados e encontrou uma diferença estatisticamente significativa ($P = 0{,}04$) entre o valor da média das apólices vendidas em 2001 e das vendidas em 2002. A diferença entre os valores da média era de $0,98. Embora fosse estatisticamente significativa, a gerência não a percebeu como uma diferença importante, já que uma apólice típica é vendida por mais de $1000. Por outro lado, uma melhoria negociável de 10% numa taxa de rendimento para um novo medicamento para dor pode não ser estatisticamente significativa a não ser que uma grande quantidade de pessoas seja testada. O efeito, que é economicamente significativo, talvez não seja estatisticamente significativo.

É uma boa prática relatar a magnitude da diferença entre o valor da estatística observada e o valor da hipótese nula (nas mesmas unidades de dados) junto com o valor-P no qual você baseou sua decisão sobre a significância estatística.

11.7 Valores críticos

Quando construímos um intervalo de confiança, encontramos um **valor crítico**, z^*, que delimita o nosso nível de confiança selecionado. Os valores críticos também podem ser usados como um atalho para os testes de hipóteses. Antes da popularidade dos computadores e calculadoras, os valores-P eram difíceis de se obter. Era mais fácil selecionar alguns níveis alfa comuns (0,05, 0,01, 0,001, por exemplo) e descobrir os valores críticos correspondentes para o modelo Normal (isto é, os valores críticos correspondentes aos intervalos de confiança 0,95, 0,99 e 0,999, respectivamente). Em vez de encontrar a probabilidade que corresponde à nossa estatística observada, você apenas calculava quantos desvios padrão estavam longe do valor hipotético e comparava aquele valor diretamente a esses valores z^* (lembre que, sempre que medimos a distância de um valor até a média em desvios padrões, encontramos um escore-z). Qualquer escore-z maior em magnitude (isto é, mais extremo) do que um valor crítico específico deve ser menos provável, assim, ele terá um valor-p menor que o alfa correspondente.

Se você tiver de tomar uma decisão rápida, sem auxílio tecnológico, lembre do "2". Ele é nosso velho amigo da Regra 68-95-99,7. É aproximadamente o valor crítico para testar uma hipótese *versus* uma alternativa bilateral com $\alpha = 0{,}05$. O valor crítico exato é 1,96, mas 2 está próximo o suficiente para a maioria das decisões.

Se quisermos tomar uma decisão de rejeitar/não rejeitar, comparar um escore-z observado com um valor crítico para um nível alfa especificado seria um atalho para a decisão. Para o exemplo da vantagem de jogar em casa, se escolhermos $\alpha = 0{,}05$, então para rejeitar H_0, nosso escore-z deve ser maior que o valor crítico unilateral de 1,645. A proporção observada era, na verdade, de 4,78 desvios padrões acima de 0,5, assim, obviamente rejeitamos a hipótese nula. Isso é perfeitamente correto e nos dá uma decisão de sim/não, mas fornece menos informação sobre a hipótese porque não temos de pensar sobre o valor-P. Com a tecnologia, os valores-P são fáceis de se encontrar. Como eles fornecem mais informação sobre a força da evidência, você deve relatá-los.

Eis os valores críticos z^* tradicionais do modelo Normal[2]:

α	unilateral	bilateral
0,05	1,645	1,96
0,01	2,33	2,576
0,001	3,09	3,29

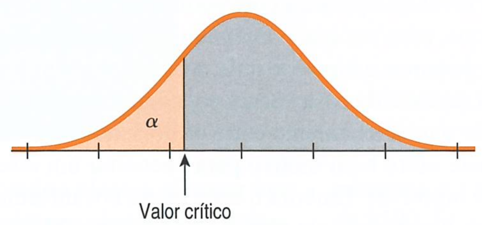


Figura 11.5 Quando a alternativa for unilateral, a probabilidade α fica em um único lado.

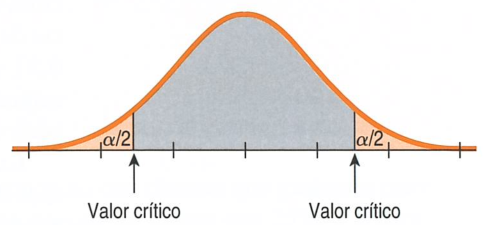


Figura 11.6 Quando a alternativa for bilateral, a probabilidade α é dividida igualmente entre os dois lados.

[2] De certa forma, esses valores são o outro lado da Regra 68-95-99,7. Naquela, escolhemos a distância estatística da média até as áreas das caudas. Aqui, selecionamos áreas convenientes das caudas (0,05, 0,01 e 0,001, em um lado ou acrescentando ambos os lados) e registramos as distâncias estatísticas correspondentes.

ALERTA DE NOTAÇÃO:

Anexamos símbolos a muitos dos *p*. Vamos esclarecê-los.

p é o parâmetro da população – a verdadeira proporção na população.

p_0 *é o valor hipotético de* p.

$\hat{p}$ é uma proporção observada.

p^* é um valor crítico de uma proporção para um α específico (veja a página 337).

11.8 Intervalos de confiança e testes de hipóteses

Os intervalos de confiança e os testes de hipóteses são construídos com os mesmos cálculos. Eles têm as mesmas suposições e condições. Como acabamos de ver, você pode aproximar um teste de hipótese examinando o intervalo de confiança. Apenas pergunte se o valor da hipótese nula é consistente com o intervalo de confiança para o parâmetro no nível de confiança correspondente. Visto que os intervalos de confiança são naturalmente bilaterais, eles correspondem aos testes bilaterais. Por exemplo, um intervalo de confiança de 95% corresponde ao teste de hipótese bilateral em $\alpha = 5\%$. Em geral, um intervalo de confiança com um nível $C\%$ de confiança corresponde ao teste de hipótese bilateral com um nível α de $100 - C\%$.

Os relacionamentos entre intervalos de confiança e os testes de hipótese unilaterais fornecem uma escolha. Para um teste unilateral com $\alpha = 5\%$, você pode construir um nível de confiança unilateral de 95%, deixando 5% em uma cauda.

Um intervalo de confiança unilateral deixa um lado sem limites. Assim, no exemplo sobre jogar em casa, queríamos saber se o fator local fornece ao time da casa uma *vantagem*. Portanto, nosso teste era, naturalmente, unilateral. Um intervalo de confiança unilateral de 95% seria construído de um lado do intervalo de confiança bilateral associado:

$$0{,}5463 - 1{,}645 \times 0{,}0101 = 0{,}530.$$

Para deixar 5% em um lado, usamos o valor z^* 1,645, que deixa 5% em uma cauda. Escrever o intervalo unilateral como $(0{,}530, \infty)$ permite afirmar com 95% de confiança que sabemos que o time local irá vencer, em média, pelo menos 53,0% das vezes. Para testar a hipótese H_0: $p = 0{,}50$, observamos que o valor 0,50 não está nesse intervalo. O limite inferior de 0,53 está claramente acima de 0,50, mostrando a conexão entre a hipótese e os intervalos de confiança.

Como vimos no Capítulo 10, não é exatamente verdadeiro que os testes de hipótese e intervalos de confiança sejam equivalentes para proporções. Para um intervalo de confiança, estimamos o desvio padrão de $\hat{p}$ do próprio $\hat{p}$, tornando-o um *erro padrão*. Para o teste de hipótese correspondente, usamos o *desvio padrão* do modelo para $\hat{p}$ com base no valor da hipótese nula p_0. Quando $\hat{p}$ e p_0 estão próximos, esses cálculos fornecem resultados similares. Quando eles diferem, você provavelmente irá rejeitar H_0 (já que a proporção observada está longe do valor hipotético). Nesse caso, será melhor construir seu intervalo de confiança com um erro padrão estimado a partir dos dados do que depender do modelo que acabou de rejeitar.

Por conveniência e para fornecer mais informação, entretanto, às vezes relatamos um intervalo de confiança bilateral mesmo se estamos interessados num teste unilateral. Para o exemplo do fator local, poderíamos determinar um intervalo de confiança de 90% como sendo:

$$0{,}5463 \pm 1{,}645 \text{x} 0{,}0101 = (0{,}530;\ 0{,}563).$$

Observe que *igualamos* o ponto da esquerda deixando α em *ambos* os lados, o que criou o nível de confiança correspondente de 90%. Ainda podemos ver a correspondência de que, uma vez que o intervalo de confiança (bilateral) de 95% para $\hat{p}$ não contém 0,50, rejeitamos a hipótese nula, mas ele também informa que é improvável que o percentual do time da casa vencer seja maior que 56,3% e auxiliou a compreensão. Você pode ver o relacionamento entre os dois intervalos de confiança na Figura 11.7.

Existe outro bom motivo para encontrar um intervalo de confiança junto com um teste de hipóteses. Embora o teste possa nos informar se a estatística observada difere do valor hipotético, ele não diz por quanto. Muitas vezes, as decisões de negócios dependem não somente da existência ou não de uma diferença estatisticamente significativa, mas também da importância da diferença. Para a vantagem de jogar em casa, o intervalo de confiança correspondente mostra que, durante toda a temporada, a vantagem acrescenta uma média de aproximadamente duas a seis vitórias extras para um time. Isso representa uma diferença significativa tanto na posição do time na liga quanto no tamanho da torcida.

"Afirmações extraordinárias requerem prova extraordinária."

—Carl Sagan

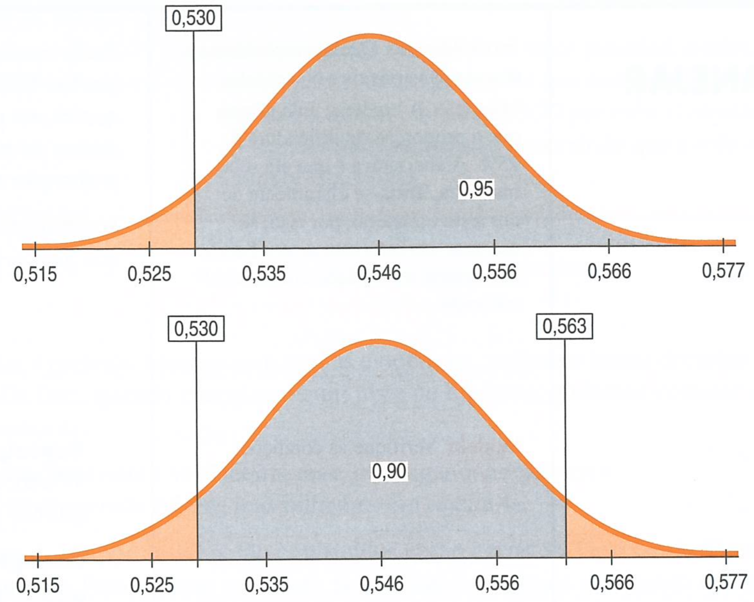


Figura 11.7 O intervalo de confiança de 95% (acima) deixa 5% em um lado (neste caso, à esquerda), mas o outro lado fica sem limites. O intervalo de 90% é simétrico e se equipara com o intervalo unilateral no lado de interesse. Ambos os intervalos indicam que um teste unilateral de $p = 0{,}50$ seria rejeitado com um $\alpha = 0{,}05$ para qualquer valor de $\hat{p}$ maior que 0,530.

TESTE RÁPIDO

4 Um banco testa um novo modelo para que seus clientes devedores paguem suas contas atrasadas do cartão de crédito. A maneira padrão era enviar uma carta (com um custo aproximado de $0,60 cada) solicitando o pagamento, o que funcionava 30% das vezes. O banco quer testar um novo método que envolve o envio de um DVD ao cliente, estimulando-o a contatar o banco e estabelecer um plano de pagamento. Desenvolver e enviar o DVD custa aproximadamente $10,00 por cliente. Qual é o parâmetro de interesse? Qual é a hipótese nula e qual é a alternativa?

5 O banco organiza um experimento para testar a eficácia do DVD. O DVD é enviado a vários clientes devedores aleatoriamente selecionados e os empregados mantêm registros de quantos clientes contatam o banco para quitar as dívidas. O banco acabou de receber os resultados do seu teste da estratégia do DVD. Um intervalo de confiança de 90% para a taxa de sucesso é (0,29; 0,45). Seu antigo método de enviar uma carta funcionava 30% das vezes. Você pode rejeitar a hipótese nula e concluir que o método aumenta a proporção com um $\alpha = 0{,}05$? Explique.

6 Dado o intervalo de confiança que o banco encontrou no experimento do envio do DVD, o que você recomendaria que fosse feito? O banco deveria descartar a estratégia do DVD?

EXEMPLO ORIENTADO — *Promoção do cartão de crédito*

Uma empresa de cartão de crédito planeja oferecer um programa de incentivo especial aos clientes que gastarem pelo menos $500 no próximo mês. O departamento de *marketing* pegou uma amostra de 500 clientes do mesmo mês do ano anterior e observou que o gasto médio foi de $478,10 e o gasto mediano foi de $216,48. O departamento de finanças afirma que a única quantia relevante é a proporção dos clientes que gastaram mais de $500. Se a proporção for menor que 25%, o programa irá perder dinheiro.

Entre os 500 clientes, 148 ou 29,6% deles gastaram $500 ou mais. Podemos usar um intervalo de confiança para testar se o objetivo de 25% para todos os clientes foi satisfeito?

continua...

PLANEJAR

Especificação Declare o problema e discuta as variáveis e o contexto.

Queremos saber se 25% ou mais dos clientes irão gastar \$500 ou mais no próximo mês e se eles se qualificam para o programa especial. Usaremos os dados do mesmo mês do ano passado para estimar a proporção e ver se ela é de pelo menos 25%.

Hipótese A hipótese nula afirma que a proporção qualificadora é 25%. A alternativa é que ela seja mais alta. Trata-se claramente de um teste unilateral, por isso, se usarmos um intervalo de confiança, deveremos ter cuidado com o nível utilizado.

A estatística é $\hat{p} = 0{,}296$, a proporção dos clientes que gastaram \$500 ou mais.

$$H_0: p = 0{,}25$$

$$H_A: p > 0{,}25$$

Modelo Verifique as condições.

✓ **Suposição da independência:** Os clientes não influenciam uns aos outros quando se refere aos gastos com seus cartões de crédito.

✓ **Condição de aleatoriedade:** Esta é uma amostra aleatória do banco de dados da empresa.

✓ **Condição dos 10%:** A amostra é menor que 10% de todos os clientes da empresa.

✓ **Condição de sucesso/fracasso:** Houve 148 sucessos e 352 fracassos, ambos estão acima de 10. A amostra é grande o suficiente.

Esclareça o método. Aqui, estamos usando um intervalo de confiança para testar uma hipótese.

Sob essas condições, o modelo amostral é Normal. Determinaremos um intervalo-z para uma proporção.

FAZER

Mecânica Escreva a informação dada e determine a proporção amostral.

Para usar um intervalo de confiança, precisamos de um nível de confiança que corresponda ao nível alfa do teste. Se usarmos $\alpha = 0{,}05$, devemos construir um intervalo de confiança de 90%, porque se trata de um teste unilateral. Isso deixará 5% em cada lado da proporção observada.

Determine o erro padrão da proporção amostral e a margem de erro. O valor crítico é $z^* = 1{,}645$.

O intervalo de confiança é: estimativa ± a margem de erro.

$n = 500$, assim

$$\hat{p} = \frac{148}{500} = 0{,}296 \text{ e}$$

$$EP(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0{,}296)(0{,}704)}{500}} = 0{,}020$$

$$ME = z^* \times SE(\hat{p})$$

$$= 1{,}645(0{,}020) = 0{,}033$$

O intervalo de confiança de 90% é: $0{,}296 \pm 0{,}033$ ou $(0{,}263; 0{,}329)$.

RELATAR

Conclusão Vincule o intervalo de confiança à sua decisão sobre a hipótese nula e, então, escreva sua conclusão no contexto.

Memorando:

Re: Promoção do cartão de crédito

Nosso estudo de uma amostra de registros de clientes indica que entre 26,3 e 32,9% deles gastam \$500 ou mais. Estamos 90% confiantes de que esse intervalo inclui o valor verdadeiro. Visto que o valor mínimo adequado de 25% está abaixo desse intervalo, concluímos que ele não é um valor plausível, assim, rejeitamos a hipótese nula de que somente 25% dos clientes gastam mais que \$500 por mês. O objetivo parece ter sido alcançado, assumindo que o mês estudado seja típico.

11.9 Dois tipos de erros

Ninguém é perfeito. Mesmo com muitas evidências, podemos tomar decisões equivocadas. De fato, quando executamos um teste de hipótese, podemos cometer erros de *duas maneiras*.

I. A hipótese nula é verdadeira, mas, por engano, a rejeitamos.
II. A hipótese nula é falsa, mas falhamos em rejeitá-la.

Esses dois tipos de erros são conhecidos como **erros** do **Tipo I** e do **Tipo II**, respectivamente. Para lembrar os nomes, pense que começamos assumindo que a hipótese nula é verdadeira, então um erro do Tipo I é o primeiro tipo de erro que poderemos cometer.

Nos testes com doenças, a hipótese nula é normalmente a suposição de que a pessoa é saudável. A alternativa é que ela tem a doença que queremos testar. Assim, um erro do Tipo I é um *falso positivo* – uma pessoa saudável é diagnosticada como tendo a doença. Um erro do Tipo II é uma pessoa doente ser declarada saudável, um *falso negativo*. Esses erros têm outros nomes, dependendo da disciplina e do contexto.

Qual tipo de erro é o mais sério depende da situação. Num julgamento, um erro do Tipo I ocorre se o júri condena uma pessoa inocente. Um erro do Tipo II ocorre se o júri falha em condenar uma pessoa culpada. Qual parece mais sério? Num diagnóstico médico, um falso negativo poderia significar que um paciente doente não é tratado. Um falso positivo poderia significar que a pessoa deve fazer mais testes.

No planejamento de negócios, um resultado falso positivo pode significar que o dinheiro será investido num projeto que não terá retorno. Um resultado falso negativo pode significar não investir em um projeto que será lucrativo. Qual erro é o pior, o investimento perdido ou a oportunidade perdida? A resposta sempre depende da situação, do custo e do seu ponto de vista.

Eis uma ilustração das situações:

		A Verdade: H_0 Verdadeira	A Verdade: H_0 Falsa
Minha decisão	Rejeitar H_0	Erro do Tipo I	OK
Minha decisão	Não foi possível rejeitar H_0	OK	Erro do Tipo II

Figura 11.8 Os dois tipos de erros ocorrem na diagonal, onde a verdade e a decisão não se equiparam. Lembre que começamos assumindo que H_0 *é verdadeira, por isso um erro cometido (rejeitá-la)* quando H_0 é verdadeira é chamado de erro do Tipo I. Um erro do Tipo II é cometido quando H_0 é falsa (e não a rejeitamos).

Qual é a frequência de um erro do Tipo I? Ele acontece quando a hipótese nula é verdadeira e tivemos a má sorte de coletar uma amostra incomum. Para rejeitar H_0, o

ALERTA DE NOTAÇÃO:

Em estatística, α é a probabilidade do erro do Tipo I e β é a probabilidade do erro do Tipo II.

A hipótese nula especifica um único valor para o parâmetro. Assim, fica fácil calcular a probabilidade de cometermos o erro do Tipo I. No entanto, a hipótese alternativa fornece inúmeras opções de valores e podemos querer determinar o valor de β para muitos deles.

Vimos formas de encontrar o tamanho da amostra pela especificação da margem de erro. Escolher um tamanho de amostra de modo a cometer um determinado erro β (para um valor específico da hipótese alternativa) às vezes é mais adequado, mas os cálculos são mais complexos e estão além do escopo deste livro.

valor-P deve estar abaixo de α. Quando a H_0 for verdadeira, isso acontece *exatamente* com a probabilidade α. Portanto, quando você escolhe o nível α, está determinando que a probabilidade do erro do Tipo I seja igual a α.

E se a H_0 não for verdadeira? Então, possivelmente não poderemos cometer um erro do Tipo I. Você não consegue um falso positivo de uma pessoa doente. Um erro do Tipo I pode acontecer somente quando a H_0 for verdadeira.

Quando a H_0 for falsa e a rejeitamos, fizemos o certo. A capacidade do teste de detectar uma hipótese falsa é chamada de **poder** do teste. Num julgamento, o poder é a medida da capacidade do sistema da justiça criminal de condenar pessoas culpadas. Falaremos mais sobre o poder em breve.

Quando a H_0 for falsa, mas não foi possível rejeitá-la, cometemos um erro do Tipo II. Designamos a letra β à probabilidade deste erro. Qual é o valor de β? Ele é mais difícil de avaliar do que α,porque não sabemos qual é realmente o valor do parâmetro, nessa situação. Quando a H_0 for verdadeira, ela especifica um único valor do parâmetro. Entretanto, quando a H_0 é falsa, não temos um valor específico; há muitos valores possíveis. Podemos calcular a probabilidade β para qualquer valor do parâmetro em H_A, mas a opção de qual escolher nem sempre é clara.

Uma maneira de focar nossa atenção é pensar sobre o *tamanho do efeito*. Ou seja, perguntar: que tamanho faria a diferença? Suponha que uma instituição beneficente quer testar se colocar etiquetas personalizadas com o endereço num envelope junto com um pedido para uma doação aumenta a taxa de respostas acima da linha base de 5%. Se a resposta mínima que cobriria os custos das etiquetas de endereços é 6%, a instituição calcularia β para uma alternativa $p = 0{,}06$.

É claro, podemos reduzir β para *todos* os valores de parâmetro alternativos aumentando o valor de α. Tornando mais fácil a rejeição da hipótese nula, seria mais provável rejeitá-la se fosse verdadeira ou falsa. A única maneira de reduzir *ambos* os tipos de erro é coletar mais evidências ou, em termos estatísticos, coletar mais dados. De outra forma, apenas acabamos trocando um erro por outro. Sempre que você projetar uma pesquisa ou experimento, calcule β (para um nível α razoável). Use um valor alternativo do parâmetro que corresponda a um tamanho do efeito que você quer ser capaz de detectar. Os estudos falham com frequência porque os tamanhos das amostras são pequenos demais para detectar a mudança que eles estão procurando.

TESTE RÁPIDO

7 Você se lembra do banco que está enviando DVDs com o objetivo de que seus clientes quitem pagamentos atrasados? Ele busca evidências de que a estratégia dispendiosa do DVD produza uma taxa de sucesso maior que as cartas que estavam sendo enviadas. Explique o que um erro do Tipo I representa nesse contexto e quais seriam as consequências para o banco.

8 O que é um erro do Tipo II no contexto do experimento do banco e quais seriam as consequências?

9 Se a estratégia do DVD *realmente* funcionar – conseguindo que 60% das pessoas paguem seus débitos – o poder do teste seria mais alto ou mais baixo, comparado com os 32% da taxa do pagamento da dívida? Explique brevemente.

*11.10 Poder

Lembre-se de que nunca podemos provar que a hipótese nula é verdadeira. Podemos dizer apenas que não foi possível rejeitá-la. Mas quando não conseguimos rejeitar a hipótese nula, é natural querer saber se, de fato, fizemos o possível. É possível que a hipótese nula seja falsa e que nosso teste seja fraco demais para detectá-la?

Quando a hipótese nula realmente *é* falsa, esperamos que nosso teste seja poderoso o suficiente para rejeitá-la. Queremos saber qual é a nossa probabilidade de sucesso. O poder do teste é uma forma de pensar nisso. O poder de um teste é a probabilidade que

ele rejeite corretamente uma hipótese nula falsa. Quando o poder for alto, podemos estar confiantes de que fizemos o suficiente. Sabemos que β é a probabilidade de que um teste *não* rejeite uma hipótese nula; portanto, o poder do teste é o complemento $1 - \beta$. Bastaria escrever $1 - \beta$, mas o poder é um conceito tão importante que tem nome próprio.

Sempre que um estudo é incapaz de rejeitar a hipótese nula, o poder do teste é questionado. O tamanho da amostra era grande o suficiente para detectar um efeito, caso existisse um? Poderíamos não ter notado um efeito grande o suficiente para ser interessante, apenas porque falhamos na coleta de dados suficientes ou porque havia muita variabilidade nos dados que coletamos? Será que o experimento simplesmente não tem o poder adequado para detectar o efeito?

Quando calculamos o poder, imaginamos que a hipótese nula seja falsa. O valor do poder depende de quão longe está o valor verdadeiro do valor da hipótese nula. Chamamos a distância entre o valor da hipótese nula, p_0, e o valor verdadeiro, p, de **tamanho do efeito.** O poder depende diretamente do tamanho do efeito. É mais fácil ver efeitos maiores, então quanto mais distante p_0 estiver de p, maior será o poder.

Como podemos decidir que poder precisamos? A escolha do poder é uma decisão mais financeira ou científica do que estatística, já que, para calcular o poder, precisamos especificar o valor "verdadeiro" do parâmetro no qual estamos interessados. Em outras palavras, o poder é calculado para um tamanho do efeito específico e ele muda dependendo do tamanho do efeito que queremos detectar.

Faça um gráfico!

Quanto maior for o tamanho do efeito, mais fácil deverá ser vê-lo. Obter um tamanho da amostra maior diminui a probabilidade de um erro do Tipo I, portanto, aumenta o poder. Quanto mais dispostos estivermos a aceitar um erro do Tipo I, menos provável será cometermos um erro do Tipo II.

A Figura 11.9 pode ajudá-lo a visualizar as relações entre esses conceitos. Suponha que estejamos testando H_0: $p = p_0$ *versus* a alternativa H_A: $p > p_0$. Rejeitaremos

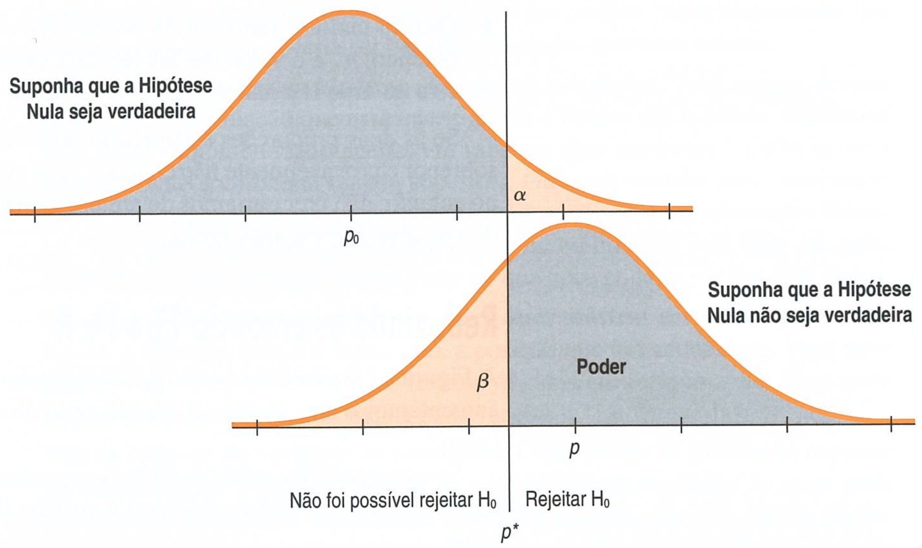


Figura 11.9 O poder de um teste é a probabilidade de ele rejeitar a falsa hipótese nula. A figura superior mostra o modelo considerando a hipótese nula verdadeira. Rejeitaríamos a hipótese nula num teste unilateral se observássemos um valor na região escura à direita do valor crítico, p^*. A figura de baixo mostra o modelo verdadeiro. Se o valor verdadeiro de p for maior que p_0, é mais provável que observemos um valor que exceda o valor crítico e que tomaremos a decisão correta de rejeitar a hipótese nula. O poder do teste é a região mais clara à direita da figura de baixo. É claro que, mesmo coletando amostras cujas proporções observadas estejam distribuídas em volta de p, às vezes conseguiremos um valor na região escura à esquerda e cometeremos um erro do Tipo II ao não rejeitar a hipótese nula.

a hipótese nula se a proporção observada, $\hat{p}$, for grande o suficiente. Por *grande o suficiente*, queremos dizer $\hat{p} > p^*$ para um valor crítico p^* (mostrado como a região escura na cauda direita da curva superior). O modelo superior mostra uma figura do modelo de distribuição amostral para a proporção quando a hipótese nula for verdadeira. Se a hipótese nula fosse verdadeira, isso seria uma figura daquele valor verdadeiro. Cometeríamos um erro do Tipo I sempre que a amostra fornecesse $\hat{p} > p^*$, porque rejeitaríamos a hipótese nula (verdadeira). Amostras incomuns como aquela somente aconteceriam com a probabilidade α.

Entretanto, a hipótese nula raras vezes é *exatamente* verdadeira. O modelo de probabilidade inferior supõe que H_0 não seja verdadeira. De fato, ele supõe que o valor verdadeiro seja p e não p_0. Ele mostra uma distribuição de possíveis valores $\hat{p}$ observados em volta desse valor verdadeiro. Por causa da variabilidade amostral, às vezes $\hat{p} < p^*$, e não é possível rejeitar a (falsa) hipótese nula. Então cometemos um erro do Tipo II. A área abaixo da curva à esquerda de p^*, no modelo da figura inferior, representa a frequência com que isso acontece. A probabilidade é β. Nesta figura, β é menos que a metade da outra região, assim, na maioria das vezes, tomamos a decisão certa. O *poder* do teste – a probabilidade de que tomemos a decisão certa – está mostrado como a região à direita de p^*. Ele é igual a $1 - \beta$.

Calculamos p^* com base no modelo da primeira figura porque p^* depende somente do modelo nulo e do nível alfa. Não importa o que acontecer, a proporção verdadeira, p^*, não muda. Afinal, não *sabemos* o valor verdadeiro, logo não podemos usá-lo para determinar o valor crítico. Porém, sempre rejeitamos H_0 quando $\hat{p} > p^*$.

Com que frequência rejeitamos H_0 quando ela *é falsa* depende do tamanho do efeito. Podemos ver a partir da figura que, se a proporção verdadeira estivesse mais longe do valor hipotético, a curva inferior se moveria para a direita, tornando o poder maior.

Podemos ver várias relações importantes a partir dessa figura:

- Poder = $1 - \beta$.
- Mover o valor crítico, p^*, para a direita, reduz α, a probabilidade de um erro do Tipo I, mas aumenta β, a probabilidade de um erro do Tipo II. Assim, diminui o poder.
- Quanto maior o tamanho do efeito verdadeiro, a diferença real entre o valor hipotético, p_0, e o valor verdadeiro da população, p, menor a chance de cometer um erro do Tipo II e maior o poder do teste.

Se as duas proporções estão muito distantes, os dois modelos dificilmente irão se sobrepor e provavelmente não cometeremos qualquer erro do Tipo II – provavelmente, no entanto, não precisaríamos de um procedimento formal para testar a hipótese a fim de ver uma diferença tão óbvia.

Reduzindo os erros do Tipo I e II

A Figura 11.9 parece mostrar que, se reduzirmos o erro do Tipo I, automaticamente aumentamos o erro do Tipo II. Existe uma forma de reduzir ambos. Você saberia dizer como?

Se conseguirmos tornar as duas curvas mais estreitas, como mostra a Figura 11.10, a probabilidade dos erros do Tipo I e do Tipo II diminuiria e o poder do teste aumentaria.

Como podemos fazer isso? A única forma é reduzir os desvios padrão, aumentando o tamanho da amostra (lembre: essas são figuras de modelos de distribuição amostral, não de dados). Aumentar o tamanho da amostra funciona independentemente dos parâmetros verdadeiros da população. Mas lembre-se da maldição dos lucros decrescentes. O desvio padrão da distribuição amostral diminui somente com a *raiz quadrada* do tamanho da amostra, assim, para dividir por dois o desvio padrão, devemos *quadruplicar* o tamanho da amostra.

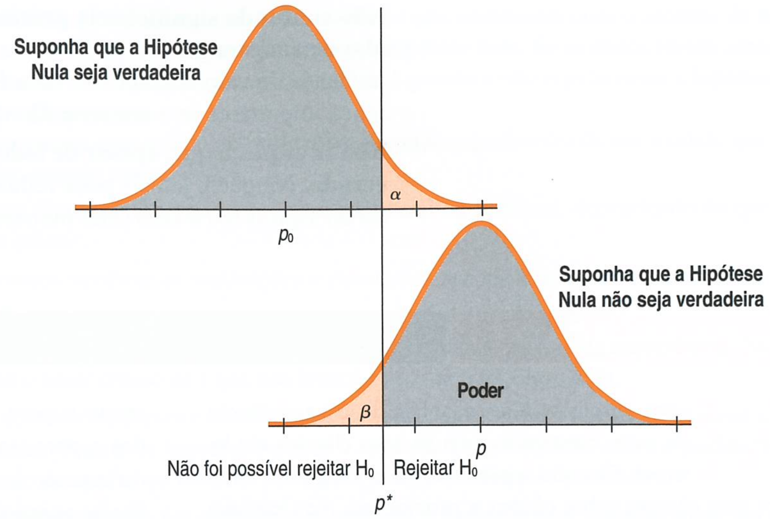


Figura 11.10 Tornar os desvios padrão menores aumenta o poder sem mudar o nível alfa ou o valor crítico z correspondente. As médias estão tão distantes quanto estavam na Figura 11.9, mas as taxas de erro estão reduzidas.

O QUE PODE DAR ERRADO?

- **Não baseie sua hipótese nula no que você vê nos dados.** Você não pode primeiro analisar os dados e depois tentar ajustar a sua hipótese nula para que ela seja rejeitada. Se o seu valor da amostra resulta em $\hat{p} = 51{,}8\%$, com um desvio padrão de 1%, não forme a hipótese nula como H_0: $p = 49{,}8\%$, sabendo que isso o fará rejeitá-la. A sua hipótese nula descreve um cenário "nada interessante" ou "nada mudou" e não deveria ser baseada nos dados que você coletou.
- **Também não baseie sua hipótese alternativa nos dados.** Você sempre deveria pensar sobre a situação que está investigando e basear sua hipótese alternativa nela. Você está interessado apenas em saber se algo *aumentou*? Então escreva uma alternativa unilateral (cauda superior). Ou você também está interessado numa mudança em ambas as direções? Então você quer uma alternativa bilateral. Você deve decidir se faz um teste unilateral ou bilateral com base em quais resultados seriam do seu interesse e não no que você poderia ver nos seus dados.
- **Não faça da sua hipótese nula o que você quer mostrar ser verdadeiro.** Lembre-se, a hipótese nula é o *status quo*, a posição que um cético teria. Você deve analisar se o conjunto dos dados levanta dúvidas em relação a ela. Você pode rejeitar a hipótese nula, mas nunca pode "aceitá-la" ou "prová-la".
- **Não se esqueça de verificar as condições.** O raciocínio de inferência depende da aleatorização. Nenhum cuidado no cálculo de um resultado do teste pode salvá-lo de uma amostra tendenciosa. As probabilidades que você calcula dependem da suposição de independência. Sua amostra deve ser grande o suficiente para justificar o uso de um modelo Normal.
- **Não acredite tanto em níveis alfa arbitrários.** Não existe muita diferença entre um valor-P de 0,051 e um valor-P de 0,049, mas, às vezes, ele é estimado como uma diferença entre noite (ter de reter H_0) e dia (ser capaz de gritar para o mundo que seus resultados são "estatisticamente significativos"). Seria melhor relatar o valor-P e o intervalo de confiança e deixar o mundo (talvez seu gerente ou o cliente) decidir com você.

- **Não confunda significância prática com significância estatística.** Um tamanho da amostra grande pode tornar mais fácil enxergar mesmo uma pequena mudança do valor da hipótese nula. Por outro lado, você pode não ver uma diferença importante se o seu teste não tem poder suficiente.
- **Não se esqueça que, apesar de todo o cuidado, você pode tomar uma decisão errada.** Ninguém jamais pode reduzir a probabilidade de um erro do Tipo I (α) ou do Tipo II (β) a zero (mas aumentar o tamanho da amostra ajuda).

ÉTICA EM AÇÃO

Muitos varejistas têm reconhecido a importância de estar conectado com os seus clientes via Internet. Eles não somente usam a Web para informar seus clientes sobre ofertas e promoções, mas também enviam cupons eletrônicos resgatáveis para descontos. Shellie Cooper, proprietária de longa data de uma pequena loja de alimentos orgânicos, é especializada em produtos orgânicos produzidos no local. Ao longo dos anos, sua base de clientes tem se mantido estável, consistindo principalmente em indivíduos interessados na saúde, que não tendem a ser muito sensíveis ao preço, optando por pagar mais por produtos orgânicos locais de melhor qualidade. Entretanto, enfrentando um aumento da competição por parte das cadeias de mercearias que oferecem opções orgânicas, Shellie está pensando em oferecer cupons e precisa decidir entre o jornal e a Internet. Ela leu recentemente que o percentual de consumidores que usa cupons imprimíveis da Internet está aumentando, mas 15% é ainda muito menor que os 40% que recortam e resgatam os cupons de jornal. Entretanto, ela está interessada em saber mais sobre a Internet e marca uma reunião com Jack Kasor, um consultor de *sites*. Shellie descobre que, com um investimento inicial e uma taxa mensal contínua, Jack projetaria um *site* para a loja dela, hospedaria o *site* no seu servidor e anunciaria cupons eletrônicos para os clientes em intervalos regulares. Embora ela estivesse preocupada com a diferença nas taxas de resgate para os cupons eletrônicos *versus* cupons dos jornais, Jack garantiu que os resgates dos cupons via Web continuariam a subir e que ela deveria esperar um resgate entre 15 e 40% dos seus clientes. Shellie concordou em tentar. Depois dos seis primeiros meses, Jack informou-a de que a proporção dos clientes que resgatou os cupons eletrônicos era significativamente maior que 15%. Ele chegou a essa conclusão ao selecionar vários anúncios ao acaso e encontrar o número de resgates (483) de um número total de cupons enviados (3000). Shellie julgou a informação positiva e decidiu continuar usando os cupons eletrônicos.

PROBLEMA ÉTICO *Significância estatística* versus *significância prática. Embora seja verdadeiro que o percentual dos clientes que resgatou os cupons seja significativamente maior que 15%, na verdade, o percentual está um pouco acima de 16%. Essa diferença atinge aproximadamente 33 clientes a mais do que os 15% e pode não ter significância prática para Shellie (relacionado ao Item A, ASA Ethical Guidelines). Mencionar um intervalo de 15 a 40% ilude Shellie a esperar um valor no meio desses percentuais.*

SOLUÇÃO ÉTICA *Jack deveria relatar a diferença entre o valor observado e o valor hipotético a Shellie, especialmente porque existem custos associados à continuidade dos cupons eletrônicos. Talvez ele devesse recomendar que ela considerasse o uso do jornal.*

O que aprendemos?

Aprendemos a usar o que vemos numa amostra aleatória para testar uma hipótese especial sobre o mundo. Esse é nosso segundo passo em inferência estatística, complementando o uso de intervalos de confiança.

Aprendemos que o teste de uma hipótese envolve propor um modelo e ver se os dados que observamos são consistentes com ele ou são tão incomuns que devemos rejeitá-lo. Para tanto, encontramos o valor-P – a probabilidade de que dados como os nossos tenham ocorrido se o modelo for correto. Se os dados não combinam com o modelo da hipótese nula, o valor-P será pequeno e rejeitaremos a hipótese nula. Se os dados forem consistentes com o modelo da hipótese nula, o valor-P será grande e não rejeitaremos a hipótese nula.

Aprendemos que:

- Começamos com a *hipótese nula* especificando o parâmetro de um modelo que iremos testar, usando nossos dados.
- Nossa *hipótese alternativa* pode ser unilateral ou bilateral, dependendo do que queremos saber.
- Devemos verificar as *suposições* e *condições* apropriadas antes de continuar o nosso teste.
- O *nível de significância* do teste estabelece o nível de prova que requeremos. Isso determina o valor crítico de z que nos levará a rejeitar a hipótese nula.
- Os *testes de hipóteses* e os *intervalos de confiança* são duas formas de analisar a mesma questão. O teste de hipótese fornece a resposta a uma decisão sobre o parâmetro; o intervalo de confiança informa os valores plausíveis daquele parâmetro.
- Se a hipótese nula for realmente verdadeira e a rejeitarmos, cometeremos um *erro do Tipo I*; o nível alfa do teste é a probabilidade de que isso aconteça.
- Se a hipótese nula for realmente falsa e não for possível rejeitá-la, cometeremos um *erro do Tipo II*.

*Seções opcionais

- O *poder* do teste é a probabilidade de rejeitarmos a hipótese nula quando ela é falsa. Quanto maior o tamanho do efeito que estamos testando, maior o poder do teste em detectá-lo.
- Os testes com uma maior probabilidade de erro do Tipo I têm mais poder e menos chance de um erro do Tipo II. Podemos aumentar o poder reduzindo as chances dos dois tipos de erro por meio do aumento do tamanho da amostra.

Termos

Alternativa bilateral Uma hipótese alternativa é bilateral (H_A: $p \neq po$) quando estamos interessados nos desvios em ambas as direções do valor hipotético do parâmetro.

Alternativa unilateral Uma hipótese alternativa é unilateral (por exemplo, H_A: $p > po$ ou H_A: $p < po$) quando estamos interessados em desvios em somente uma direção do valor hipotético do parâmetro.

Erro do Tipo I Erro de rejeitar a hipótese nula quando ela de fato é verdadeira (também chamado de um "falso positivo"). A probabilidade de um erro do Tipo I é representada por α.

Erro do Tipo II Erro de não rejeitar a hipótese nula quando ela de fato é falsa (também chamado de "falso negativo"). A probabilidade de um erro do Tipo II é representada por β e depende do tamanho do efeito.

Hipótese alternativa Hipótese que propõe o que devemos concluir se acharmos que a hipótese nula é improvável.

Hipótese nula Afirmação a ser avaliada num teste de hipóteses. Geralmente, a hipótese nula é uma declaração do tipo "nenhuma mudança do valor tradicional", "nenhum efeito", "nenhuma diferença" ou "nenhum relacionamento." Para uma declaração ser uma hipótese nula testável, ela deve especificar o valor de um parâmetro de uma população que forma a base da distribuição amostral da estatística teste.

Nível alfa Valor-P que determina quando rejeitamos a hipótese nula. Usando um nível alfa de α, se observarmos uma estatística cujo valor-P baseado na hipótese nula for menor que α, rejeitaremos a hipótese nula.

Nível de significância Outro termo para o valor alfa, usado mais frequentemente numa expressão como "num nível de significância de 5%".

Poder Probabilidade de que o teste de hipótese rejeitará corretamente a hipótese nula falsa. Para encontrar o poder de um teste, devemos especificar um valor particular alternativo para o parâmetro como sendo o valor "verdadeiro". Para qualquer valor específico da hipótese alternativa, o poder é $1 - \beta$.

Tamanho do efeito Diferença entre o valor da hipótese nula e o valor verdadeiro de um parâmetro do modelo.

Teste-*z* de uma proporção Teste da hipótese nula em que a proporção de uma única amostra se iguala a um valor especificado (H_0: $p = p0$) comparando a estatística ao modelo Normal padrão.

Valor crítico Valor da distribuição amostral da estatística cujo valor-P é igual ao nível alfa. Qualquer valor da estatística mais afastado do valor da hipótese nula do que o valor crítico terá um valor-P menor do que α e levará à rejeição da hipótese nula. O valor crítico é geralmente denotado com um asterisco, como z^*, por exemplo.

Valor-P Probabilidade de observar um valor para uma estatística teste pelo menos tão distante do valor hipotético quanto o valor da estatística realmente observada, caso a hipótese nula seja verdadeira. Um valor-P pequeno indica que a observação obtida é improvável dada a hipótese nula e, dessa forma, fornece evidência contra a hipótese nula.

Habilidades

PLANEJAR

- Ser capaz de declarar a hipótese nula e a alternativa para o teste-*z* de uma proporção.
- Saber como pensar sobre as suposições e suas condições associadas. Examinar os dados para violações dessas condições.
- Ser capaz de identificar e usar a hipótese alternativa quando testar hipóteses. Entender como escolher entre uma hipótese alternativa unilateral e uma bilateral e ser capaz de explicar sua escolha.

FAZER

- Saber como executar um teste-*z* de uma proporção.
- Ser capaz de interpretar os resultados de um teste-*z* de uma proporção.

RELATAR

- Ser capaz de interpretar o significado de um valor-P numa linguagem não técnica, deixando claro que a afirmação da probabilidade diz respeito aos valores calculados sob a suposição de que o modelo nulo seja verdadeiro e não sobre a população do parâmetro de interesse.

Ajuda tecnológica

Os testes de hipóteses para proporções são tão fáceis e naturais que muitos pacotes estatísticos não oferecem comandos especiais para eles. A maioria dos programas estatísticos quer saber o *status* de "sucesso" ou "fracasso" para cada caso. Geralmente, esses são dados como 1 ou 0, mas podem ser uma categoria de palavras como "sim" e "não". Muitas vezes, conhecemos apenas a proporção de sucessos, $\hat{p}$, e a contagem total, *n*. Os pacotes computacionais não lidam naturalmente com dados resumidos como esses, mas veja na próxima página duas importantes exceções (Minitab e JMP).

Em alguns programas, você pode reconstruir os valores originais. No entanto, mesmo quando você tiver reconstruído (ou puder reconstruir) os valores a partir dos dados brutos, em geral, não terá *exatamente* o mesmo teste estatístico de um pacote computacional que você encontraria calculando à mão. Isso acontece porque os pacotes fazem algumas aproximações ao tratar a proporção como uma média. O resultado é muito próximo, mas não igual. Se você usar um pacote computacional, irá notar pequenas discrepâncias entre as suas respostas e as respostas do final do livro, mas elas não são importantes.

Os relatórios sobre testes de hipóteses criados por tecnologias não seguem uma forma padrão. A maioria irá nomear o teste e fornecer o valor da estatística teste, seu desvio padrão e o valor-P. No entanto, esses elementos podem não estar identificados claramente. Por exemplo, a expressão "Prob > |z|" significa que a probabilidade (a "Prob") de observar uma estatística teste cuja magnitude (o valor absoluto nos diz isso) é maior que o (o "z") encontrado nos dados (o qual segue um modelo Normal, pois está escrito como "z"). Essa é uma maneira especial (e não muito clara) de apresentar o valor-P. Em alguns pacotes, você pode especificar que o teste é unilateral. Outros podem relatar três valores-P, abrangendo os testes unilaterais e os bilaterais.

Às vezes, os resultados de um teste de hipótese e o intervalo de confiança são dados automaticamente juntos. O intervalo de confiança deve ser para um nível de confiança correspondente a $1 - \alpha$.

Muitas vezes, o desvio padrão da estatística é chamado de "erro padrão"; em geral, isso é apropriado, porque tivemos de estimar seu valor a partir dos dados. Entretanto, esse não é o caso para proporções: obtemos o desvio padrão para uma proporção do valor suposto na hipótese nula. Todavia, você pode ver o desvio padrão ser chamado de "erro padrão" mesmo para testes com proporções.

É comum que os pacotes estatísticos e as calculadoras apresentem mais dígitos de "precisão" do que poderia ter sido encontrado dos dados. Você pode ignorá-los. Arredonde valores como os do desvio padrão para um dígito a mais do que os existentes nos seus dados.

Eis os tipos de resultados que você poderá ver em saída comum de computador.

Geralmente, o teste recebe um nome

$\hat{p}$

```
Test of p = 0,5
              Value     Test Stat   Prob > |z|
Estimate      0,467     -0,825      0,42
Std Err       0,04073
Upper 95%     0,547
Lower 95%     0,387
```

Na verdade, um desvio padrão porque este é um teste

Estes são limites para o IC de 95% porque α = 0,05 – um fato não declarado claramente

Valor da estatística teste

Valor-P

Alternativa bilateral

Excel

Métodos de inferência para proporções não fazem parte do conjunto de ferramentas padrão do Excel.

Comentários

Para dados resumidos, digite os cálculos em qualquer célula e os avalie.

Você pode usar o suplemento DDXL para realizar o teste de hipóteses para proporções. Selecione as variáveis a serem testadas. Do menu **DDXL**, escolha **Hypothesis Tests**. Na caixa de diálogo **Hypothesis Tests**, escolha **1 Var Prop Test**. Especifique a variável a ser testada e clique em **OK**.

Minitab

Escolha **Basic Statistics** do menu **Stat.**
- Escolha **1 Proportion** do submenu **Basic Statistics**.
- Se os dados forem nomes categóricos em uma variável, atribua a variável da caixa com a lista de variáveis à caixa **Samples in columns.**
- Se você tiver dados resumidos, clique em **Summarized Data** e preencha o número de tentativas e o de sucessos.
- Clique em **Options** e especifique os detalhes restantes.
- Se você tiver uma amostra grande, marque a opção: **Use test and Interval based on Normal distribution.**
- Clique em **OK**.

Comentários

Ao trabalhar com uma variável categórica, o MINITAB trata a última categoria como a categoria "sucesso". Você pode especificar como as categorias devem ser ordenadas.

SPSS

O SPSS não fornece testes de hipóteses para proporções.

JMP

Para uma variável **categórica** que tem rótulos categóricos, a plataforma **Distribution** inclui testes e intervalos para proporções. Para dados resumidos, coloque os nomes das categorias em uma variável e as frequências em uma variável ao lado. Especifique que a coluna da frequência faça o **papel da frequência**. Então, use a plataforma ***Distribution***.

Comentário

O JMP usa métodos levemente diferentes para inferências de proporções do que os discutidos neste livro. Suas respostas provavelmente serão um pouco diferentes.

DATA DESK

O Data Desk não oferece métodos embutidos para fazer inferência com proporções. O comando **Replicate Y by X** no menu **Manip** irá "reconstruir" frequências de dados que foram agrupados para que você possa exibi-las.

Comentários

Para dados resumidos, abra um **Scrachtpad** a fim de calcular o desvio padrão e a margem de erro digitando os cálculos. Depois, execute o teste com o **z-interval for individual s** encontrado no comando **Test**.

Projetos de estudo de pequenos casos

Produção de metal

Lingotes são enormes pedaços de metal, geralmente ultrapassando 20000 libras, feitos em um molde gigante. Eles devem ser moldados em grandes pedaços para serem utilizados na fabricação de estruturas de carros e aviões. Se eles racharem durante a fabricação, a rachadura pode propagar-se até a zona necessária para a peça, comprometendo sua integridade. Os fabricantes de aeronaves exigem que o metal para seus aviões não tenha defeitos; portanto, o lingote deve ser feito novamente, se uma rachadura for detectada no molde.

Embora o metal do lingote rachado seja reciclado, o custo da sucata atinge dezenas de milhares de dólares. Os fabricantes de metais querem evitar rachaduras sempre que possível. Porém, o processo de moldagem é complicado e nem tudo pode ser controlado. Numa usina, apenas aproximadamente 75% dos lingotes estão livres de rachaduras. Numa tentativa de reduzir a proporção das rachaduras, os engenheiros e cientistas da usina recentemente (em janeiro de 2006) fizeram mudanças no processo de moldagem. Os dados de 500 lingotes produzidos desde que as mudanças ocorreram são encontrados no arquivo **ch11_MCSP_Ingots**. A variável rachadura (*Crack*) indica se uma rachadura foi encontrada (1) ou não (0). Selecione uma amostra aleatória de 100 lingotes e teste a afirmação de que a taxa de rachaduras diminui em 25%. Encontre também o intervalo de confiança para a taxa de rachaduras. Selecione uma amostra aleatória de 1000 lingotes, teste a afirmação e encontre o intervalo de confiança novamente. Compare os dois testes e os intervalos e prepare um breve relatório sobre o que você encontrou, incluindo as diferenças (se houver alguma) que você percebeu nas duas amostras.

Programa de lealdade

Um gerente de *marketing* enviou 10 mil correspondências a uma amostra aleatória de clientes para testar um novo programa de fidelidade baseado na rede. Os clientes ou não receberam nada (Nenhuma Oferta), ou ganharam uma passagem aérea de uma companhia associada ou receberam um seguro aéreo grátis no próximo voo (Seguro Grátis). O responsável pela seleção dos 10 mil clientes assegurou ao gerente de *marketing* que a amostra é representativa de vários segmentos da base de clientes. Entretanto, o gerente está preocupado, com o fato de a oferta não ter sido enviada a clientes suficientes do segmento *Viagem*, que representa 25% de toda a base (Variável *Spending.Segment*). Além disso, ele está preocupado porque acredita que menos de um terço dos clientes daquele segmento de fato não receberam oferta alguma. Usando os dados encontrados em **ch11_MCSP_Loyalty_Program,** escreva um breve relatório testando a hipótese apropriada e resumindo seus achados. Inclua no seu relatório um intervalo de confiança de 95% para a proporção dos clientes que responderam à oferta se registrando para o programa de fidelidade. (A variável *Response* [Resposta] indica 1 para respondentes e 0 para não respondentes.)

EXERCÍCIOS

1. Hipóteses. Escreva a hipótese nula e a alternativa para testar cada uma das seguintes situações.

a) Uma empresa de roupas *on-line* está preocupada com a pontualidade da entrega dos seus produtos. A vice-presidente de operações e *marketing* recentemente declarou que queria que o percentual dos produtos enviados dentro do prazo fosse de pelo menos 90%. Ela deseja saber se a empresa obteve sucesso.

b) Uma imobiliária recentemente anunciou que a proporção de casas que levam mais de três meses para serem vendidas é, agora, maior que 50%.

c) O setor de contabilidade de uma empresa financeira relata uma taxa de erro abaixo de 2%.

2. Mais hipóteses. Escreva a hipótese nula e a alternativa para testar cada uma das seguintes situações.

a) Um artigo de uma revista de negócios relata que, em 1990, 35% dos CEOs tinham MBA. Esse percentual mudou?

b) Recentemente, 20% dos carros de certo modelo exigiram reparos caros na transmissão depois de atingirem 50 mil e 100 mil milhas. O fabricante do carro espera que o novo projeto de um componente de transmissão tenha solucionado o problema.

c) Um pesquisador de mercado para um fabricante de cola decide fazer um teste de campo de um novo sabor de refrigerante, planejando comercializá-lo somente se tiver certeza de que 60% das pessoas irão gostar do sabor.

3. Entregas. A empresa de roupas do Exercício 1a analisa a amostra dos relatórios de entrega. Ela testa a hipótese de que 90% das entregas são feitas dentro do prazo *versus* a alternativa de que mais de 90% são entregues dentro do prazo e encontram o valor-P de 0,22. Qual destas conclusões é apropriada?

a) Existe uma probabilidade de 22% de que 90% das entregas sejam feitas dentro do prazo.

b) Existe uma probabilidade de 78% de que 90% das entregas sejam feitas dentro do prazo.

c) Existe uma probabilidade de 22% de que a amostra coletada mostre o percentual correto das entregas dentro do prazo.

d) Existe uma probabilidade de 22% de que a variação amostral natural produza uma amostra com uma proporção observada de entregas dentro do prazo como a obtida se, de fato, 90% das entregas são feitas dentro do prazo.

4. Vendas de casas. A imobiliária do Exercício 1b analisa uma amostra recente de casas que vendeu. Testando a hipótese nula de que 50% das casas levam mais de três meses para serem vendidas *versus* a hipótese de que mais de 50% das casas levam mais de três meses para serem vendidas, eles encontram um valor-P de 0,034. Qual das seguintes conclusões é adequada?

a) Existe uma probabilidade de 3,4% de que 50% das casas demorem mais de três meses para serem vendidas.

b) Se 50% das casas levam mais de três meses para serem vendidas, existe uma probabilidade de 3,4% de que uma amostra aleatória produza uma proporção amostral tão alta quanto a que eles obtiveram.

c) Existe uma probabilidade de 3,4% de que a hipótese nula esteja correta.

d) Existe uma probabilidade de 96,6% de que 50% das casas levem mais de três meses para serem vendidas.

5. Valor-P. As multas mais severas e as campanhas publicitárias aumentaram o uso do cinto de segurança entre motoristas e passageiros? As observações do trânsito de pessoas indo e vindo do trabalho falharam em encontrar evidências de uma mudança significativa comparada a três anos atrás. Explique o que o valor-P de 0,17 do estudo significa nesse contexto.

6. Outro valor-P. Uma empresa que desenvolve escâneres para procurar armas escondidas nos aeroportos conclui que um novo dispositivo é significativamente melhor do que o escâner atual. A empresa tomou essa decisão com base num valor-P de 0,003. Explique o significado do valor-P nesse contexto.

7. Campanha de publicidade. Uma analista da tecnologia da informação acredita estar perdendo clientes do seu *site* devido ao complicado sistema de compras e pagamento. Ela adiciona o recurso de um clique ao *site*, para torná-lo mais amigável, mas descobre que somente 10% dos clientes o usam. Ela decide lançar uma campanha publicitária de conscientização para divulgar aos consumidores o novo recurso na esperança de aumentar o percentual. Ao não perceber uma diferença significativa de comportamento, ela contrata um consultor para ajudá-la. O consultor seleciona uma amostra aleatória de compradores recentes, testa a hipótese de que as campanhas não produziram qualquer mudança *versus* a alternativa de que o percentual que usa o recurso de um clique é agora maior que 10% e encontra um valor-P de 0,22. Qual conclusão é apropriada? Explique.

a) Existe uma probabilidade de 22% de que a campanha tenha funcionado.

b) Existe uma probabilidade de 78% de que a campanha tenha funcionado.

c) Existe uma probabilidade de 22% de que a hipótese nula seja verdadeira.

d) Existe uma probabilidade de 22% de que a variação amostral poderia produzir resultados de pesquisa de opinião como esses se o uso do recurso de um clique tivesse aumentado.

e) Existe uma probabilidade de 22% de que a variação amostral poderia produzir resultados de opinião como esses caso realmente não tenha havido mudança no uso do *site*.

8. Fundos mútuos. Uma gerente de fundos mútuos afirma que pelo menos 70% das ações que ela seleciona irão subir de preço no próximo ano. Examinamos uma amostra de 200 ações da sua seleção dos últimos três anos. Nosso valor-P é 0,03. Teste uma hipótese apropriada. Qual conclusão é adequada? Explique.

a) Existe uma probabilidade de 3% de que a gerente do fundo esteja correta.

b) Existe uma probabilidade de 97% de que a gerente do fundo esteja correta.

c) Existe uma probabilidade de 3% de que uma amostra aleatória possa produzir os resultados que observamos, assim, é razoável concluir que a gerente do fundo está correta.

d) Existe uma probabilidade de 3% de que uma amostra aleatória possa produzir os resultados que observamos se $p = 0,7$, assim, é razoável concluir que a gerente do fundo não está correta.

e) Existe uma probabilidade de 3% de que a hipótese nula esteja correta.

9. Eficácia do produto. A antiga fórmula do antiácido de uma empresa farmacêutica dá alívio a 70% das pessoas que a consomem. A empresa testa uma fórmula nova para verificar se ela é melhor e obtém um valor-P de 0,27. É razoável concluir que a nova fórmula e a antiga são igualmente eficazes? Explique.

10. Vendas de carros. Uma empresa de automóveis alemã espera vender uma quantidade maior de carros ao segmento mais jovem do mercado – motoristas com menos de 20 anos. Os pesquisadores de mercado da empresa fazem uma pesquisa para investigar se a proporção dos formandos do Ensino Médio que têm o seu próprio carro é maior ou não do que era há uma década. Eles encontram um valor-P de 0,017. É razoável concluir que mais formandos do Ensino Médio têm carros? Explique.

11. Afirmações falsas? Uma empresa de doces afirma que num pacote grande dos M&M's® de Natal, metade dos doces é vermelha e metade é verde. Você escolhe os doces ao acaso de um pacote e descobre que, dos primeiros 20 que comeu, 12 eram vermelhos.

a) Se for verdade que metade é vermelha e metade é verde, qual é a probabilidade de que você tenha encontrado pelo menos 12 dos 20 vermelhos?

b) Você acha que metade dos M&M's® no pacote são realmente vermelhos? Explique.

12. Raspadinha. Um varejista oferece uma promoção da "raspadinha". Entrando na loja, você recebe um cartão. Ao pagar, pode raspá-lo. A empresa informa que a metade dos cartões é vencedora e tem um prêmio imediato de \$5 (os demais oferecem \$1 para qualquer compra futura de café na cafeteria). Você não está seguro de que o percentual é realmente de 50% de vencedores.

a) Na primeira vez em que faz compras na loja, você ganha o cupom do café. Você tenta novamente e ganha o cupom do café. Dois fracassos seguidos o convencem de que a fração verdadeira dos vencedores não é de 50%? Explique.

b) Você tenta uma terceira vez e ganha café novamente! Qual é a probabilidade de não ganhar o dinheiro três vezes seguidas se a metade dos cartões realmente oferece dinheiro?

c) Três perdas seguidas o convenceriam de que a loja está trapaceando?

d) Quantas vezes seguidas você teria de ganhar o cupom do café, em vez do dinheiro, para estar certo de que a empresa não está cumprindo o percentual de vencedores prometido? Justifique sua resposta calculando a probabilidade e explicando o que ela significa.

13. Pesquisa de opinião da Spike. Em agosto de 2004, a revista *Time* publicou os resultados de uma pesquisa de opinião por telefone encomendada pela rede de comunicações Spike. Dos 1302 homens que responderam, somente 39 disseram que a medida mais importante de sucesso era o seu trabalho.

a) Estime o percentual de todos os homens norte-americanos que avaliam o sucesso principalmente pelo seu trabalho. Use um intervalo de confiança de 98%. Não se esqueça de verificar primeiro as condições.

b) Alguns acreditam que poucos homens contemporâneos julgam o seu sucesso principalmente pelo seu trabalho. Suponha que queremos conduzir um teste de hipótese para ver se a fração está abaixo da marca de 5%. O que seu intervalo de confiança indica? Explique.

c) Qual é o nível de significância para esse teste? Explique.

14. Ações. Uma jovem investidora do mercado de ações se preocupa com o fato de que investir no mercado de ações é, na verdade, apostar, já que a chance de o mercado de ações subir em um dia qualquer é de 50%. Ela decide monitorar sua ação preferida por 250 dias e descobre que, em 140 dias, a ação esteve em "alta".

a) Encontre o intervalo de confiança para a proporção dos dias em que a ação esteve em "alta." Não esqueça de primeiro verificar as condições.

b) O seu intervalo de confiança fornece qualquer evidência de que o mercado não é aleatório? Explique.

c) Qual é o nível de significância para esse teste? Explique.

15. Economia. Em 2008, a Gallup Poll perguntou a 2336 adultos norte-americanos com 18 anos ou mais como eles avaliavam as condições econômicas. Numa pesquisa de opinião conduzida de 27 de janeiro a 1° fevereiro de 2008, 24% avaliaram a economia como Excelente/Boa. Um meio de comunicação declarou recentemente que o percentual de norte-americanos que achavam o estado da economia Excelente/Bom era, de fato, 28%. A Gallup Poll apoia esta afirmação?

a) Teste a hipótese apropriada. Encontre um intervalo de confiança de 95% para a proporção amostral de adultos norte-americanos que avaliaram a economia como Excelente/Boa. Verifique as condições.

b) O intervalo de confiança fornece evidência para apoiar a afirmação?

c) Qual é o nível de significância do teste em b? Explique.

16. Economia, parte 2. Os mesmos dados da Gallup Poll do Exercício 15 também relataram que 33% dos pesquisados avaliaram a economia como Ruim. O mesmo meio de comunicação declarou que a proporção verdadeira era 30%. A Gallup Poll apoia esta afirmação?

a) Teste a hipótese apropriada. Encontre um intervalo de confiança de 95% para a proporção amostral dos adultos norte-americanos que avaliaram a economia como Ruim. Verifique as condições.

b) O seu intervalo de confiança fornece evidência para apoiar essa afirmação?

c) Qual é o nível de significância para o teste em b? Explique.

17. Alfa conveniente. Um executivo júnior entusiasta executou um teste do seu novo programa de *marketing*. Ele relata ter provocado um aumento "significativo" nas vendas. Uma nota de rodapé no seu relatório explica que ele usou um nível alfa de 7,2% para o seu teste. Presumivelmente, executou um teste de hipótese *versus* a hipótese nula de nenhuma mudança nas vendas.

a) Se ele tivesse usado um nível alfa de 5%, seria mais ou menos provável que rejeitasse a hipótese nula? Explique.

b) Se ele tivesse escolhido o nível alfa de 7,2% para poder declarar significância estatística, explique por que isso não seria ético.

18. Segurança. O fabricante de uma nova pílula para dormir suspeita que ela possa aumentar o risco de sonambulismo, o que poderia ser perigoso. Num teste do medicamento, não foi possível rejeitar a hipótese nula do aumento do sonambulismo quando testado com um alfa = 0,01.

a) Se o teste tivesse sido executado com um alfa = 0,05, teria maior ou menor probabilidade de rejeitar a hipótese nula de que não houve aumento do sonambulismo?

b) Qual nível alfa você acha que o fabricante deveria usar? Por quê?

19. Teste de um produto. Visto que muitas pessoas têm dificuldade em programar seus videocassete, uma empresa de eletrônica desenvolveu instruções mais simples. O objetivo é que pelo menos 96% dos seus clientes tenham sucesso na programação dos seus vídeos. A empresa testa o novo sistema com 200 pessoas e 188 delas têm sucesso em programar o aparelho. Isso é uma forte evidência de que o novo sistema falha na satisfação do objetivo da empresa? Um teste dessa hipótese, executado por um estudante, é mostrado aqui. Quantos erros você consegue achar?

$H_0: \hat{p} = 0{,}96$

$H_A: \hat{p} \neq 0{,}96$

Condição de pelo menos 10 S/F, $0{,}96(200) > 10$

$$\frac{188}{200} = 0{,}94;\ DP(\hat{p}) = \sqrt{\frac{(0{,}94)(0{,}06)}{200}} = 0{,}017$$

$$z = \frac{0{,}96 - 0{,}94}{0{,}017} = 1{,}18$$

$$\mathrm{P} = P(z > 1{,}18) = 0{,}12$$

Existe uma forte evidência de que o novo sistema não funciona.

20. *Marketing*. Em novembro de 2001, o boletim informativo da *Ag Globe Trotter* relatou que 90% dos adultos tomam leite. Uma organização regional de fazendeiros, planejando uma nova campanha de *marketing* em vários municípios, faz uma pesquisa de opinião com uma amostra aleatória de 750 adultos moradores da região. Da amostra, 657 pessoas relataram que tomam leite. Essas respostas fornecem uma forte evidência de que o percentual de 90% não é correto para essa região? Corrija os erros na tentativa abaixo, de um estudante, de verificar a hipótese apropriada.

$H_0: \hat{p} = 0{,}9$

$H_A: \hat{p} < 0{,}9$

Condição de pelo menos 10 S/F, $750 > 10$

$$\frac{657}{750} = 0{,}876;\ SD(\hat{p}) = \sqrt{\frac{(0{,}88)(0{,}12)}{750}} = 0{,}012$$

$$z = \frac{0{,}876 - 0{,}94}{0{,}012} = -2$$

$$\mathrm{P} = P(z > -2) = 0{,}977$$

Existe uma probabilidade maior do que 97% de que o percentual declarado esteja correto para essa região.

21. Meio ambiente. Nos anos 1980, acreditava-se, de modo geral, que as anormalidades congênitas afetavam 5% das crianças do país. Alguns acreditam que o aumento no número de produtos químicos no meio ambiente leva a um aumento na incidência de anormalidades. Um estudo recente examinou 384 crianças e encontrou que 46 delas mostravam sinais de anormalidade. Isso é uma evidência forte de que o risco aumentou? (Consideramos um valor-P de aproximadamente 5% para representar uma evidência razoável.)

a) Escreva a hipótese apropriada.
b) Verifique as suposições necessárias.
c) Execute a mecânica do teste. Qual é o valor-P?
d) Explique cuidadosamente o que o valor-P significa nesse contexto.
e) Qual é a sua conclusão?
f) Os produtos químicos no meio ambiente causam anormalidades congênitas?

22. Empresa de cobrança. Uma empresa que cobra contas para os consultórios médicos na região está preocupada com o fato de o percentual das contas pagas pela Medicare ter aumentado. Historicamente, o percentual tem sido de 31%. Um exame em 8368 contas recentes revelou que 32% delas estão sendo pagas pela Medicare. Isso é evidência de uma mudança no percentual de contas pagas pela Medicare?

a) Escreva a hipótese adequada.
b) Verifique as suposições e condições.
c) Execute o teste e encontre o valor-P.
d) Declare sua conclusão.
e) Você acha que essa diferença é significativa? Explique.

23. Educação. O Centro Nacional para a Estatística Educacional monitora vários aspectos da educação primária e secundária dos Estados Unidos. Os números de 1996 são geralmente usados como referência para avaliar mudanças. Em 1996, 34% dos estudantes não faltaram nenhuma aula durante o mês anterior. Na pesquisa de 2000, as respostas de 8302 estudantes mostraram que esse número caiu para 33%. As autoridades ficaram preocupadas com o fato de a frequência dos estudantes estar declinando. Esses números fornecem uma evidência do decréscimo da frequência dos estudantes?

a) Escreva a hipótese apropriada.
b) Verifique as suposições e condições.
c) Execute o teste e encontre o valor-P.
d) Declare a sua conclusão.
e) Você acha que essa diferença é significativa? Explique.

24. Confiança do consumidor. Várias vezes, em 2007, quando perguntados se as condições econômicas estavam melhores ou piores, mais de 20% dos adultos norte-americanos disseram que estavam melhores. De 19 a 20 de janeiro de 2008, quando a Gallup fez uma pesquisa de opinião entre 2590 adultos norte-americanos, somente 13% afirmaram que as condições estavam melhorando. Essas respostas dão evidência de que a confiança do consumidor diminuiu do nível de 2007?

a) Escreva a hipótese apropriada.
b) Verifique as suposições e condições.
c) Execute o teste e encontre o valor-P.
d) Declare a sua conclusão.
e) Essa diferença é significativa? Explique.

25. Aposentadoria. Um levantamento de dados com 1000 trabalhadores indicou que aproximadamente 520 investiram num plano individual de aposentadoria. Os dados nacionais sugerem que 44% dos trabalhadores investem em planos individuais de aposentadoria.

a) Crie um intervalo de confiança de 95% para a proporção de trabalhadores que investiram em planos individuais de aposentadoria com base no levantamento de dados.
b) Isso fornece evidência de uma mudança no comportamento dos trabalhadores? Usando o seu intervalo de confiança, teste uma hipótese apropriada e declare a sua conclusão.

26. Satisfação do cliente. Uma empresa espera melhorar a satisfação dos clientes, estabelecendo como objetivo não mais que 5% de comentários negativos. Um levantamento aleatório com 350 clientes encontrou somente 10 clientes com reclamações.

a) Crie um intervalo de confiança de 95% para o nível verdadeiro de descontentamento entre os clientes.

b) Isso fornece evidência de que a empresa alcançou seu objetivo? Usando o seu intervalo de confiança, teste a hipótese apropriada e formule a sua conclusão.

27. Custo de manutenção. Uma empresa de limusines está preocupada com o aumento dos custos de manutenção de sua frota de 150 carros. Após um teste, a empresa descobriu que os sistemas de emissão de 7 em 22 carros que eles testaram não satisfizeram as diretrizes do controle de poluição. Eles previram custos assumindo que um total de 30 carros precisaria de modificações para satisfazer as diretrizes atuais. Isso é uma evidência forte de que mais de 20% da frota pode estar fora das especificações? Teste a hipótese apropriada e formule sua conclusão. Certifique-se de que as suposições e condições apropriadas estejam satisfeitas antes de você proceder.

28. Mercadorias danificadas. Um fabricante de eletrodomésticos armazena suas lavadoras e secadoras num grande depósito para o envio às lojas varejistas. Às vezes, os aparelhos são danificados no seu manuseio. Embora o dano seja pequeno, a empresa deve vender essas máquinas a preços muito menores. O objetivo da empresa é manter a proporção das máquinas danificadas abaixo de 2%. Certo dia, um inspetor verifica aleatoriamente 60 lavadoras e descobre que 5 delas têm arranhões ou estão amassadas. Isso é uma evidência forte de que o depósito está falhando na satisfação do objetivo da empresa? Teste a hipótese apropriada e formule sua conclusão. Tenha certeza de que as suposições e condições apropriadas estejam satisfeitas antes de proceder.

29. Produtos com defeito. Um relatório interno de uma empresa de manufatura indicou que aproximadamente 3% de todos os produtos estavam com defeito. Os dados de um lote encontraram somente 7 produtos defeituosos em 469 produtos avaliados. Isso está consistente com o relatório? Teste a hipótese apropriada e formule sua conclusão. Tenha certeza de que as suposições e condições apropriadas estejam satisfeitas antes de proceder.

30. Empregos. O departamento de contabilidade de uma grande universidade estadual quer anunciar publicamente que mais de 50% dos seus formandos obtiveram uma oferta de emprego antes da graduação. Uma amostra de 240 graduados recentes indicou que 138 tiveram uma oferta de emprego antes da graduação. Teste a hipótese apropriada e formule sua conclusão. Tenha certeza de que as suposições e condições apropriadas estejam satisfeitas antes de proceder.

31. *WebZine*. Uma revista chamada *WebZine* está considerando o lançamento de uma edição *on-line*. A revista planeja investir somente se estiver convencida de que mais de 25% dos leitores atuais fariam uma assinatura. A revista entrevistou uma amostra aleatória simples de 500 assinantes atuais e 137 deles demonstraram interesse. O que ela deveria fazer? Teste a hipótese apropriada e formule sua conclusão. Certifique-se de que as suposições e condições apropriadas estejam satisfeitas antes de proceder.

32. A verdade na propaganda. Um centro de jardinagem quer armazenar sobras de pacotes de sementes de vegetais para vender na próxima primavera, mas está preocupado que as sementes não germinem na mesma taxa do ano passado. O gerente encontra um pacote de sementes de ervilhas do ano passado e as planta, a fim de testá-las. Embora o pacote alegue uma taxa de germinação de 92%, somente 171 das 200 sementes testadas germinaram. Isso é uma evidência de que as sementes perderam vitalidade durante o ano em armazenamento? Teste a hipótese apropriada e formule sua conclusão. Certifique-se de que as suposições e condições apropriadas estejam satisfeitas antes de proceder.

33. Mulheres executivas. Uma empresa é criticada porque somente 13 das 43 pessoas em posições de liderança são mulheres. A empresa explica que, embora essa proporção seja mais baixa do que ela gostaria, não é surpreendente, dado que somente 40% dos seus empregados são mulheres. O que você acha? Teste a hipótese apropriada e formule sua conclusão. Certifique-se de que as suposições e condições apropriadas estejam satisfeitas antes de proceder.

34. Júri. Os dados do censo para certo condado norte-americano mostram que 19% dos residentes adultos são hispânicos. Suponha que 72 pessoas são chamadas para servir como jurados e somente 9 delas sejam hispânicas. Essa aparente falta de representação dos hispânicos questiona a integridade do sistema de seleção dos jurados? Explique.

35. Sem fins lucrativos. Uma empresa sem fins lucrativos, preocupada com as taxas de abandono dos estudos nos Estados Unidos, designou um programa de monitoramento direcionado a alunos entre 16 e 18 anos de idade. O Centro Nacional para a Estatística da Educação relatou que a taxa de abandono do Ensino Médio no país para o ano de 2000 foi de 10,9%. Uma escola do distrito, que adotou o programa de monitoramento sem fins lucrativos e cuja taxa de abandono tem sempre estado muito próxima à média nacional, relatou em 2004 que 175 dos seus 1782 estudantes abandonaram a escola. A experiência deles é uma evidência de que o programa de monitoramento é eficaz? Explique.

36. Setor imobiliário. Uma revisa nacional do setor imobiliário anunciou que 15% das pessoas que compram a sua primeira casa têm uma renda familiar abaixo de \$40000. Uma empresa imobiliária nacional acredita que esse percentual é muito baixo e amostra 100 dos seus registros. A empresa encontra que 25 dos compradores da primeira casa realmente tinham uma renda familiar abaixo de \$40000. A amostra sugere que a proporção de compradores da primeira casa com uma renda menor do que \$40000 é maior que 15%? Comente e escreva suas próprias conclusões com base num intervalo de confiança apropriado e num teste de hipóteses. Inclua qualquer suposição que você tenha feito sobre os dados.

37. Relações públicas. De acordo com o Departamento de Transportes dos Estados Unidos (DOT), os passageiros fizeram mais reclamações sobre o serviço das companhias aéreas em 2007 do que em 2006. O departamento de relações públicas de uma companhia aérea relata que sua empresa raramente perde a bagagem. Além disso, afirma que, quando isso ocorre, 90% das vezes as malas são recuperadas num período de 24 horas. Um grupo de consumidores faz um levantamento de dados com um grande grupo de passageiros e encontra que 103 de 122 pessoas que perderam sua bagagem receberam seus itens perdidos dentro de 24 horas. Isso lança dúvidas sobre a afirmação da companhia aérea? Explique.

38. Comerciais na TV. Uma nova empresa quer comercializar uma impressora. Ela decide apostar em comerciais durante o Super Bowl. A empresa espera que o reconhecimento da sua marca justifique o alto custo dos comerciais. O objetivo da empresa é que 40% do público reconheça o nome da sua marca e o associe a equipamentos de computador. No primeiro dia após o jogo, um pesquisador de opinião pública escolhe aleatoriamente 420 adultos e descobre que 181 deles sabem que essa empresa manufatura impressoras. Você recomendaria que a empresa continuasse a anunciar durante o Super Bowl? Explique.

39. Ética nos negócios. Um estudo relata que 30% dos MBAs recém-contratados enfrentam práticas de negócios contrárias à ética durante o seu primeiro ano no emprego. A diretora de uma faculdade de economia deseja saber se os seus formandos de MBA tinham experiência similar. Ela fez um levantamento de dados com graduados recentes do programa da sua faculdade e descobriu que 27% dos 120 graduados do ano anterior declararam ter se deparado com a prática de negócios contrária à ética no seu ambiente de trabalho. Podemos concluir que as experiências dos seus graduados são diferentes?

40. Ações, parte 2. Um jovem investidor acredita que possa vencer no mercado de ações escolhendo ações que irão aumentar seu valor. Assuma que, em média, 50% das ações selecionadas por um gerente de portfólio irão aumentar durante 12 meses. Das 25 ações que o jovem investidor comprou durante os últimos 12 meses, 14 aumentaram o seu valor. Ele pode afirmar que é melhor na previsão dos aumentos das ações do que um gerente de portfólio comum?

41. Política nos Estados Unidos. As eleições nacionais em 2008 estão aparentemente atraindo mais interesse e debate entre os eleitores do que as eleições anteriores. Uma amostra nacional de 2020 adultos norte-americanos, com 18 anos ou mais, pesquisados por telefone (utilizando linhas fixas e celulares), entre 30 de janeiro e 2 de fevereiro de 2008, pela Gallup, revelou que 71% têm pensado "muito" na próxima eleição para presidente. Existe alguma evidência de que o percentual tenha mudado do marco histórico relatado de 58% durante o mesmo espaço de tempo em 2004?

a) Encontre o escore-z da proporção observada.
b) Compare o escore-z ao valor crítico para um nível de significância de 0,1% usando a alternativa bilateral.
c) Explique sua conclusão.

42. Confiabilidade do iPod. A MacInTouch relatou que várias versões do iPod registraram taxas de falhas de 20% ou mais. De um levantamento de dados dos clientes, o iPod colorido, lançado em 2004, apresentou 64 falhas em 517. Existe evidência de que a taxa de falha para esse modelo possa ser mais baixa do que a taxa de 20% dos modelos anteriores?

a) Encontre o escore-z da proporção observada.
b) Compare o escore-z ao valor crítico para um nível de significância de 0,1% usando a alternativa bilateral.
c) Explique sua conclusão.

43. Testando carros. Um critério de ar limpo requer que as emissões de gás não excedam os limites especificados para vários poluentes. Muitos estados requerem que os carros sejam testados anualmente para garantir que satisfaçam os critérios. Suponha que os controladores do estado façam uma dupla verificação de uma amostra aleatória de carros que uma oficina mecânica suspeita tenha certificado como dentro da norma. Eles irão revogar a licença da mecânica se encontrarem evidências significativas de que a mecânica esteja certificando veículos que não satisfazem os critérios.

a) Nesse contexto, o que é um erro do Tipo I?
b) Nesse contexto, o que é um erro do Tipo II?
c) Que tipo de erro o dono da mecânica consideraria mais sério?
d) Que tipo de erro os ambientalistas considerariam mais sério?

44. Controle de qualidade. Os gerentes de produção de uma linha de montagem devem monitorar a saída para ter certeza de que o nível de produtos defeituosos permaneça pequeno. Eles inspecionam periodicamente uma amostra aleatória de itens produzidos. Se encontrarem um aumento significativo na proporção dos itens que devem ser rejeitados, irão parar o processo de montagem até que o problema possa ser identificado e consertado.

a) Escreva a hipótese nula e alternativa para esse problema.
b) Qual é o erro do Tipo I e do Tipo II nesse contexto?
c) Qual tipo de erro o dono da fábrica consideraria mais sério?
d) Qual tipo de erro os consumidores considerariam mais sério?

45. Testando carros, novamente. Como no Exercício 43, os controladores do estado estão verificando as oficinas mecânicas para ver se elas certificam veículos que não satisfazem os critérios de poluição.

a) Nesse contexto, o que é pretendido pelo poder do teste que os reguladores estão conduzindo?
b) O poder será maior se eles testarem 20 ou 40 carros? Por quê?
c) O poder será maior se eles usarem um nível de significância de 5 ou de 10%? Por quê?
d) O poder será maior se os inspetores da mecânica forem mais ou menos exigentes? Por quê?

46. Controle de qualidade, parte 2. Considere novamente a tarefa dos inspetores de controle qualidade do Exercício 44.

a) Nesse contexto, o que é pretendido pelo poder do teste que os inspetores conduzem?
b) Eles estão atualmente testando cinco itens a cada hora. Alguém propôs que eles testassem dez itens a cada hora. Quais são as vantagens e desvantagens de tal chance?
c) O teste atualmente usa um nível de significância de 5%. Quais são as vantagens e desvantagens de mudar o nível de significância para 1%?
d) Suponha que, com o passar dos dias, uma das máquinas na linha de produção produza mais itens com defeitos. Como isso afetará o poder do teste?

47. *Software* de estatística. Um professor de estatística observou que, ao longo de vários anos, aproximadamente 13% dos estudantes que se matricularam no seu curso de Estatística Introdutória desistem antes do final do semestre. Um vendedor sugeriu que ele tente um pacote estatístico que envolva mais os alunos com o computador, com o intuito de reduzir a taxa de desistência. Como o *software* é caro, o vendedor sugere ao professor que ele o use por um semestre, para ver se a taxa de desistência baixa de modo significativo. O professor terá de pagar pelo *software* somente se continuar com ele.

a) Esse é um teste unilateral ou bilateral? Explique.
b) Escreva a hipótese nula e a alternativa.
c) Nesse contexto, explique o que aconteceria se o professor cometesse um erro do Tipo I.
d) Nesse contexto, explique o que aconteceria se o professor cometesse um erro do Tipo II.
e) O que é pretendido pelo poder desse teste?

48. Anúncios no rádio. Uma empresa quer renovar seu contrato de publicidade com uma estação de rádio local somente se a estação provar que mais de 20% dos residentes da cidade ouviram o anúncio e reconhecem o produto da empresa. A estação de rádio conduz um levantamento de dados aleatório com 400 pessoas por telefone.

a) Quais são as hipóteses?
b) A estação de rádio planeja conduzir esse teste usando um nível de significância de 10%, mas a empresa quer que o nível de significância seja baixado para 5%. Por quê?
c) O que é pretendido pelo poder desse teste?
d) Para qual nível de significância o poder desse teste será maior? Por quê?
e) Eles finalmente chegam ao acordo de $\alpha = 0{,}05$, mas a empresa propõe que a estação de rádio ligue para 600 pessoas, em vez das 400 inicialmente propostas. Isso irá aumentar ou diminuir o risco de um erro do Tipo II? Explique.

49. *Software* estatístico. Do Exercício 47, 203 estudantes se matricularam para o curso de Estatística. Eles usaram o *software* sugerido pelo vendedor e somente 11 desistiram do curso.

a) O professor deve investir no *software*? Sustente sua recomendação com um teste apropriado.
b) Explique o que o seu valor-P significa nesse contexto.

50. Anúncios no rádio, parte 2. A empresa do Exercício 48 entrevista 600 pessoas selecionadas ao acaso e 133 lembram o anúncio.

a) A empresa deve renovar o contrato? Sustente sua recomendação com um teste apropriado.
b) Explique cuidadosamente o que o valor-P significa nesse contexto.

T **51. Gastos dos consumidores, parte 2.** No Capítulo 10, o Exercício 51 construiu um intervalo de confiança para a proporção de clientes que se qualificavam para uma oferta especial gastando mais de $1000 por mês no cartão. Historicamente, o percentual tem sido de 11% e o departamento de finanças quer saber se ele aumentou. Teste a hipótese apropriada e escreva algumas frases com as suas conclusões.

T **52. Arrecadação de fundos.** No Capítulo 10, o Exercício 52 encontrou um intervalo de confiança para a proporção de doadores que tinham 50 anos ou mais. O chefe das finanças diz que o anúncio na revista da *American Association of Retired People* (Associação Americana dos Aposentados – AARP) não valerá o investimento a não ser que pelo menos dois terços dos doadores tenham 50 anos ou mais. Teste a hipótese apropriada e escreva algumas frases com as suas conclusões.

RESPOSTAS DO TESTE RÁPIDO

1 Você não pode concluir que a hipótese nula é verdadeira. Você pode concluir somente que o experimento não foi capaz de rejeitar a hipótese nula. Eles não foram capazes, com base em 12 pacientes, de mostrar que a aspirina era eficaz.

2 A hipótese nula é H_0: $p = 0{,}75$.

3 Com um valor-P de 0,0001, isso é uma evidência muito forte contra a hipótese nula. Podemos rejeitar a H_0 e concluir que a versão aprimorada do medicamento dá alívio a uma proporção maior de pacientes.

4 O parâmetro de interesse é a proporção, p, de todos os clientes inadimplentes que irão pagar suas contas. H_0: $p = 0{,}30$ e H_A: $p > 0{,}30$.

5 Em $\alpha = 0{,}05$, você pode rejeitar a hipótese nula, porque 0,30 está contido no intervalo de confiança de 90% – é plausível que enviar os DVDs não seja mais eficaz do que enviar as cartas.

6 O intervalo de confiança é de 29 a 45%. A estratégia do DVD é mais cara e pode não valer a pena. Não podemos distinguir a taxa de sucesso de 30%, dados os resultados desse experimento, mas 45% representariam um grande progresso. O banco deveria considerar outro teste, aumentando a amostra para conseguir um intervalo de confiança mais estreito.

7 Um erro do Tipo I significaria decidir que a taxa de sucesso do DVD é maior que 30%, quando ela não é. O banco adotaria um método mais caro de cobrança que não é melhor que o método original, uma estratégia menos cara.

8 Um erro do Tipo II significaria decidir que não existe evidência suficiente para afirmar que a estratégia do DVD funciona, quando de fato funciona. O banco falharia na descoberta de um método eficaz para aumentar sua receita das contas dos devedores.

9 Mais alta; quanto maior o tamanho do efeito, maior o poder. É mais fácil detectar uma melhora numa taxa de sucesso de 60% do que numa taxa de 32%.