7 COMPARAÇÃO DE DOIS GRUPOS

A comparação de dois grupos é um tipo muito comum de análise nas ciências sociais e comportamentais. Um estudo pode comparar a renda média para homens e mulheres que têm empregos e experiências similares. Outro estudo pode comparar a proporção de norte-americanos e canadenses que são a favor de leis de controle de armas de fogo. As médias são comparadas para as variáveis quantitativas e as proporções para as variáveis categóricas.

A Seção 7.1 introduz alguns conceitos básicos para a comparação de grupos. A Seção 7.2 ilustra esses conceitos para a comparação de proporções, e a Seção 7.3 para a comparação de médias. O restante do capítulo mostra alguns métodos alternativos úteis para casos especiais.

7.1 PRELIMINARES DA COMPARAÇÃO DE GRUPOS

As mulheres tendem a gastar mais tempo nas tarefas de casa do que os homens? Se for assim, quanto tempo mais? Na Grã-Bretanha em 2005, a Time Use Survey[1] estudou como uma amostra aleatória de britânicos passava o seu tempo em um dia típico. Para aqueles que disseram que trabalhavam o dia todo, a Tabela 7.1 relata a média e o desvio padrão do número médio relatado de minutos gastos por dia cozinhando e limpando. Usamos a Tabela 7.1 para apresentar alguns conceitos básicos para a comparação de grupos.

Análise bivariada com variáveis resposta e explicativa

Dois grupos sendo comparados constituem uma variável **binária** – uma variável que tem somente duas categorias é, algumas vezes, também chamada de **dicotômica**. Na comparação do tempo médio do trabalho doméstico, homens e mulheres são duas categorias da variável binária, sexo. Os métodos para comparar os dois grupos são os casos especiais dos métodos estatísticos **bivariados** – uma variável de saída de algum tipo é analisada para cada categoria de uma variável de entrada.

Tabela 7.1 Minutos diários gastos cozinhando e limpando por homens e mulheres que trabalham em período integral na Grã-Bretanha

Sexo	Tamanho da amostra	Minutos cozinhado e limpando: Média	Minutos cozinhado e limpando: Desvio padrão
Homens	1219	23	32
Mulheres	733	37	16

Da Seção 3.5 (página 73), lembre que uma variável de saída sobre a qual as comparações são feitas é chamada de **variável resposta**. A variável que define o grupo é chamada de **variável explicativa**. Na Tabela 7.1 o tempo gasto cozinhando e limpando é a variável resposta. O sexo do respondente é a variável explicativa.

Amostras dependentes e independentes

Alguns estudos comparam médias ou proporções em dois ou mais pontos no tempo. Por exemplo, um **estudo longitudinal** observa sujeitos várias vezes. Um exemplo é o Framingham Heart Study, que a cada dois anos, desde 1948, observa muitas características de saúde em mais de 5000 adultos de Framingham, Massachusetts. As amostras que contêm os mesmos sujeitos são ditas **dependentes**.

Em um sentido amplo, duas amostras são *dependentes* quando uma equiparação natural ocorre entre cada sujeito nas duas amostras. Geralmente isso acontece quando cada amostra tem os mesmos sujeitos. Mas a equiparação pode também ocorrer quando as duas amostras têm sujeitos diferentes. Um exemplo é a comparação do trabalho de casa de maridos e esposas, os maridos formando uma amostra e suas esposas, a outra.

Mais comumente, as comparações usam **amostras independentes**. Isso significa que as observações em uma amostra são *independentes* daquelas da outra amostra. Os sujeitos nas duas amostras são diferentes e não existe equiparação entre amostras. Um exemplo é a Tabela 7.1. Os sujeitos foram selecionados aleatoriamente e então classificados quanto ao sexo e mensurados em relação ao tempo por eles gasto em diversas atividades. As amostras de homens e mulheres são independentes.

Suponha que você planeja analisar se um programa com tutoria melhora o entendimento de matemática. O delineamento de um estudo administra um teste em uma amostra de estudantes antes e depois do programa. A amostra dos escores resultantes do teste após o programa é, então, *dependente* porque as duas amostras são formadas pelos mesmos sujeitos.

Outro delineamento de um estudo divide aleatoriamente a classe dos estudantes em dois grupos, um dos quais participa do programa com tutoria (o grupo *experimental*) e o outro grupo não (o grupo de *controle*). Após o curso, ambos os grupos fazem um teste em matemática e os escores médios são comparados. As duas amostras são, portanto, *independentes* porque contêm sujeitos diferentes e não existe equiparação entre as amostras.

Esses dois estudos são *experimentais*. Como foi mencionado no final da Seção 2.2 (página 30), muitos estudos das ciências sociais são, ao contrário, *observacionais*. Por exemplo, muitas comparações dos grupos resultam da divisão da amostra em subamostras de acordo com a classificação em uma variável, como sexo, raça ou partido político. A Tabela 7.1 é um exemplo disso. Tais casos são exemplos de estudos **transversais**, que usam um único levantamento de dados para comparar grupos. Se toda a amostra foi selecionada aleatoriamente, então as subamostras são amostras aleatórias independentes das sub-populações correspondentes.

Por que distinguimos entre amostras *independentes* e *dependentes*? Porque as fórmulas do erro padrão para as estatísticas que comparam médias ou comparam proporções são diferentes para os dois tipos de amostras. Com amostras dependentes, respostas equiparadas provavelmente estão associadas. No estudo sobre o programa de tutoria, os estudantes que tiveram um desempenho relativamente bom em um exame provavelmente tendem a ter um desempenho bom no segundo exame. Isso afeta o erro padrão das estatísticas na comparação dos grupos.

Diferenças de estimativas e seu erro padrão

Para comparar duas populações, podemos estimar a diferença entre seus parâmetros. Para comparar as médias populacionais μ_1 e μ_2, tratamos $\mu_2 - \mu_1$ como um parâmetro e o estimamos pela diferença das médias amostrais, $\bar{y}_2 - \bar{y}_1$. Para a Tabela 7.1, a diferença estimada entre o tempo médio populacional diário de cozinhar e limpar para homens e mulheres é igual a $\bar{y}_2 - \bar{y}_1 = 37 - 23 = 14$ minutos.

A distribuição amostral do estimador $\bar{y}_2 - \bar{y}_1$ tem valor esperado $\mu_2 - \mu_1$. Para amostras aleatórias grandes ou para amostras aleatórias pequenas de populações com distribuições normais, esta distribuição amostral tem uma forma normal, como é apresentado na Figura 7.1.

Uma estimativa tem um erro padrão que descreve quão precisamente ele estima um parâmetro. Da mesma forma ocorre para a diferença entre estimativas de duas amostras que apresentam um erro padrão. Para a Tabela 7.1, o erro padrão da distribuição amostral de $\bar{y}_2 - \bar{y}_1$ descreve quão precisamente $\bar{y}_2 - \bar{y}_1 = 14$ estima $\mu_2 - \mu_1$. Se muitos estudos tivessem sidos conduzidos na Grã-Bretanha comparando o tempo diário cozinhando e limpando para homens e mulheres, a estimativa $\bar{y}_2 - \bar{y}_1$ não teria sido igual a 14 minutos para cada um deles. A estimativa teria variado de estudo para estudo. O erro padrão descreve a variabilidade das estimativas de estudos potenciais diferentes do mesmo tamanho.

A seguinte regra geral permite encontrar o erro padrão quando comparamos estimativas de amostras independentes:

> ☑ **Erro padrão da diferença entre duas estimativas**
>
> Para duas estimativas de amostras independentes que têm erros padrão estimados ep_1 e ep_2, a distribuição amostral da sua diferença tem um erro padrão estimado $= \sqrt{(ep_1)^2 + (ep_2)^2}$.

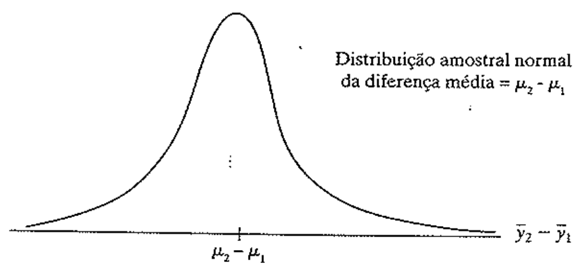
Cada estimativa tem erro amostral e as variabilidades se somam para determinar o erro padrão da diferença das estimativas. A fórmula do erro padrão para amostras dependentes difere desta fórmula e a Seção 7.4 a apresenta.

Lembre que o erro padrão estimado de uma média amostral é igual a

$$ep = \frac{s}{\sqrt{n}},$$

onde s é o desvio padrão da amostra. Considere n_1 o tamanho da primeira amostra e n_2 da segunda. Considere s_1 e s_2 os desvios padrão das duas amostras respectivamente. A diferença $\bar{y}_2 - \bar{y}_1$ entre duas médias amostrais, de amostras independentes, tem o erro padrão estimado igual a:

$$ep = \sqrt{(ep_1)^2 + (ep_2)^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$



☑ **Figura 7.1** Para amostras aleatórias, a distribuição amostral da diferença entre as médias amostrais $\bar{y}_2 - \bar{y}_1$ é aproximadamente normal em torno de $\mu_2 - \mu_1$.

Por exemplo, da Tabela 7.1, o erro padrão estimado da diferença em minutos entre as médias amostrais do tempo de cozinhar e limpar para mulheres e homens é igual a:

$$ep = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(32)^2}{1219} + \frac{(16)^2}{733}} = 1{,}1.$$

Para amostras grandes, a estimativa $\bar{y}_2 - \bar{y}_1$ não iria variar muito de estudo para estudo.

Note-se que o erro padrão da diferença é maior do que o erro padrão para cada estimativa individualmente. Por que isso? Em termos práticos, $(\bar{y}_2 - \bar{y}_1)$ geralmente está mais longe de $(\mu_2 - \mu_1)$ do que $\bar{y}_1$ está de μ_1 ou $\bar{y}_2$ está de μ_2. Por exemplo, suponha que $\mu_1 = \mu_2 = 30$ (desconhecido para nós), mas as médias amostrais são $\bar{y}_1 = 23$ e $\bar{y}_2 = 37$. Então, os erros da estimação seriam

$$\bar{y}_1 - \mu_1 = 23 - 30 = -7 \text{ e}$$
$$\bar{y}_2 - \mu_2 = 37 - 30 = 7,$$

cada estimativa se distanciando 7 unidades da hipótese nula. Mas a estimativa $(\bar{y}_2 - \bar{y}_1) = 37 - 23 = 14$ está 14 de $(\mu_2 - \mu_1) = 0$. O erro de tamanho 14 para a diferença é maior do que o erro de tamanho 7 para cada média individualmente. Suponha que uma média amostral que está 7 unidades distante da média populacional esteja no final da cauda de uma distribuição amostral para uma única média amostral. Então, a diferença entre as médias amostrais que está 14 unidades distante da diferença entre as médias populacionais está no final da cauda da distribuição amostral de $\bar{y}_2 - \bar{y}_1$.

A razão dos parâmetros

Outra forma de comparar duas proporções ou duas médias usa a sua *razão*. A razão é igual a 1,0 quando os parâmetros são iguais. As razões distantes de 1,0 representam efeitos maiores.

Na Tabela 7.1, a razão da média amostral do tempo de cozinhar e limpar para mulheres e homens é 37/23 = 1,61. A média amostral para as mulheres é 1,61 vezes maior que a média amostral para homens. Isso também pode ser expresso dizendo que a média para mulheres é 61% maior do que a média dos homens.

A razão de duas proporções é geralmente chamada de **risco relativo**, porque geralmente é usada nas aplicações de saúde pública para comparar razões de um resultado não desejado entre dois grupos. A razão é geralmente mais informativa do que a diferença quando as duas proporções estão próximas de zero.

Por exemplo, de acordo com dados recentes das Nações Unidas, a taxa anual de homicídios com armas de fogo é 62,4 por milhão de residentes nos Estados Unidos e 1,3 por milhão de residentes na Grã-Bretanha. Na forma de proporção, os resultados são 0,0000624 nos Estados Unidos e 0,0000013 na Grã-Bretanha. A diferença entre as proporções é de 0,0000624 − 0,0000013 = 0,0000611, extremamente pequena. Em contraposição, a razão é 0,0000624/0,0000013 = 624/13 = 48. A proporção de pessoas mortas por armas de fogo nos Estados Unidos foi 48 vezes a proporção na Grã-Bretanha. Neste sentido, o efeito é grande.

Um *software* pode formar um intervalo de confiança para uma razão populacional de médias ou proporções. As fórmulas são complexas e não iremos cobri-las neste livro.

7.2 DADOS CATEGÓRICOS: COMPARANDO DUAS PROPORÇÕES

Vamos, agora, aprender como comparar as proporções de forma inferencial. Considere π_1 a representação da proporção para a primeira população e π_2 a proporção para a segunda população. Considere $\hat{\pi}_1$ e $\hat{\pi}_2$ a representação das proporções amostrais. Você talvez queira revisar as Seções 5.2 (página 134) e 6.3 (página 182) sobre inferências para proporções para o caso de uma amostra.

EXEMPLO 7.1 A oração ajuda pacientes de cirurgia coronária?

Um estudo usou pacientes em seis hospitais dos Estados Unidos que passariam por uma cirurgia de ponte de safena.[2] Os pacientes foram aleatoriamente designados a dois grupos. Para um grupo, voluntários cristãos foram instruídos para rezar por uma cirurgia bem-sucedida com uma recuperação rápida e sem complicações. A oração iniciou na noite anterior à cirurgia e continuou por duas semanas. A resposta foi a ocorrência de complicações médicas que ocorreram em um período de 30 dias após a cirurgia. A Tabela 7.2 resume os resultados.

Existe uma diferença nas taxas de complicações pós-operatórias entre os dois grupos? Considere π_1 a probabilidade de complicações para os pacientes que tiveram um grupo de oração e π_2 dos que não tiveram. Considerando a Tabela 7.2, as proporções amostrais são iguais a:

$$\hat{\pi}_1 = \frac{315}{604} = 0{,}522, \quad \hat{\pi}_2 = \frac{304}{597} = 0{,}509.$$

■

☑ **Tabela 7.2** Complicações que ocorreram com pacientes que sofreram uma cirurgia cardíaca que tinham ou não um grupo de oração

Oração	Complicações	Sem complicações	Total
Sim	315	289	604
Não	304	293	597

Comparamos as probabilidades usando a sua diferença, $\pi_2 - \pi_1$. A diferença das proporções amostrais, $\hat{\pi}_2 - \hat{\pi}_1$, estima $\pi_2 - \pi_1$. Se n_1 e n_2 são relativamente grandes, o estimador $\hat{\pi}_2 - \hat{\pi}_1$ tem uma distribuição amostral que é aproximadamente normal. Veja a Figura 7.2. A média da distribuição amostral é o parâmetro $\pi_2 - \pi_1$ a ser estimado.

Da regra no quadro da Seção 7.1 (página 214), o erro padrão da diferença das proporções amostrais é igual à raiz quadrada da soma dos erros padrão ao quadrado de proporções amostrais separadas. Lembre que o erro padrão estimado de uma única proporção amostral é:

$$ep = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

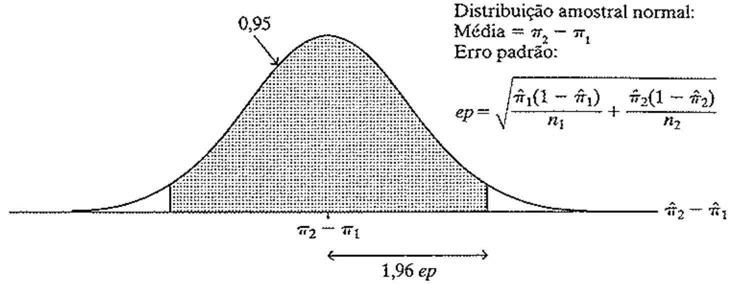
Portanto, a diferença entre as duas proporções tem o erro padrão estimado igual a:

$$ep = \sqrt{(ep_1)^2 + (ep_2)^2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$

Para a Tabela 7.2, $\hat{\pi}_2 - \hat{\pi}_1$ tem o erro padrão estimado igual a:

$$ep = \sqrt{\frac{(0{,}522)(0{,}478)}{604} + \frac{(0{,}509)(0{,}491)}{597}} = 0{,}0288.$$

Para amostras com esses tamanhos, a diferença de proporções amostrais não iria variar muito de estudo para estudo.



☑ **Figura 7.2** Para amostras aleatórias grandes, a distribuição amostral do estimador $\hat{\pi}_2 - \hat{\pi}_1$ da diferença de proporções é aproximadamente normal.

Intervalo de confiança para a diferença de proporções

Assim como com uma única proporção, o intervalo de confiança pega a estimativa por ponto, soma e subtrai a margem de erro que é um escore-z vezes o erro padrão estimado, isto é

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm 1{,}96(ep)$$

para o intervalo de 95% de confiança.

> ☑ **Intervalo de confiança para $\pi_2 - \pi_1$**
>
> Para amostras aleatórias grandes e independentes, um intervalo de confiança $\pi_2 - \pi_1$ entre duas proporções populacionais é:
>
> $$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(ep), \text{ onde}$$
>
> $$ep = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$
>
> O escore-z depende do nível de confiança. Ele vale 1,96 para uma confiança de 95%.

A amostra é grande o suficiente para usar a fórmula se pelo menos dez observações estão na categoria para a qual a proporção é estimada e pelo menos dez observações estão na outra. A maioria dos estudos facilmente satisfaz isso.

EXEMPLO 7.2 Oração e pacientes de cirurgia coronária (continuação)

Para a Tabela 7.2, estimamos a diferença $\pi_2 - \pi_1$ entre a probabilidade de complicações para os pacientes com e sem orações. Visto que $\hat{\pi}_1 = 0{,}522$ e $\hat{\pi}_2 = 0{,}509$, a diferença estimada é igual a $\hat{\pi}_2 - \hat{\pi}_1 = -0{,}013$. Houve uma queda de 0,013 na proporção de complicações entre aqueles que não receberam orações.

Para determinar a precisão dessa estimativa, formamos um intervalo de confiança. Previamente determinamos o $ep = 0{,}0288$. Um intervalo de 95% de confiança para $\pi_2 - \pi_1$ é, então:

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm 1{,}96(ep), \text{ ou } (0{,}509 - 0{,}522) \pm 1{,}96(0{,}0288) = -0{,}013 \pm 0{,}057 \text{ ou } (-0{,}07; 0{,}04)$$

Parece que a diferença está próxima de 0, assim a probabilidade de complicações é similar para os dois grupos. ■

Interpretando um intervalo de confiança a partir da comparação de proporções

Quando o intervalo de confiança para $\pi_2 - \pi_1$ contém 0, como no exemplo anterior, é plausível que $\pi_2 - \pi_1 = 0$. Isto é, é possível que $\pi_1 = \pi_2$. Não existe evidência suficiente para concluir se π_1 ou π_2 é maior. Para o intervalo de confiança para $\pi_2 - \pi_1$ de $(-0{,}07; 0{,}04)$, inferimos que π_2 pode ser tão menor quanto 0,07 ou tão maior quanto 0,04 do que π_1.

Quando um intervalo de confiança para $\pi_2 - \pi_1$ contém somente valores *negativos*, isso sugere que $\pi_2 - \pi_1$ é negativo. Em outras palavras, inferimos que π_2 é *menor* do que π_1. Quando um intervalo de confiança para $\pi_2 - \pi_1$ contém somente valores *positivos*, concluímos que $\pi_2 - \pi_1$ é positivo; isto é, π_2 é *maior* do que π_1.

A denominação de Grupo 1 e Grupo 2 é feita de forma arbitrária. Se considerarmos o Grupo 1 como o sem oração em vez do grupo com oração, então a diferença estimada seria +0,013 em vez de −0,013. O intervalo de confiança teria sido $(-0{,}04; 0{,}07)$, os simétricos dos pontos extremos que obtivemos. Da mesma forma, não interessa se formamos um intervalo de confiança para $\pi_2 - \pi_1$ ou para $\pi_1 - \pi_2$. Se o intervalo de confiança para $\pi_2 - \pi_1$ é $(-0{,}07; 0{,}04)$, então o intervalo de confiança para $\pi_1 - \pi_2$ é $(-0{,}04; 0{,}07)$.

A magnitude dos valores no intervalo de confiança diz a você quão grande é qualquer diferença verdadeira. Se todos os valores no intervalo de confiança estão próximos de 0, como no intervalo $(-0{,}07; 0{,}04)$, inferimos que $\pi_2 - \pi_1$ é pequena, em termos práticos, mesmo se não for exatamente igual a 0.

Como no caso de uma amostra, tamanhos da amostra maiores contribuem para um *ep* menor, uma margem de erro menor e intervalos de confiança mais precisos (menores). Além disso, níveis de eonfiança mais altos geram intervalos de confiança maiores. Para o estudo da oração, um intervalo de 99% de confiança seria igual a $(-0{,}09; 0{,}06)$. Ele é mais amplo (menos preciso) do que um intervalo de 95% de confiança de $(-0{,}07; 0{,}04)$.

Testes de significância sobre $\pi_2 - \pi_1$

Para comparar as proporções populacionais π_1 e π_2, um teste de significância especifica que H_0: $\pi_1 = \pi_2$. Para a diferença das proporções dos parâmetros, esta hipótese é H_0: $\pi_2 - \pi_1 = 0$, *nenhuma diferença* ou *nenhum efeito*.

Sob a suposição H_0 de que $\pi_1 = \pi_2$, estimamos o valor comum de π_1 e π_2 pela proporção amostral das duas amostras. Esse valor é indicado por $\hat{\pi}$. Para ilustrar, utilizando os dados da Tabela 7.2 do estudo da oração, tem-se: $\hat{\pi}_1 = 315/604 = 0{,}522$ e $\hat{\pi}_2 = 304/597 = 0{,}509$. Para as duas amostras consideradas como um todo, segue que:

$$\hat{\pi} = (315 + 304)/(604 + 597) = 619/1201 = 0{,}515.$$

A proporção $\hat{\pi}$ é chamada de *estimativa agrupada ou combinada*, porque agrupa observações das duas amostras.

A estatística-teste mensura o número de erros padrão entre a estimativa e o valor H_0. Tratando $\pi_2 - \pi_1$ como o parâmetro, testamos se $\pi_2 - \pi_1 = 0$; isto é, o valor da hipótese nula do parâmetro $\pi_2 - \pi_1$ é 0. O valor estimado de $\pi_2 - \pi_1$ é $\hat{\pi}_2 - \hat{\pi}_1$. A estatística-teste é:

$$z = \frac{\text{Estimativa} - \text{valor da hipótese nula}}{\text{Erro padrão}} = \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{ep_0}.$$

Em vez de usar o erro padrão do intervalo de confiança, você deveria usar uma fórmula alternativa baseada na suposição declarada em H_0 de que $\pi_1 = \pi_2$. Usamos a notação ep_0 porque é um ep que ocorre na suposição de H_0. Este erro padrão é:

$$ep_0 = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Para os dados da Tabela 7.2, a estimativa para o erro padrão para o teste é igual a:

$$ep_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0{,}515(0{,}485)\left(\frac{1}{604} + \frac{1}{597}\right)} = \sqrt{0{,}000832} = 0{,}0288.$$

A estatística-teste para H_0: $\pi_1 = \pi_2$ é igual a:

$$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{ep_0} = \frac{0{,}509 - 0{,}522}{0{,}0288} = -0{,}43.$$

O valor-*p* depende, na forma usual, de o teste ser bilateral, H_a: $\pi_1 \neq \pi_2$ (isto é, $\pi_2 - \pi_1 \neq 0$) ou unilateral, H_a: $\pi_1 > \pi_2$ (isto é, $\pi_2 - \pi_1 < 0$), ou H_a: $\pi_1 < \pi_2$ (isto é, $\pi_2 - \pi_1 > 0$). A mais comum é a alternativa bilateral. Seu valor-*p* é a soma das probabilidades das duas caudas da distribuição normal padrão que está além do valor da estatística-teste observada. Um escore-*z* de −0,43 tem um valor-*p* bilateral igual a 0,67. Assim, não existe muita evidência contra H_0.

Em resumo, é plausível que a probabilidade de complicações seja a mesma para as condições de oração e sem oração. Entretanto, este estudo não refuta o poder da oração. Exceto pelo fato de que não podemos aceitar a hipótese nula, o experimento não pode controlar muitos fatores, como, por exemplo, se amigos e familiares também estavam rezando para os pacientes.

O teste *z* para comparar proporções funciona muito bem mesmo para tamanhos da amostra relativamente pequenos. Daremos diretrizes detalhadas na Seção 8.2 quando estudaremos um teste mais geral para comparar vários grupos. Para simplificar, você pode usar a diretriz para intervalos de confiança que comparam proporções, ou seja, que cada amostra deva ter pelo menos 10 resultados em cada categoria. Na prática, testes *bilaterais* são robustos e funcionam bem se cada amostra tiver pelo menos cinco resultados em cada categoria.

Tabelas de contingência e probabilidades condicionais

A Tabela 7.2 é um exemplo de uma **tabela de contingência**. Cada linha é uma categoria da variável explicativa (se houve oração) que define os dois grupos comparados. Cada coluna é uma categoria da variável resposta (se complicações ocorreram). As **células** da tabela contêm frequências para as quatro combinações possíveis dos resultados.

Os parâmetros π_1 e π_2 estimados usando a tabela de contingência são chamados de **probabilidades condicionais**. Esse termo se refere às probabilidades para uma variável resposta avaliada sob duas condições, a saber, os dois níveis da variável explicativa. Por exemplo, sob a condição de que estão sendo feitas orações para o sujeito, a probabilidade condicional de desenvolver complicações é estimada em 315/604 = 0,52.

Esta seção estudou as variáveis respostas binárias. Ao contrário, a variável resposta poderia ter várias categorias. Por exemplo, as categorias da resposta poderiam ser (sem complicações, complicações leves, complicações severas). Então, iríamos comparar os dois grupos em termos das probabilidades condicionais das observações em cada uma das três categorias. Da mesma forma, o número de grupos comparados poderia exceder a dois. O Capítulo 8 mostra como analisar tabelas de contingência que apresentam mais do que duas linhas ou colunas.

7.3 DADOS QUANTITATIVOS: COMPARANDO DUAS MÉDIAS

Para comparar duas médias da população μ_1 e μ_2, podemos fazer inferências sobre a sua diferença. Seria interessante, talvez, revisar as Seções 5.3 (página 140) e 6.2 (página 172) sobre inferência para a média no caso de uma amostra.

Intervalo de confiança para $\mu_2 - \mu_1$

Para amostras aleatórias grandes ou pequenas de distribuições com populações normais, a distribuição amostral de $(\bar{y}_2 - \bar{y}_1)$ tem uma forma normal. Como de costume, a inferência para médias com erros padrão *estimados* usa a distribuição t para estatísticas-teste e para a margem de erro em intervalos de confiança. Um intervalo de confiança pega a estimativa por ponto, soma e subtrai a margem de erro que é um escore-t vezes o erro padrão.

> ☑ **Intervalo de confiança para $\mu_2 - \mu_1$**
>
> Para amostras aleatórias independentes de dois grupos que têm distribuições populacionais normais, um intervalo de confiança para $\mu_2 - \mu_1$ é
>
> $$(\bar{y}_2 - \bar{y}_1) \pm t(ep),$$
>
> $$\text{onde } ep = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$
>
> O escore-t é escolhido de modo a fornecer o nível de confiança desejado.

A fórmula para os graus de liberdade para o escore-t, chamada de *aproximação de Welch-Satterthwaite*, é complexa. Os gl dependem dos desvios padrão das amostras s_1 e s_2 assim como dos tamanhos das amostras n_1 e n_2. Se $s_1 = s_2$ e $n_1 = n_2$, isto se simplifica a $gl = (n_1 + n_2 - 2)$. Esta é a soma dos valores dos gl para a inferência separada de cada grupo; isto é, $gl = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$. Geralmente, o gl está em algum lugar entre $n_1 + n_2 - 2$ e o mínimo de $(n_1 - 1)$ e $(n_2 - 1)$. Um *software* pode facilmente encontrar esse valor do gl, o escore-t e o intervalo de confiança.

Na prática, o método é robusto a violações da suposição de normalidade. Isto é especialmente verdadeiro quando ambos n_1 e n_2 são pelo menos aproximadamente 30, pelo Teorema Central do Limite.

Como de costume, você deve ser cauteloso com valores atípicos ou assimetria extrema que podem tornar a média uma medida inadequada para representar os dados.

EXEMPLO 7.3 Comparando o trabalho doméstico de homens e mulheres

Para a Tabela 7.1 (página 212), sobre o tempo diário que trabalhadores de tempo integral gastam cozinhando e limpando, representamos a média da população da Grã-Bretanha por μ_1 para homens e μ_2 para mulheres. Aquela tabela apresentou médias amostrais de 23 minutos para 1219 homens e de 37 minutos para 733 mulheres com desvios padrão amostrais de 32 e 16. A estimativa por ponto de $\mu_2 - \mu_1$ é igual a $\bar{y}_2 - \bar{y}_1 = 37 - 23 = 14$. A Seção 7.1 encontrou que o erro padrão estimado desta diferença se iguala a:

$$ep = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(32)^2}{1219} + \frac{(16)^2}{733}} = 1{,}09.$$

Os tamanhos da amostra são muito grandes, assim o escore-t para a margem de erro é essencialmente o escore-z. Portanto, o intervalo de 95% de confiança para $\mu_2 - \mu_1$ é:

$$(\bar{y}_2 - \bar{y}_1) \pm 1{,}96(ep) = 14 \pm 1{,}96(1{,}09)$$
$$\text{ou ainda } 14 \pm 2, \text{ que é } (12, 16).$$

Podemos estar 95% confiantes de que a quantia média do tempo diário da população gasto cozinhando e limpando está entre 12 e 16 minutos a mais para as mulheres do que para homens. ■

Interpretando um intervalo de confiança na comparação de médias

O intervalo de confiança (12, 16) contém somente valores positivos. Visto que calculamos a diferença entre a média para mulheres e a média para homens, podemos concluir que a média populacional é maior para as mulheres. Um intervalo de confiança para $\mu_2 - \mu_1$ que contém somente valores positivos sugere que $\mu_2 - \mu_1$ é positivo, significando que μ_2 é maior do que μ_1. Um intervalo de confiança para $\mu_2 - \mu_1$ que contém somente valores negativos sugere que μ_2 é menor do que μ_1. Quando o intervalo de confiança contém 0, não existe evidência o suficiente para concluir se μ_1 ou μ_2 é maior. É plausível, então, que $\mu_1 = \mu_2$.

A identificação de qual é o grupo 1 e qual é o grupo 2 é arbitrário, como quando estimamos $\mu_2 - \mu_1$ ou $\mu_1 - \mu_2$. Por exemplo, um intervalo de confiança de (12, 16) para $\mu_2 - \mu_1$ é equivalente a um de $(-16, -12)$ para $\mu_1 - \mu_2$.

Testes de significância sobre $\mu_2 - \mu_1$

Para comparar as médias populacionais μ_1 e μ_2, podemos, também, conduzir um teste de significância de H_0: $\mu_1 = \mu_2$. Para a diferença das médias dos parâmetros, esta hipótese é H_0: $\mu_2 - \mu_1 = 0$ (sem efeito).

Como sempre, a estatística-teste mensura o número de erros padrão entre a estimativa e o valor H_0,

$$t = \frac{\text{Estimativa do parâmetro} - \text{valor da hipótese nula}}{\text{Estimativa do erro padrão}}.$$

Tratando $\mu_2 - \mu_1$ como o parâmetro, testamos que $\mu_2 - \mu_1 = 0$. Sua estimativa é $\bar{y}_2 - \bar{y}_1$. O erro padrão é o mesmo do intervalo de confiança. A estatística-teste t é

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{ep},$$

$$\text{onde } ep = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

EXEMPLO 7.4 Teste comparando a média do trabalho doméstico para homens e mulheres

Usando os dados da Tabela 7.1 (página 212), testamos, agora, a diferença entre a o tempo médio de cozinhar e lavar da população, μ_1 para homens e μ_2 para mulheres. testamos H_0: $\mu_1 = \mu_2$ contra H_a: $\mu_1 \neq \mu_2$. Vimos que a estimativa $\bar{y}_2 - \bar{y}_1 = 37 - 23 = 14$ tem $ep = 1{,}09$.

A estatística-teste é igual a:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{ep} = \frac{(37 - 23)}{1{,}09} = 12{,}8.$$

Com amostras grandes, visto que a distribuição t é essencialmente a mesma que a normal padrão, $t = 12{,}8$ é enorme. Ela fornece um valor-p que é zero com muitas casas decimais. Concluímos que as médias das populações diferem. As médias amostrais mostram que a diferença toma a direção de uma média maior para as mulheres. ■

Na prática, os testes de significância são muito mais comuns para comparações de duas amostras do que para análises de uma amostra. Geralmente, é artificial testar se a média da população é igual a um determinado valor em particular, como, por exemplo, testar H_0: $\mu = \mu_0$. Entretanto, é normalmente relevante testar se existe uma *diferença* entre duas médias populacionais, como no teste de H_0: $\mu_1 = \mu_2$. Por exemplo, podemos não ter ideia do que supor para o tempo médio de trabalho doméstico para homens, mas podemos querer saber se essa média (qualquer que seja seu valor) é a mesma, maior ou menor do que o tempo médio para as mulheres.

Correspondência entre intervalos de confiança e testes

Para médias, a equivalência entre testes bilaterais e intervalos de confiança mencionada nas Seções 6.2 (página 172) e 6.4 e (página 185) também se aplica ao caso de duas amostras. Por exemplo, visto que o valor-p bilateral do Exemplo 7.4 é menor do que 0,05, rejeitamos H_0: $\mu_2 - \mu_1 = 0$, ao nível $\alpha = 0{,}05$. De forma similar, um intervalo de 95% de confiança para $\mu_2 - \mu_1$ não contém 0, o valor H_0. Aquele intervalo é igual a (12, 16).

Como na inferência de uma amostra, os intervalos de confiança são mais informativos do que os testes. O intervalo de confiança nos diz não somente que a média da população difere para homens e mulheres, mas também nos mostra quão grande é provável que essa diferença seja e qual a sua direção.

7.4 COMPARANDO MÉDIAS COM AMOSTRAS DEPENDENTES

Amostras dependentes ocorrem quando cada observação na amostra 1 se equipara com a observação na amostra 2. Os dados são, geralmente, chamados de dados de **pares emparelhados** em virtude dessa equiparação.

Diferenças dos escores pareadas para amostras emparelhadas

A dependência ocorre geralmente quando cada amostra tem os mesmos sujeitos. Exemplos são estudos observacionais *longitudinais* que observam a resposta de uma pessoa em vários pontos no tempo e estudos experimentais que tomam *medidas repetidas* dos sujeitos. Um exemplo do último é um estudo *transversal*, no qual um sujeito recebe um tratamento por um período e, depois, outro tratamento. O próximo exemplo ilustra essa situação.

EXEMPLO 7.5 Uso do telefone celular e tempo de reação do motorista

Um experimento recente[3] usou uma amostra de estudantes universitários para investigar se o uso do telefone celular prejudica a reação dos motoristas. Em uma máquina que simulou situações de direção, em períodos irregulares um alvo brilhava na cor vermelha ou verde. Os participantes foram instruídos para pressionar o freio o mais rápido possível assim que detectassem uma luz verde. Sob a condição do telefone celular, o estudante mantinha uma conversa sobre política no telefone celular com alguém em uma sala separada. Na condição de controle, eles escutavam uma transmissão no rádio ou livros em fita enquanto simulavam estar dirigindo.

Para cada estudante, em uma condição particular, o resultado registrado na Tabela 7.3 é o seu tempo médio de resposta (em milésimos de segundos) em várias tentativas. A Figura 7.3 mostra os diagramas de caixa e bigodes para as duas situações. O estudante 28 é um valor atípico sob cada condição. ■

Para os dados emparelhados, cada observação de uma amostra é pareada com uma observação na outra amostra. Para cada par, formamos:

Diferença = Observação na amostra 2 − Observação na amostra 1.

A Tabela 7.3 mostra a diferença dos escores para o experimento do telefone celular.

Considere $\bar{y}_d$ a representação da média amostral das diferenças dos escores. Isto estima μ_d, a diferença média na população. Na verdade, o parâmetro μ_d é idêntico a $\mu_2 - \mu_1$, diferença entre as médias populacionais para os dois grupos. A média das diferenças é igual à diferença entre as médias.

> ☑ Para dados emparelhados, a diferença entre as médias dos dois grupos é igual à média da diferença dos escores.

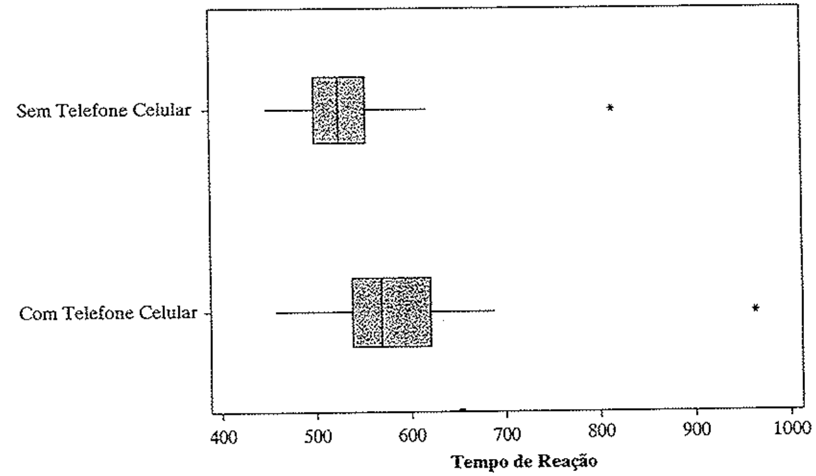
Inferências comparando médias a partir de diferenças pareadas

Podemos basear as análises sobre $\mu_2 - \mu_1$ nas inferências sobre μ_d, usando uma amostra da diferença dos escores. Isso simplifica a análise porque reduz um problema de duas amostras para um problema de uma amostra.

Considere n a representação do número de observações em cada amostra. Isso se iguala ao número da diferença dos escores. O intervalo de confiança para μ_d é:

☑ **Tabela 7.3** Tempos de reação (em milésimos de segundos) na tarefa de dirigir sobre a condição de uso ou não do telefone celular. A diferença do escore é o tempo de reação para frear usando e não usando o telefone celular, tal como 636 − 604 = 32 milésimos de segundo

	Telefone Celular?				Telefone Celular?		
Estudante	Não	Sim	Diferença	Estudante	Não	Sim	Diferença
1	604	636	32	17	525	626	101
2	556	623	67	18	508	501	−7
3	540	615	75	19	529	574	45
4	522	672	150	20	470	468	−2
5	459	601	142	21	512	578	66
6	544	600	56	22	487	560	73
7	513	542	29	23	515	525	10
8	470	554	84	24	499	647	148
9	556	543	−13	25	448	456	8
10	531	520	−11	26	558	688	130
11	599	609	10	27	589	679	90
12	537	559	22	28	814	960	146
13	619	595	−24	29	519	558	39
14	536	565	29	30	462	482	20
15	554	573	19	31	521	527	6
16	467	554	87	32	543	536	−7



☑ **Figura 7.3** Diagramas de caixa e bigodes do experimento do uso ou não do telefone celular nos tempos de reação para frear.

$$\bar{y}_d \pm t\left(\frac{s_d}{\sqrt{n}}\right).$$

Aqui, $\bar{y}_d$ e s_d são média amostral e desvio padrão da diferença dos escores, e t é o escore-t para o nível de confiança escolhido, tendo $gl = n - 1$. Este intervalo de confiança tem a mesma forma do que o da Seção 6.3 (página 182) para uma única amostra. Aplicamos a fórmula para uma única amostra de n diferenças em vez de sobre dois conjuntos originais de observações.

Para testar H_0: $\mu_1 = \mu_2$, expressamos a hipótese em termos de diferenças dos escores como H_0: $\mu_d = 0$. A estatística-teste é:

$$t = \frac{\bar{y}_d - 0}{ep}, \quad \text{onde } ep = s_d/\sqrt{n},$$

que compara a média amostral das diferenças ao valor da hipótese nula de 0, em termos do número de erros padrão entre eles. O erro padrão é o mesmo usado para um intervalo de confiança. Visto que esse teste usa a diferença dos escores para pares de observações, ele é chamado de **teste *t* para diferenças pareadas**.

EXEMPLO 7.6 Uso do telefone celular e tempo de reação do motorista (continuação)

Agora, analisamos os dados de pares equiparados da Tabela 7.3 para o experimento de dirigir com e sem o uso do telefone celular. Os tempos médios de reação foram 534,6 milésimos de segundo sem o uso do celular e 585,2 milésimos de segundo usando o celular. As 32 diferenças dos escores (32, 67, 75,...) da Tabela 7.3 têm uma média amostral de:

$$\bar{y}_d = (32 + 67 + 75 + \dots + (-7))/32 = 50{,}6.$$

Isto se iguala à diferença entre as médias amostrais de 585,2 e 534,6 para as duas condições. O desvio padrão amostral das 32 diferenças dos escores é:

$$s_d = \sqrt{\frac{(32 - 50{,}6)^2 + (67 - 50{,}6)^2 + \cdots}{32 - 1}} = 52{,}5.$$

O erro padrão de $\bar{y}_d$ é $ep = s_d/\sqrt{n} = 52{,}5/\sqrt{32} = 9{,}28$.

Para um intervalo de 95% de confiança para $\mu_d = \mu_2 - \mu_1$ com $gl = n - 1 = 31$, usamos $t_{0,025} = 2{,}04$. O intervalo de confiança é igual a

$$\bar{y}_d \pm 2{,}04(ep) = 50{,}6 \pm 2{,}04(9{,}28), \text{ que é } (31{,}7; 69{,}5).$$

Inferimos que o tempo médio da reação da população usando o celular está entre 32 e 70 milésimos de segundo acima do que aquele sem o uso celular. O intervalo de confiança não contém 0. Concluímos, assim, que o tempo médio de reação da população é maior quando envolve o uso do celular.

A seguir, considere o teste de significância H_0: $\mu_d = 0$ (e, por conseguinte, médias iguais da população para as duas condições) contra H_a: $\mu_d \neq 0$. A estatística-teste é

$$t = \frac{(\bar{y}_d - 0)}{ep} = \frac{50{,}6}{9{,}28} = 5{,}5,$$

com $gl = 31$. O valor-p bilateral é igual a 0,000005. Existe uma forte evidência de que o tempo médio de reação é maior quando o celular é usado. A Tabela 7.4 mostra como o SPSS apresenta esses resultados para a opção do teste t para amostras emparelhadas. ■

As inferências para diferenças pareadas fazem as suposições usuais para os procedimentos t: as observações (as diferenças dos escores) são obtidas aleatoriamente da população que é normal. Os intervalos de confiança e testes bilaterais funcionam bem, mesmo se a suposição de normalidade é violada (propriedade da *robustez*), a não ser que o tamanho da amostra seja pequeno e a distribuição seja altamente assimétrica ou tenha valores atípicos severos. Para o estudo sobre dirigir com o uso ou não de celulares, um sujeito era um valor atípico nos dois tempos de reação. Contudo, a diferença dos escores para esse sujeito, que é o valor usado na análise, não é um valor atípico. O artigo sobre o estudo não indicou se os sujeitos foram selecionados aleatoriamente. Os sujeitos no experimento eram provavelmente de uma amostra voluntária, portanto as conclusões inferenciais são aproximadas.

Tabela 7.4 Saída do SPSS para a análise de valores pareados comparando os tempos de reação do motorista com e sem o uso do telefone celular

Testes t para Amostras Pareadas

Variável	Nº de pares	Média	DP	EP da média
SEM CELULAR		534,56	66,45	11,75
	32			
COM CELULAR		585,19	89,65	15,85

Diferenças pareadas

Média	DP	EP da média	valor t	gl	Sig. bilateral
50,63	52,49	9,28	5,46	31	0,000

IC de 95% (31,70; 69,55)

Amostras independentes *versus* amostras dependentes

Usar amostras dependentes pode ter certos benefícios. Primeiro, fontes conhecidas de tendenciosidade potencial são controladas. Usar os mesmos sujeitos em cada amostra, por exemplo, mantém outros fatores fixos que poderiam afetar a análise. Suponha que sujeitos mais jovens tendam a ter os tempos de reação mais rápidos. Se o grupo 1 tem uma média amostral mais baixa do que o grupo 2, não é porque os sujeitos do grupo 1 sejam mais jovens, porque ambos os grupos têm os mesmos sujeitos.

Segundo, o erro padrão de $\bar{y}_2 - \bar{y}_1$ pode ser menor com amostras dependentes. No estudo do telefone celular, o erro padrão era de 9,3. Se tivéssemos observado amostras *independentes* com os mesmos escores que os da Tabela 7.3, o erro padrão de $\bar{y}_2 - \bar{y}_1$ seria de 19,7. Isso ocorre porque a variabilidade da diferença dos escores tende a ser menor do que a variabilidade nos escores originais quando os escores nas duas amostras estão fortemente correlacionados. Na verdade, para os dados na Tabela 7.3, a correlação (lembre-se da Seção 3.5, página 73) entre os tempos de reação com e sem celular é de 0,81, mostrando uma associação positiva forte.

7.5 OUTROS MÉTODOS PARA COMPARAR MÉDIAS*

A Seção 7.3 apresentou inferência para comparar duas médias com amostras independentes. Um método de inferência levemente diferente pode ser usado quando esperamos variabilidade similar para os dois grupos. Por exemplo, sob a hipótese nula de "sem efeito", geralmente esperamos que todas as distribuições da variável resposta sejam idênticas para os dois grupos. Assim, esperamos que tanto os desvios padrão quanto as médias sejam idênticos.

Comparando médias com a suposição de desvios padrão iguais

Na comparação das médias populacionais, esse método faz a suposição adicional de que os desvios padrão populacionais são iguais, isto é, $\sigma_1 = \sigma_2$. Nesse caso, existe uma expressão mais simples para o *gl* do valor *exato* da distribuição *t* para essa estatística. Embora pareça desagradável fazer uma suposição adicional, os intervalos de confiança e testes bilaterais são bem robustos contra violações dessa hipótese e a da normalidade, principalmente quando os tamanhos das amostras são similares e não muito pequenos.

O valor comum σ de σ_1 e σ_2 é estimado por:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{\Sigma(y_1 - \bar{y}_1)^2 + \Sigma(y_2 - \bar{y}_2)^2}{n_1 + n_2 - 2}}.$$

Aqui, $\Sigma(y_1 - \bar{y}_1)^2$ representa a soma dos quadrados em torno da média para os valores do primeiro exemplo e $\Sigma(y_2 - \bar{y}_2)^2$ representa a soma dos quadrados em torno da média para as observações do segundo exemplo. A estimativa *s* junta as informações dos dois exemplos para fornecer uma única estimativa da variabilidade. Ela é chamada de *estimativa combinada*. O termo dentro da raiz quadrada é uma média ponderada das variâncias das duas amostras. Quando $n_1 = n_2$, é a média aritmética simples. A estimativa *s* está entre s_1 e s_2. Com *s* como a estimativa de σ_1 e σ_2, o erro padrão estimado de $\bar{y}_2 - \bar{y}_1$ fica reduzido a:

$$ep = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

O intervalo de confiança para $\mu_2 - \mu_1$ tem a forma usual:

$$(\bar{y}_2 - \bar{y}_1) \pm t(ep).$$

O escore-*t* vem da tabela *t* para o nível de confiança desejado, com $gl = n_1 + n_2 - 2$. O *gl* é igual ao número total de observações $(n_1 + n_2)$ menos o número dos parâmetros estimados para calcular *s* (a saber, as duas médias, μ_1 e μ_2, estimadas por $\bar{y}_1$ e $\bar{y}_2$).

Para testar H_0: $\mu_1 = \mu_2$, a estatística-teste tem a forma usual:

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{ep}.$$

Agora, o *ep* usa a fórmula combinada, como no intervalo de confiança. A estatística-teste tem uma distribuição *t* com $gl = n_1 + n_2 - 2$.

EXEMPLO 7.7 Comparando uma terapia ao grupo de controle

Os Exemplos 5.5 (página 144) e 6.4 (página 177) descreveram um estudo que usou uma terapia cognitivo-comportamental para tratar de uma amostra de meninas adolescentes que sofriam de anorexia. O estudo observou a mudança média no peso após um período do tratamento. Estudos desse tipo geralmente também têm um grupo de controle que recebe o tratamento ou um tratamento padrão. Então, os pesquisadores podem analisar como a mudança no peso se comporta em relação ao grupo do tratamento e o grupo de controle.

Na verdade, o grupo da anorexia tinha um grupo de controle. As meninas adolescentes do estudo foram designadas aleatoriamente ao tratamento cognitivo-comportamental (Grupo 1) ou ao grupo de controle (Grupo 2). A Tabela 7.5 resume os resultados. (Os dados de ambos os grupos são exibidos na Tabela 12.21 na página 439.)

Se H_0 é verdadeira, que o tratamento tem o mesmo efeito do que o grupo de controle, então esperaríamos que os grupos apresentassem médias iguais e desvios padrão iguais. Para esses dados, a estimativa combinada do desvio padrão comum assumido é igual a:

Tabela 7.5 Resumo dos resultados comparando o grupo de tratamento com o de controle para o estudo da anorexia

Grupo	Tamanho da amostra	Média	Desvio padrão
Tratamento	29	3,01	7,31
Controle	26	−0,45	7,99

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{28(7{,}31)^2 + 25(7{,}99)^2}{29 + 26 - 2}} = \sqrt{\frac{3092{,}2}{53}} = 7{,}64.$$

Agora, $\bar{y}_1 - \bar{y}_2 = 3{,}01 - (-0{,}45) = 3{,}46$ tem um erro padrão estimado de:

$$ep = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 7{,}64\sqrt{\frac{1}{29} + \frac{1}{26}} = 2{,}06.$$

Considere μ_1 e $\mu 2$ representando os ganhos médios de peso para essas terapias para as populações hipotéticas de onde as amostras foram retiradas. Testamos H_0: $\mu_1 = \mu_2$ contra H_a: $\mu_1 \neq \mu_2$.

$$t = \frac{\bar{y}_1 - \bar{y}_2}{ep} = \frac{3{,}01 - (-0{,}45)}{2{,}06} = 1{,}68.$$

A estatística tem $gl = n_1 + n_2 - 2 = 29 + 26 - 2 = 53$. Da tabela *t* (Tabela B) o valor-*p* bilateral é 0,10. Existe somente uma fraca evidência de que o uso da terapia cognitivo-comportamental tenha sucesso.

Quando $gl = 53$, o escore-*t* para um intervalo de 95% de confiança para $(\mu_1 - \mu_2)$ é $t_{0,025} = 2{,}006$. O intervalo é:

$$(\bar{y}_1 - \bar{y}_2) \pm t(ep) = 3{,}46 \pm 2{,}006(2{,}06),$$
$$\text{que é } 3{,}46 \pm 4{,}14 \text{ ou } (-0{,}7; 7{,}6).$$

Concluímos que a mudança média do peso para a terapia cognitivo-comportamental poderá ser tanto quanto 0,7 libras mais baixas ou tanto quanto 7,6 libras mais altas do que a mudança média do peso para o grupo de controle. Visto que o intervalo contém 0, é plausível que as médias populacionais sejam idênticas. Isso é consistente com o valor-*p* excedendo a 0,05 no teste. Se a mudança média no peso da população é menor do que a do grupo de terapia cognitivo-comportamental, ela é um pouco menor (menos do que 1 libra), mas se a mudança média da população é maior, ela poderá ser de aproximadamente 8 libras. Como os tamanhos da amostra não são grandes, o intervalo de confiança é relativamente grande. ■

Delineamentos aleatorizados por blocos *versus* completamente aleatorizados

O estudo da anorexia usou um delineamento experimental *completamente aleatorizado*: os sujeitos foram aleatoriamente designados às duas terapias. Com esse delineamento, existe uma chance de que os sujeitos selecionados para uma terapia possam diferir de uma forma importante dos sujeitos selecionados para a outra terapia. Para amostras moderadas a grandes, os fatores que poderiam influenciar os resultados (como o peso inicial) tendem a se equilibrar em virtude da aleatorização. Para amostras pequenas, um desequilíbrio pode ocorrer.

Um delineamento experimental alternativo *equipara* os sujeitos nas duas amostras, como, por exemplo, pegar duas meninas com o mesmo peso e aleatoriamente decidir qual menina receberá determinada terapia. Esse plano com pares equiparados é um exemplo simples de um **delineamento aleatorizado por blocos**. Cada par de sujeitos forma um *bloco* e dentro dos blocos os sujeitos são aleatoriamente designados aos tratamentos. Com esse delineamento, usamos os métodos da seção anterior para variáveis dependentes.

Inferências relatadas pelo *software*

A Tabela 7.6 ilustra a forma como o SPSS apresenta os resultados dos testes t para duas amostras. A tabela mostra os resultados dos dois testes para comparar médias, que diferem na forma de assumirem os desvios padrão iguais da população. O teste t recém apresentado assume que $\sigma_1 = \sigma_2$. A estatística t que o *software* apresenta para o caso de "variâncias supostamente diferentes" é a estatística t da Seção 7.3,

$$t = (\bar{y}_2 - \bar{y}_1)/ep, \text{ com } ep = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Quando $n_1 = n_2$, as estatísticas-teste supondo "variâncias iguais" e "variâncias diferentes" coincidem. Eles são geralmente semelhantes se n_1 e n_2 estão próximos ou se s_1 e s_2 estão próximos.

Se os dados mostram evidência de uma diferença potencialmente grande nos desvios padrão (com, digamos, um desvio padrão sendo, pelo menos, o dobro do outro) é melhor usar o teste t aproximado (Seção 7.3) que não necessita da suposição de $\sigma_1 = \sigma_2$. Ele pode gerar um valor estatístico t muito diferente do método que supõe $\sigma_1 = \sigma_2$ se s_1 e s_2 são bem diferentes e os tamanhos das amostras desiguais.

Muitos livros e *softwares* apresentam uma estatística representada por F para testar que os desvios padrão populacionais são iguais. Não é apropriado conduzir esse teste para determinar qual teste t usar. Na verdade, não recomendamos esse teste mesmo se o seu propósito principal é comparar a variabilidade de dois grupos. Esse teste supõe que as distribuições populacionais são normais e ele não é robusto a violações dessa suposição.

Tamanho do efeito

No Exemplo 7.7, do estudo da anorexia, a diferença estimada entre os ganhos médios de peso de 3,46 é grande ou pequena em termos práticos? Lembre que o tamanho de uma diferença estimada depende das unidades da mensuração. Esses dados estavam em libras, mas, se convertermos para quilogramas, a diferença estimada seria de 1,57 e, se convertida a onças, seria de 55,4.

Uma forma padronizada para descrever a diferença é dividi-la pelo desvio padrão estimado para cada grupo. Isto é chamado de **tamanho do efeito**. Com médias amostrais de 3,01 e −0,45 libras e uma estimativa combinada do desvio padrão de $s = 7{,}64$ libras, a diferença padronizada é:

$$\text{Tamanho do efeito} = \frac{\bar{y}_1 - \bar{y}_2}{s} = \frac{3{,}01 - (-0{,}45)}{7{,}64} = 0{,}45.$$

Tabela 7.6 Saída do SPSS para os testes t de duas amostras

		Teste t para a igualdade de médias				
		t	gl	Sig. (bilateral)	Diferença média	Erro padrão da diferença
MUDANÇA NO PESO	Supondo variâncias iguais	1,68	53	0,099	3,46	2,06
	Supondo variâncias diferentes	1,67	50	0,102	3,46	2,07

A diferença entre as médias amostrais é menor do que a metade de um desvio padrão, que é um valor relativamente pequeno. Obteríamos o mesmo valor para o tamanho do efeito se mensurássemos esses dados em unidades diferentes, como quilogramas ou onças.

Um modelo para médias

Na segunda parte deste livro, aprenderemos sobre métodos avançados para analisar associações entre variáveis. Basearemos as análises explicitamente em um *modelo*. Para duas variáveis, um **modelo** é uma aproximação simples para o relacionamento verdadeiro entre essas variáveis na população.

Considere $N(\mu, \sigma)$ a representação de uma distribuição normal com média μ e desvio padrão σ. Considere y_1 a representação de uma observação aleatoriamente selecionada do Grupo 1 e y_2 uma observação aleatoriamente selecionada do Grupo 2. A hipótese testada acima para comparar médias supondo que $\sigma_1 = \sigma_2$ pode ser expressa como o modelo:

H_0: ambos y_1 e y_2 têm uma distribuição $N(\mu, \sigma)$.

H_a: y_1 tem uma distribuição $N(\mu_1, \sigma)$ e y_2 tem uma distribuição $N(\mu_2, \sigma)$, com $\mu_1 \neq \mu_2$.

Sob H_0, as médias populacionais são iguais, com um valor comum μ. Sob H_a, as médias populacionais diferem. Este é um caso especial de um modelo que o Capítulo 12 usa para comparar *várias* médias.

As distribuições amostrais e as inferências resultantes são derivadas sob a estrutura suposta do modelo. Mas os modelos são meramente simplificações convenientes da realidade. Não esperamos que as distribuições sejam exatamente normais, por exemplo. Uma das partes-chave de se ficar mais à vontade usando os métodos estatísticos é estar mais informado sobre quais suposições são mais importantes em um modelo e como verificar tais suposições. Geralmente, existem benefícios no uso de modelos mais simples. Eles têm menos parâmetros para estimar e as inferências podem ser mais poderosas. Entretanto, quando tal modelo está muito errado, é melhor usar um modelo mais complexo.

O primeiro teste de significância que discutimos para comparar médias usou um modelo um pouco mais complexo:

H_0: y_1 tem uma distribuição $N(\mu, \sigma_1)$, y_2 tem uma distribuição $N(\mu, \sigma_2)$.

H_a: y_1 tem uma distribuição $N(\mu_1, \sigma_1)$, y_2 tem uma distribuição $N(\mu_2, \sigma_2)$, com $\mu_1 \neq \mu_2$.

Novamente, sob H_0 as médias da população são iguais. Mas, agora, nenhuma suposição é feita sobre os desvios padrão serem iguais. Se existe motivo para esperar que os desvios padrão sejam muito diferentes ou se os dados indicam isto (com um dos desvios padrão da amostra sendo pelo menos o dobro do outro), então é melhor usarmos análises baseadas neste modelo. Se os dados mostram que mesmo este modelo pode estar muito errado, como, por exemplo, quando as distribuições dos dados amostrais são tão assimétricas que a média é um representante inadequado, é melhor, então, usar um modelo diferente. A seção final deste capítulo apresenta um modelo que não assume a normalidade.

7.6 OUTROS MÉTODOS PARA COMPARAR PROPORÇÕES*

A Seção 7.2 apresentou métodos para amostras grandes para comparar proporções com amostras independentes. Esta seção apresenta métodos para comparar proporções com (1) amostra dependente e (2) amostras pequenas.

Comparando proporções dependentes

A Seção 7.4 apresentou métodos para comparar médias com amostras dependentes. O exemplo seguinte ilustra métodos para comparar proporções com amostras dependentes.

EXEMPLO 7.8 Comparando dois sistemas de reconhecimento da voz

Nos últimos anos tem havido aprimoramentos impressionantes nos sistemas para reconhecimento automático da voz. Quando você, nos dias de hoje, liga para vários centros de serviços, antes de falar com um ser humano é solicitado a você que responda a várias perguntas verbalmente, enquanto no passado você tinha que usar o disco do telefone.

As pesquisas comparando a qualidade dos diferentes sistemas de reconhecimento da voz geralmente usam, como um teste de avaliação, uma série de palavras isoladas, verificando quão seguido cada sistema comete erros no reconhecimento da palavra. A Tabela 7.7 mostra um exemplo[4] de um desses testes, comparando dois sistemas de reconhecimento da voz, chamados de segmentação generalizada de mínima distorção (SGMD) e densidade contínua do modelo oculto de Markov (DCMOM).

As linhas da Tabela 7.7 são as categorias (correto, incorreto) para cada palavra usando o SGMD. As colunas são as mesmas categorias para o DCMOM. As frequências marginais das linhas (1979, 21), os totais de (correto, incorreto) para o SGMD. As frequências marginais das colunas (1937, 63) são os totais para a DCMOM.

Tabela 7.7 Resultados do teste de avaliação utilizando 2000 palavras para dois sistemas de reconhecimento da voz

	DCMOM		
SGMD	**Correto**	**Incorreto**	**Total**
Correto	1921	58	1979
Incorreto	16	5	21
Total	1937	63	2000

Iremos comparar a proporção das respostas corretas para esses dois sistemas de reconhecimento da voz. As amostras são dependentes porque os dois sistemas usaram as mesmas 2000 palavras. Consideraremos estas 2000 palavras como uma amostra aleatória de palavras possíveis nas quais o sistema poderia ter sido testado. Considere π_1 a representação da proporção populacional de palavras corretamente identificadas com o SGMD e considere π_2 a representação da proporção populacional de palavras corretamente identificadas com o DCMOM. As estimativas da amostra são $\hat{\pi}_1 = 1979/2000 = 0{,}9895$ e $\hat{\pi}_2 = 1937/2000 = 0{,}9685$.

Se as proporções de palavras corretamente identificadas foram idênticas para os dois sistemas, o número de observações da primeira linha da Tabela 7.7 seria igual ao número de observações da primeira coluna. A primeira célula (aquela contendo 1921 na Tabela 7.7) é comum tanto na primeira linha quanto na primeira coluna, assim a outra frequência da célula na primeira linha seria igual à outra frequência na primeira coluna. Isto é, o número de palavras julgadas corretas pelo SGMD, mas incorretas pela DCMOM. Podemos testar H_0: $\pi_1 = \pi_2$ usando as frequências dessas duas células. Se H_0 é verdadeira, então, destas palavras, nós esperamos que ½ sejam corretas para o SGMD e incorretas para DCMOM e ½ sejam incorretas para SGMD e corretas para o DCMOM.

Como no teste emparelhado para uma média, reduzimos a inferência para um único parâmetro. Para a população nas duas células recém mencionadas, testamos se metade está em cada célula. Na Tabela 7.7, dos 58 + 16 = 74 palavras julgadas corretamente por um sistema, mas incorretas pelo outro, a proporção amostral 58/74 = 0,784 está correta com SGMD. Sob a hipótese nula de que a proporção da população é de 0,50, o erro padrão da proporção amostral para essas 74 observações é $\sqrt{(0{,}50)(0{,}50)/74} = 0{,}058$.

Da Seção 6.3 (página 182), a estatística z para testar que a proporção populacional é igual a 0,50 é:

$$z = \frac{\text{proporção amostral} - \text{proporção } H_0}{\text{erro padrão}} = \frac{0{,}784 - 0{,}50}{0{,}058} = 4{,}88.$$

O valor-p bilateral é igual a 0,000. Isto fornece uma forte evidência contra H_0: $\pi_1 = \pi_2$. Baseado nas proporções amostrais, a evidência favorece a proporção populacional maior de reconhecimentos corretos pelo sistema SGMD. ■

Teste McNemar para comparar proporções dependentes

Existe uma fórmula simples para essa estatística-teste z para comparar duas proporções dependentes. Para uma tabela da forma da Tabela 7.7, represente as frequências das células nas duas células relevantes por n_{12} para aquelas na linha 1 e na coluna 2 e por n_{21} para aquelas na linha 2 e na coluna 1. A estatística-teste é igual a:

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

Quando $n_{12} + n_{21}$ excede 20, essa estatística tem aproximadamente uma distribuição normal padrão se H_0 é verdadeira. Este teste é geralmente chamado de **teste McNemar**. Para amostras pequenas, use a distribuição binomial para realizar o teste.

Para a Tabela 7.7, teste McNemar usa $n_{12} = 58$, o número de palavras reconhecidas corretamente pelo SGMD e incorretamente pelo DCMOM, e $n_{21} = 16$, o número para o inverso. A estatística-teste é igual a:

$$z = \frac{58 - 16}{\sqrt{58 + 16}} = 4{,}88.$$

O valor-p é 0,000.

Intervalo de confiança para a diferença das proporções dependentes

Um intervalo de confiança para a diferença das proporções é mais informativo do que um teste de significância. Para amostras grandes, ele é:

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(ep),$$

onde o erro padrão é estimado usando:

$$ep = \sqrt{(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n}/n.$$

Para a Tabela 7.7, $\hat{\pi}_1 = 1979/2000 = 0{,}9895$ e $\hat{\pi}_2 = 1937/2000 = 0{,}9685$. A diferença entre $\hat{\pi}_1 - \hat{\pi}_2 = 0{,}9895 - 0{,}9685 = 0{,}021$. Para $n = 2000$ observações com $n_{12} = 58$ e $n_{21} = 16$,

$$ep = \sqrt{(58 + 16) - (58 - 16)^2/2000}/2000 = 0{,}0043.$$

Um intervalo de 95% de confiança para $\pi_1 - \pi_2$ é igual a $0{,}021 \pm 1{,}96(0{,}0043)$ ou (0,013; 0,029). Concluímos que a proporção da população correta com o sistema SGMD está entre aproximadamente 0,01 e 0,03 maior do que a proporção da população correta com o sistema DCMOM. Em resumo, a diferença entre as proporções da população parece ser muito pequena.

O teste exato de Fisher para comparar proporções

As inferências para proporções com amostras independentes introduzidas na Seção 7.2 são válidas para amostras relativamente grandes. A seguir, estudaremos métodos para amostras pequenas.

O teste de significância bilateral para comparar proporções com a estatística z funciona muito bem se cada amostra tem pelo menos aproximadamente 5 a 10 resultados de cada tipo (isto é, pelo menos 5 a 10 observações em cada célula da tabela de contingência). Para tamanhos amostrais menores, a distribuição amostral de $\hat{\pi}_2 - \hat{\pi}_1$ pode não estar próximo da normalidade. Você pode, então, comparar duas proporções π_1 e π_2 usando um método chamado de **teste exato de Fisher**, atribuído ao eminente estatístico Ronald A. Fisher.

Os cálculos para o teste exato de Fisher são complexos e além do alcance deste livro. O princípio por trás do teste é claro, entretanto, como o Exercício 7.57 mostra. Um *software* estatístico fornece o seu valor-*p*. Como de costume, o valor-*p* é a probabilidade do resultado da amostra ou um resultado ainda mais extremo, sob a suposição de que H_0 é verdadeira. Para detalhes sobre o teste exato de Fisher, veja Agresti (2007, p. 45-48).

EXEMPLO 7.9 Depressão e suicídio entre pessoas infectadas com HIV

Um estudo recente[5] examinou as taxas de depressão e suicídio para infectados pelo HIV e pessoas não infectadas na China. O estudo usou uma amostra voluntária. Em uma tentativa de tornar a amostra mais representativa, os sujeitos foram recrutados de clínicas em duas regiões bem diferentes da China, uma urbana e outra rural. A Tabela 7.8 mostra os resultados baseados em uma entrevista-diagnóstico que perguntava se o sujeito já havia tentado suicídio. A tabela também mostra a saída da condução do teste exato de Fisher.

Represente a proporção da população que já tentou suicídio por π_1 para aqueles que eram HIV positivo e π_2 para aqueles que eram HIV negativo. Então, $\hat{\pi}_1 = 10/28 = 0{,}36$ e $\hat{\pi}_2 = 1/23 = 0{,}04$. Testamos H_0: $\pi_1 = \pi_2$ contra H_a: $\pi_1 > \pi_2$. Uma das quatro frequências é muito pequena, assim para estarmos seguros usamos o teste exato de Fisher.

Na saída do computador, a alternativa do lado direito se refere a H_a: $\pi_1 - \pi_2 > 0$; isto é, H_a: $\pi_1 > \pi_2$. O valor-*p* = 0,0068 dá uma forte evidência de que a proporção da população que tentou o suicídio é maior para aqueles que são HIV positivo. O valor-*p* para a alternativa bilateral é igual a 0,0075. Isto não é o dobro do valor-*p* unilateral porque, exceto em certos casos especiais, a distribuição amostral (chamada de **distribuição hipergeométrica**) não é simétrica. ■

Tabela 7.8 Comparação dos sujeitos contaminados e não contaminados pelo HIV quanto à tentativa ou não de suicídio

HIV	Tentativa de suicídio: sim	não	Total
positivo	10	18	28
negativo	1	22	23
Total	11	40	51

ESTATÍSTICAS PARA A TABELA DO HIV E TENTATIVA DE SUICÍDIO

Estatística		Prob
Teste Exato de Fisher	(Esquerda	0,9995
	(Direita)	0,0068
	(Bilateral)	0,0075

Estimação com pequenas amostras comparando duas proporções

Da Seção 7.2, o intervalo de confiança para comparar proporções com amostras grandes é:

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(ep), \text{ onde}$$

$$ep = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

Um simples ajuste desta fórmula para que ela funcione melhor, mesmo com pequenas amostras, adiciona uma observação de cada tipo para cada amostra. Para os dados na Tabela 7.8 do Exemplo 7.9, substituímos as frequências da célula (10, 18, 1, 22) por (11, 19, 2, 23). Então, as estimativas ajustadas são $\hat{\pi}_1 = (10 + 1)/(28 + 2) = 0{,}367$ e $\hat{\pi}_2 = (1 + 1)/(23 + 2) = 0{,}080$. O erro padrão ajustado (usando $n_1 = 30$ e $n_2 = 25$) é igual a 0,108 e um intervalo de 95% de confiança é:

$$(0{,}367 - 0{,}080) \pm 1{,}96(0{,}103) \text{ ou}$$
$$0{,}287 \pm 0{,}203, \text{ que é } (0{,}08; 0{,}49)$$

Observa-se, sem surpresas, que, quando as amostras são pequenas, o intervalo é bastante amplo.

7.7 ESTATÍSTICA NÃO PARAMÉTRICA PARA COMPARAR GRUPOS*

Vimos que com amostras grandes muitas estatísticas têm distribuições amostrais normais, mesmo quando as distribuições da população não são normais. Na verdade, com a amostragem aleatória, aproximadamente todas as estimativas dos parâmetros têm distribuição normal para grandes amostras. As amostras pequenas, no entanto, geralmente requerem suposições adicionais. Por exemplo, as inferências para as médias usando a distribuição *t* supõem distribuições populacionais normais.

Existe um conjunto de métodos que *não* faz suposições sobre a forma da distribuição da população. Esses métodos são chamados de **não paramétricos**. Eles contrastam com os métodos tradicionais (também chamados de *paramétricos*) que supõem populações normais. Os métodos não paramétricos são úteis, por exemplo, quando a suposição de normalidade para métodos que usam a distribuição *t* é violada. Eles são úteis principalmente para pequenas amostras, especialmente para testes unilaterais, em que os métodos paramétricos podem funcionar mal quando a suposição de normalidade for violada. Eles também são úteis quando dois grupos têm distribuições muito assimétricas, porque, nesse caso, a média pode não ser uma medida adequada para resumir os dados.

O teste de Wilcoxon-Mann-Whitney

Para ilustrar, considere o método que utiliza a distribuição *t* para comparar médias que supõem distribuições de normalidade com desvios padrão idênticos (Seção 7.5). Essas suposições são principalmente relevantes para pequenas amostras, digamos quando n_1 ou n_2 forem menor do que 20 a 30. A maioria das comparações não paramétricas de grupos também supõe formas idênticas para as distribuições da população, mas não é requerido que as formas sejam normais. O modelo para o teste é então:

H_0: ambos y_1 e y_2 têm a mesma distribuição.

H_a: as distribuições para y_1 e y_2 têm a mesma forma, mas a de y_1 está deslocada para cima, ou para baixo, quando comparada a de y_2.

O teste mais popular deste tipo é chamado de teste *Wilcoxon*. Esse teste é um método de nível ordinal, no sentido de que ele usa as classificações (postos) das observações. Os valores da amostra combinada de $n_1 + n_2$ são classificações de 1 até $n_1 + n_2$, e as médias dos postos (classificações) são calculadas para cada amostra. A estatística-teste compara as médias amostrais dos postos. Para amostras grandes, a estatística-teste z tem uma distribuição aproximadamente normal padrão. Para pequenas amostras, um valor-p exato é baseado em quão incomum é a diferença observada entre as classificações médias (sob a suposição de que H_0 é verdadeira) quando comparada a diferenças entre os postos para todas as demais classificações possíveis.

Outro teste não paramétrico é o teste de *Mann-Whitney*. Ele observa todos os pares de observações, tal que a primeira observação é de um grupo e a segunda é do outro grupo. A estatística-teste é baseada no número de pares para os quais a observação do primeiro grupo é maior. O teste é equivalente ao teste de Wilcoxon, dando o mesmo valor-p. (Frank Wilcoxon desenvolveu testes equivalentes aos de Henry Berthold Mann e Donald Ransom Whitney aproximadamente no mesmo período da década de 1940.)

Para o Exemplo 7.5, na comparação das mudanças de peso entre os grupos de terapia cognitivo-comportamental e o de controle no estudo da anorexia (página 222), o teste t paramétrico tinha um valor-p bilateral igual a 0,10. A versão para grandes amostras do teste de Wilcoxon-Mann-Whitney apresenta um resultado similar, com um valor-p de 0,11.

Alguns *softwares* também podem apresentar um intervalo de confiança correspondente para a diferença entre as medianas populacionais. O método supõe que as duas distribuições tenham a mesma forma, mas não necessariamente normal. A mudança mediana do peso foi de 1,4 libras para o grupo de terapia cognitivo-comportamental e −0,35 libras para o grupo de controle. O *software* apresenta um intervalo de 95% de confiança para a diferença entre as medianas de (−0,6; 8,1) libras.

Tamanho do efeito: proporção de melhores respostas para um grupo

A Seção 7.5 mencionou que o tamanho da diferença entre dois grupos é algumas vezes resumido pelo *tamanho do efeito*, o qual para duas amostras é definido como $(\bar{y}_1 - \bar{y}_2)/s$. Quando as distribuições apresentam assimetrias acentudas ou têm valores atípicos, as médias são tão úteis e o resultado do tamanho do efeito pode não ser apropriado. Uma medida do tamanho do efeito não paramétrica é a proporção dos pares das observações (uma para cada grupo) para a qual a observação do primeiro grupo era maior. Se y_1 representa uma observação aleatoriamente selecionada do Grupo 1 e y_2 representa uma observação aleatoriamente selecionada do Grupo 2, então essa medida estima $P(y_1 > y_2)$.

Para ilustrar, suponha que o estudo da anorexia tinha 4 meninas, 2 usando a nova terapia e 2 no grupo de controle. Suponha que as mudanças do peso foram:

Grupo com terapia (y_1): 4, 10
Grupo de controle (y_2): 2, 6

Existem quatro pares de observações, com uma em cada grupo:

$y_1 = 4, y_2 = 2$ (A do grupo 1 é maior)
$y_1 = 4, y_2 = 6$ (A do grupo 2 é maior)
$y_1 = 10, y_2 = 2$ (A do grupo 1 é maior)
$y_1 = 10, y_2 = 6$ (A do grupo 1 é maior)

A observação do Grupo 1 é maior em 3 dos 4 pares, assim a estimativa de $P(y_1 > y_2)$ é de 0,75. Se duas observações tivessem o mesmo valor, iríamos contá-la como y_1 sendo maior para ½ do par (em vez de 1 ou 0).

Sob H_0 (sem efeito), $P(y_1 > y_2) = 0{,}50$. Quanto mais longe $P(y_1 > y_2)$ está de 0,50, mais forte o efeito. Para todo o conjunto de dados da anorexia analisado no Exemplo 7.7 na página 226, a estimativa da amostra de $P(y_1 > y_2)$ é de 0,63. A probabilidade estimada de que uma menina usando a terapia cognitivo-comportamental tenha um ganho de peso maior do que uma menina usando a terapia de controle é de 0,63.

Quando os dois grupos têm distribuições normais com o mesmo desvio padrão, existe uma conexão entre esse tamanho do efeito e o do paramétrico, $(\mu_1 - \mu_2)/\sigma$. Por exemplo, quando $(\mu_1 - \mu_2)/\sigma = 0$, então $P(y_1 > y_2) = 0{,}50$; quando $(\mu_1 - \mu_2)/\sigma = 0{,}5$, então $P(y_1 > y_2) = 0{,}64$; quando $(\mu_1 - \mu_2)/\sigma = 1$, então $P(y_1 > y_2) = 0{,}71$; quando $(\mu_1 - \mu_2)/\sigma = 2$, então $P(y_1 > y_2) = 0{,}92$. O efeito é relativamente forte se $P(y_1 > y_2)$ for maior do que aproximadamente 0,70 ou menor do que aproximadamente 0,30.

Tratando as variáveis ordinais como quantitativas

Os cientistas sociais geralmente usam os métodos de estatística paramétrica para dados quantitativos com variáveis que são somente ordinais. Eles fazem isto atribuindo escores às categorias ordenadas. O Exemplo 6.2 (página 175), sobre ideologia política, mostrou um exemplo disso. Algumas vezes a escolha dos escores é direta. Para categorias (liberal, moderado, conservador) para a ideologia política, todo o conjunto de escores igualmente espaçados é lógico, como (1, 2, 3) ou (0, 5, 10). Quando a escolha não for clara, como com as categorias (não muito feliz, moderadamente feliz, muito feliz) para a felicidade, é uma boa ideia executar um estudo de sensibilidade. Escolha dois ou três conjuntos razoáveis de escores potenciais, como, por exemplo, (0, 5, 10), (0, 6, 10), (0, 7, 10), e verifique se as conclusões finais são similares para cada um. Se não, qualquer relatório deveria ressaltar como as conclusões dependem dos escores escolhidos.

De forma alternativa, os métodos não paramétricos são válidos com os dados originais. A razão é que os métodos não paramétricos não usam escores quantitativos, mas classificações das observações, e elas são informações ordinais. Entretanto, esta abordagem funciona melhor quando a variável resposta é contínua (ou aproximadamente), assim cada observação tem a sua própria classificação. Quando usados com respostas categóricas ordenadas, tais métodos são geralmente menos sensíveis do que quando usamos os métodos paramétricos que tratam a resposta como quantitativa, como o próximo exemplo ilustra.

EXEMPLO 7.10 O consumo do álcool e má-formação do bebê

A Tabela 7.9 se refere a um estudo do consumo de álcool por grávidas e da má-formação congênita. Após os três primeiros meses de gravidez, as mulheres da amostra completaram um questionário sobre o consumo do álcool. Depois do nascimento da criança, as observações foram registradas sobre a presença ou ausência de má-formação congênita dos órgãos genitais. O consumo do álcool foi mensurado como o número médio de drinques por dia.

O consumo de álcool está associado à má-formação? Uma abordagem para investigar isso é comparar o consumo médio de álcool de mães para os casos em que a má-formação ocorreu ao consumo por mães onde ela não ocorreu. O consumo de álcool foi mensurado agrupando valores de uma variável quantitativa. Para encontrar as médias, atribuímos escores ao consumo de álcool que são meios pontos das categorias; isto é, 0; 0,5; 1,5; 4,0; 7,0, o último escore (para ≥ 6) sendo arbitrário. As médias amostrais são, então, 0,28 para o grupo da ausência e 0,40 para o grupo da presença, e a estatística t de 2,56 tem um valor-p de 0,01. Existe uma forte evidência de que as mães cujos bebês sofreram má-formação tinham uma média mais alta do consumo do álcool.

Uma abordagem alternativa não paramétrica atribui classificações aos sujeitos e as usa como as categorias dos escores. Para todos os sujeitos em uma categoria,

Tabela 7.9 Má-formação dos bebês e o consumo de álcool pelas mães

	Consumo de álcool				
Má-Formação	**0**	**< 1**	**1 – 2**	**3 – 5**	**≥ 6**
Ausência	17066	14464	788	126	37
Presença	48	38	5	1	1
Total	17114	14502	793	127	38

Fonte: GRAUBARD, B. I., KORN, E. L. *Biometrics*, v. 43, p. 471-76, 1987.

atribuímos a média das classificações que se aplicaria a uma classificação completa da amostra. Elas são chamadas de *escores do meio*. Por exemplo, os 17114 sujeitos no nível 0 para o consumo de álcool dividem a classificação 1 a 17114. Atribuímos a cada um deles a média dessas classificações, que é o escore do meio (1 + 17114)/2 = 8557,5. Os 14502 sujeitos no nível < 1 para o consumo do álcool dividem as classificações 17115 a 17114 + 14502 = 31616, para um escore do meio de (17115 + 31616)/2 = 24365,5. Da mesma forma, os escores do meio para as três últimas categorias são 32013, 32473 e 32555,5. Usados, em um teste Wilcoxon para grandes amostras, esses escores geram muito menos evidência de um efeito (valor-p = 0,55).

Tabela 7.10 Resumo dos métodos de comparação para dois grupos para amostras aleatórias independentes

	Tipo de variável resposta	
	Categórica	**Quantitativa**
Estimação		
1. Parâmetro	$\pi_2 - \pi_1$	$\mu_2 - \mu_1$
2. Estimativa por ponto	$\hat{\pi}_2 - \hat{\pi}_1$	$\bar{y}_2 - \bar{y}_1$
3. Erro padrão	$ep = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$	$ep = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
4. Intervalo de confiança	$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(ep)$	$(\bar{y}_2 - \bar{y}_1) \pm t(ep)$
Teste de significância		
1. Suposições	Aleatorização ≥ 10 observações em cada categoria, para cada grupo	Aleatorização Distribuições normais (robusta, especialmente para n grande)
2. Hipóteses	$H_0: \pi_1 = \pi_2$ $(\pi_2 - \pi_1 = 0)$ $H_a: \pi_1 \neq \pi_2$	$H_0: \mu_1 = \mu_2$ $(\mu_2 - \mu_1 = 0)$ $H_a: \mu_1 \neq \mu_2$
3. Estatística-teste	$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{ep_0}$	$t = \frac{\bar{y}_2 - \bar{y}_1}{ep}$
4. Valor-p	Probabilidade bilateral da normal padrão ou t (Use a alternativa unilateral)	

Por que isso acontece? As categorias adjacentes que têm relativamente poucas observações necessariamente têm escores do meio similares. Os escores do meio (8557,5; 24365,5; 32013; 32473; 32555,5) são similares para as três categorias finais, visto que essas categorias têm consideravelmente menos observações do que as duas primeiras categorias. Uma consequência é que esse esquema de escores trata o nível 1 – 2 do consumo de álcool (categoria 3) mais próximo do nível de consumo ≥ 6 (categoria 5) do que do nível 0 de consumo (categoria 1). Isso parece inapropriado. É melhor usar seu julgamento selecionando escores que reflitam bem as distâncias entre as categorias. ■

Embora os métodos não paramétricos tenham a vantagem de suposições mais fracas, na prática os cientistas sociais não os usam tanto quanto os métodos paramétricos. Parte disso reflete os tamanhos da amostra grandes para a maioria dos estudos, para os quais as suposições sobre as distribuições da população não são vitais. Além disso, os métodos não paramétricos para conjuntos de dados multivariados não são tão desenvolvidos quanto os métodos paramétricos. A maioria dos métodos não paramétricos está além do alcance deste livro. Para mais detalhes, veja Hollander e Wolfe (1999).

7.8 RESUMO DO CAPÍTULO

Este capítulo introduziu métodos para a comparação de dois grupos. Para variáveis resposta quantitativas, as inferências se aplicam à diferença $\mu_2 - \mu_1$ entre as médias populacionais. Para variáveis resposta categóricas, as inferências se aplicam à diferença $\pi_2 - \pi_1$ entre as proporções populacionais.

Em cada caso, o teste de significância analisa se 0 é uma diferença plausível. Se o intervalo de confiança contém 0, é plausível que os parâmetros sejam iguais. A Tabela 7.10 resume os métodos para amostras aleatórias **independentes**, para as quais as observações nas duas amostras não são equiparadas. Este é o caso mais comum na prática.

- Tanto para a diferença de proporções quanto para a de médias, os intervalos de confiança têm a forma:

  Diferença estimada ± (escore)(ep),

  usando um escore-z para as proporções e escore-t para as médias. Em cada caso, a estatística-teste é igual à diferença estimada dividida pelo erro padrão.
- Para amostras **dependentes**, cada observação em uma amostra se equipara a uma na outra amostra. Para variáveis quantitativas, comparamos médias analisando a diferença das médias dos escores calculada entre as observações pareadas. O intervalo das **diferenças pareadas** e procedimentos dos testes são métodos de uma amostra dos Capítulos 5 e 6 aplicados à diferença dos escores.
- Outra abordagem para comparar médias faz a suposição extra de que as distribuições da população normal têm desvios padrão iguais. Essa abordagem agrupa os desvios padrão das duas amostras para encontrar uma estimativa comum.
- Para comparar proporções, com amostras independentes e pequenas o teste adequado é o **exato de Fisher**. Para amostras dependentes, o **teste de McNemar** compara o número de sujeitos que estão na categoria 1 na primeira amostra e na categoria 2 na segunda ao número de sujeitos que estão na categoria 2 na primeira amostra e na categoria 1 na segunda.
- Os métodos estatísticos **não paramétricos** não fazem suposição sobre a forma da distribuição da população. A maioria dos métodos usa os escores (postos) das observações.

Nesse estágio, você pode estar confuso sobre qual método usar para qualquer situação dada. Talvez possa ajudar se você seguir a seguinte lista de verificação. Pergunte a você mesmo se a análise é sobre

- Médias ou proporções (variável resposta quantitativa ou categórica)?
- Amostras independentes ou dependentes?
- Intervalo de confiança ou teste de significância?

EXERCÍCIOS

Praticando o básico

7.1 Uma matéria da Associated Press (23 de fevereiro de 2007) sobre o levantamento de dados anual dos calouros da UCLA indicou que 73% dos calouros da faculdade em 2006 consideravam que ser financeiramente próspero era muito importante, comparado com 42% em 1966 (o primeiro ano em que o levantamento de dados foi feito). A matéria também relatou que 81% dos jovens entre 18 e 25 anos nos Estados Unidos consideram ficar rico como principal objetivo na vida. Os percentuais amostrais de 42% em 1966 e 73% em 2006 são baseados em amostras independentes ou amostras dependentes? Explique.

7.2 *Transatlantic Trends* é um levantamento de dados anual da opinião pública norte-americana e europeia (veja www.transatlantictrends.org), com uma amostra aleatória de aproximadamente 1000 adultos de cada um de 13 países europeus a cada ano. Em 2002, 38% dos europeus expressaram uma atitude positiva sobre o tratamento das questões internacionais pelo presidente George W. Bush. Em 2006, 18% expressaram uma atitude positiva.

(a) Explique o significado desses resultados se eles forem baseados em (a) amostras *independentes*, (b) amostras *dependentes*.

(b) Se compararmos os resultados de 2002 e 2006, identifique a variável resposta e a variável explicativa e especifique se a variável resposta é quantitativa ou categórica.

7.3 O National Health Interview Survey (Levantamento Nacional com Entrevistas sobre Saúde) (www.cdc.gov/nchs) estimou que os fumantes eram 41,9% dos norte-americanos adultos em 1965 e 21,5% em 2003.

(a) Estime a diferença entre as proporções de fumantes nas duas datas consideradas.

(b) Suponha que o erro padrão foi relatado como 0,020 para cada proporção. Encontre o erro padrão para a diferença de proporções e interprete.

7.4 Quando um recente levantamento de dados do Eurobarometer perguntou aos sujeitos em cada um dos países europeus se eles estariam dispostos a pagar mais pela energia produzida de fontes renováveis do que a energia produzida de outras fontes, a proporção que respondeu *sim* variou de um alto 0,52 na Dinamarca ($n = 1008$) a um baixo 0,14 na Lituânia ($n = 1002$). Para esta pesquisa:

(a) Estime a diferença das proporções populacionais de respostas *sim* entre a Dinamarca e a Lituânia.

(b) Da fórmula $ep = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ do Capítulo 5, a estimativa da proporção tem $ep = 0{,}0157$ para a Dinamarca e $ep = 0{,}110$ para a Lituânia. Use isto para encontrar o *ep* para a diferença da estimativa em (a). Interprete este *ep*.

7.5 O National Center for Health Statistics (Centro Nacional de Estatísticas da Saúde) estimou, recentemente, que o peso médio para mulheres adultas norte-americanas era de 140 libras em 1962 e 164 libras em 2002.

(a) Suponha que essas estimativas tinham erros padrão de 2 libras a cada ano. Estime o aumento do peso médio na população de 1962 a 2002 e encontre e interprete o erro padrão da estimativa.

(b) Mostre que a média estimada em 2002 era de 1,17 vezes a média estimada de 1962. Expresse isto em termos de percentual de aumento.

(c) Os pesos médios estimados para homens eram de 166 libras em 1962 e 191 libras em 2002. Encontre e interprete a diferença e a razão.

7.6 A U.S. Census Bureau (Agência do Censo dos Estados Unidos) relatou em 2002 que o rendimento líquido mediano nos Estados Unidos foi estimado em aproximadamente \$89000 para domicílios brancos e \$6000 para domicílios negros.

(a) Identifique a variável resposta e a variável explicativa.

(b) Compare os grupos usando uma (i) diferença, (ii) razão.

7.7 De acordo com o U.S. Department of Justice (Departamento de Justiça dos Estados Unidos), em 2002 a taxa de encarceramento nas prisões nacionais era de 832 por 100000 moradores masculinos e 58 por 100000 moradores femininos.

(a) Encontre o risco relativo de estar preso, comparando homens a mulheres. Interprete.

(b) Encontre a diferença das proporções encarceradas. Interprete.

(c) Qual das medidas você acha que representa melhor estes dados? Por quê?

7.8 De acordo com o U.S. National Center for Health Statistics (Centro Nacional da Estatística da Saúde dos Estados Unidos), a probabilidade anual de que um homem com idade entre 20 a 24 seja uma vítima de homicídio é aproximadamente 0,00164 para negros e 0,00015 para brancos.

(a) Compare estas taxas usando a diferença de proporções.

(b) Compare estas taxas usando o risco relativo.

(c) Qual das duas medidas parece resumir melhor os resultados quando ambas as proporções estão muito próximas a 0? Explique.

7.9 Uma matéria da Associated Press (7 de agosto de 2006) sobre uma pesquisa investigando o impacto sobre os adolescentes de letras de música com conteúdo sexual relatou: "Os adolescentes que disseram que escutavam muitas músicas com mensagens sexuais degradantes tinham aproximadamente duas vezes mais probabilidade de iniciar relações sexuais... dentro dos dois anos seguintes do que os adolescentes que escutavam pouca ou nenhuma música com mensagens sexuais degradantes". Os percentuais relatados foram 51% e 29%.

(a) Um intervalo de 95% de confiança para a diferença entre as proporções populacionais correspondentes é (0,18; 0,26). Explique como interpretá-lo.

(b) O valor-p é $< 0{,}001$ para testar a hipótese nula de que as proporções populacionais correspondentes são iguais. Interprete.

7.10 Para uma amostra aleatória de canadenses, 60% indicam que aprovam o desempenho do primeiro ministro. Uma pesquisa similar, um mês depois, teve uma avaliação favorável de 57%. Um intervalo de 99% de confiança para a mudança nas proporções populacionais foi de $(-0{,}07; 0{,}01)$. Explique por que (a) pode não ter tido mudança no apoio, (b) se uma diminuição no apoio ocorreu, pode ter sido muito importante, (c) se um aumento do apoio ocorreu, ele foi provavelmente pequeno para ter uma importância substancial.

7.11 O College Alcohol Study (Estudo do Álcool na Faculdade) da Harvard School of Public Health (Escola de Saúde Pública de Harvard) entrevistou amostras aleatórias de estudantes do quarto ano da faculdade várias vezes desde 1993. Dos estudantes que disseram consumir álcool, o percentual que alegou que beber "para ficar bêbado" é um motivo importante para consumir álcool foi de 39,9% de 12708 estudantes em 1993 e 48,2% dos 8783 estudantes em 2001.[6] Para comparar os resultados entre 1993 e 2001:

(a) Mostre que o erro padrão para a diferença da estimativa entre as proporções populacionais correspondentes em 2001 e em 1993 é igual a 0,0069.

(b) Mostre que o intervalo de 95% de confiança para a diferença é (0,07; 0,10). Interprete.

7.12 No estudo mencionado no exercício anterior, o percentual que disse que mantiveram atividades sexuais não planejadas por causa da ingestão de álcool foi de 19,2% em 1993 e 21,3% em 2001.

(a) Especifique as suposições, notação e hipóteses para um teste bilateral comparando as proporções populacionais correspondentes.

(b) A estatística-teste foi $z = 3{,}8$ e o valor-p = 0,0002. Interprete o valor-p.

(c) Alguém pode argumentar que o resultado em (b) reflete *significância estatística*, mas não *significância prática*. Explique a base desse argumento e justifique por que você tem mais informação em um intervalo de 95% de confiança, que é (0,009; 0,033).

7.13 Para o Levantamento do Uso do Tempo da Tabela 7.1 (página 212), para aqueles que trabalham em tempo integral, 55% dos 1219 homens e 74% das 733 mulheres relataram que gastam algum tempo cozinhando e limpando em um dia típico. Encontre e interprete um intervalo de 95% de confiança para a diferença na participação das taxas.

7.14 A Tabela 7.11 resume as respostas de Pesquisas Sociais Gerais de 1977 e de 2006 à pergunta ("FEFAM"): "É melhor para todos envolvidos se o homem é o empreendedor fora de casa e a mulher é a que toma conta da casa e da família". Considere π_1 a representação da proporção da população que concordou com essa afirmação em 1977 e considere π_2 a representação da proporção da população em 2006.

(a) Mostre que $\hat{\pi}_1 - \hat{\pi}_2 = 0{,}30$, com erro padrão 0,0163.

(b) Mostre que um intervalo de 95% de confiança para $\pi_1 - \pi_2$ é (0,27; 0,33). Interprete.

(c) Explique como os resultados iriam diferir para a comparação das proporções daqueles que *não* concordam nos dois anos.

Tabela 7.11

Ano	Concorda	Discorda	Total
1977	989	514	1503
2006	704	1264	1968

7.15 Considere o exercício anterior sobre o papel da mulher. Em 2004, dos 411 respondentes homens, 153 (37,2%) responderam *sim*. Das 472 respondentes mulheres, 166 (35,2%) responderam *sim*.

(a) Estabeleça uma notação e especifique as hipóteses para a suposição de nenhuma diferença entre as proporções da população de homens e mulheres que iriam responder *sim*.

(b) Estime a proporção da população presumindo H_0, encontre o erro padrão da diferença das proporções amostrais e encontre a estatística-teste.

(c) Encontre o valor-p para a alternativa bilateral. Interprete.

(d) Dos 652 respondentes tendo menos educação do que o grau universitário, 40,0% respondeu *sim*. Dos 231 respondentes tendo pelo menos grau universitário, 25,6% respondeu *sim*. Qual variável, gênero ou nível de educação parece ter tido maior influência na opinião? Em outras palavras, a opinião tende a diferir mais entre homens e mulheres ou entre aquele com maior ou menor nível de educação?

7.16 Em um levantamento de dados conduzido pela Wright State University, foi perguntado aos formandos se eles haviam, alguma vez, usado maconha. A Tabela 7.12 mostra a saída do *software*. Tratando essas observações como uma amostra aleatória da população de interesse:

(a) Faça uma pergunta de pesquisa que poderia ser respondida com essa saída.

(b) Interprete o intervalo de confiança apresentado.

(c) Interprete o valor-p apresentado.

Tabela 7.12

```
Amostra     Sim    N      Proporção
                          amostral
1. Mulher   445    1120   0,3973
2. Homem    515    1156   0,4455
Estimativa para p(1) - p(2): -0,0482
IC 95% para p(1) - p(2): (-0,0887, -0,0077)
Teste para a diferença = 0 (versus ≠ 0):
z = -2,33 valor-p = 0,020
```

7.17 Um estudo sobre o comportamento compulsivo de fazer compras (compulsão incontrolável de comprar) realizou um levantamento de dados nacional por telefone em 2004 com adultos com 18 anos e acima.[7] De 800 homens, 44 foram julgados como compradores compulsivos de acordo com a Escala de Compras Compulsivas. Das 1501 mulheres, 90 foram julgadas como compradoras compulsivas. Conduza uma inferência para analisar se um sexo tem maior probabilidade do que o outro de ser um comprador compulsivo. Interprete.

7.18 A Tabela 7.13 mostra os resultados de uma Pesquisa Social Geral recente com duas variáveis, sexo e se a pessoa acredita na vida após a morte ("AFTERLIF"). Conduza todas as etapas de um teste de significância, usando $\alpha = 0{,}05$, para comparar as proporções populacionais de mulheres e homens que responderam que acreditam na vida após a morte. Se você cometer um erro na sua decisão, que tipo de erro seria? Do Tipo I ou do Tipo II?

Tabela 7.13

	Acredita na vida após a morte		
Sexo	Sim	Não ou Indeciso	Total
Mulher	435	147	582
Homem	375	134	509

7.19 Uma PSG relatou que as 486 mulheres tinham uma média de 8,3 bons amigos (s = 15,6) e 354 homens tinham uma média de 8,9 bons amigos (s = 15,5).

(a) Um intervalo de 95% de confiança para a diferença entre as médias populacionais para os homens e para as mulheres é de (−1,5; 2,7). Interprete.

(b) Para cada sexo, parece que a distribuição dos bons amigos é normal? Explique por que isto não invalida o resultado em (a), mas pode afetar a utilidade do intervalo.

7.20 A Tabela 7.14 resume o número de horas gastas no trabalho doméstico, por semana e por gênero, baseado na PSG de 2002 (variável "RHHWORK").

(a) Estime a diferença entre as médias populacionais para mulheres e homens.

(b) Mostre que o erro padrão estimado da diferença da amostra é de 0,81. Interprete.

(c) Mostre que um intervalo de 99% de confiança para a diferença é de (2,3; 6,5). Interprete.

Tabela 7.14

	Horas do trabalho doméstico		
Gênero	Tamanho da amostra	Média	Desvio padrão
Homens	292	8,4	9,5
Mulheres	391	12,8	11,6

7.21 Um estudo de 30 dias avaliou o grau de dependência que os adolescentes têm quando iniciam sua experiência com o fumo.[8] O estudo usou uma amostra aleatória de 332 estudantes da 7ª série de duas cidades de Massachusetts que nunca haviam fumado antes do início do estudo. A variável resposta foi construída da Hooked on Nicotine Checklist (HONC). Esta é uma lista de 10 perguntas como: "Você já tentou parar, mas não conseguiu?". O escore da HONC é o número total de perguntas às quais o estudante respondeu *sim*. Quanto mais alto o escore, maior a dependência da nicotina. Havia 75 fumantes e 257 ex-fumantes no final do estudo. As médias da HONC descrevendo a dependência da nicotina foram 5,9 (s

= 3,3) para os fumantes e 1,0 (s = 2,3) para ex-fumantes.

(a) Encontre e interprete uma estimativa por ponto para comparar as médias da HONC para fumantes e ex-fumantes.

(b) O *software* relata um intervalo de 95% de confiança como sendo (4,1; 5,7). Interprete.

(c) A distribuição dos dados amostrais da HONC para os ex-fumantes era aproximadamente normal? Como isto afeta a inferência?

7.22 Considere o Exercício 7.17 sobre o comportamento compulsivo de comprar. O total da fatura do cartão de crédito apresentou uma média de $3399 com desvio padrão de $5595 para 100 compradores compulsivos e uma média de $2837 com desvio padrão de $6335 para outros 1682 respondentes.

(a) Estime a diferença entre as médias para compradores compulsivos e outros respondentes e encontre o erro padrão.

(b) Compare as médias populacionais usando um teste de significância bilateral. Interprete.

7.23 Uma PSG recente perguntou: "Em quantos dias, nos últimos 7 dias, você esteve triste?". O *software* relatou médias amostrais de 1,8 para mulheres e 1,4 para homens, com um intervalo de 95% de confiança comparando-as de (0,2; 0,6), uma estatística t de 4,8 e um valor-p de 0,000. Interprete esses resultados.

7.24 Para a PSG de 2006, uma comparação entre mulheres e homens quanto ao número de horas por dia assistindo à televisão forneceu os seguintes resultados:

Grupo	N	Média	Desvio Padrão	EP da média
Mulheres	1117	2,99	2,34	0,070
Homens	870	2,86	2,22	0,075

(a) Realize todas as etapas de um teste de significância para analisar se as médias populacionais diferem para homens e mulheres. Interprete o valor-p e relate a sua conclusão para um nível $\alpha = 0{,}05$.

(b) Se você fosse construir um intervalo de 95% de confiança comparando as médias, ele iria conter 0? Responda com base no resultado em (a), sem encontrar o intervalo.

(c) Você acha que a distribuição do número de horas assistindo à televisão é aproximadamente normal? Por que ou por que não? Isto afeta a validade de suas inferências?

7.25 Para a PSG de 2004, a Tabela 7.15 mostra a saída do *software* para avaliar o número de horas assistindo à televisão por dia por raça.

☑ **Tabela 7.15**

Raça	N	Média	Desv. padrão	EP da média
Negra	101	4,09	3,63	0,3616
Branca	724	2,59	2,31	0,0859

```
Diferença = μNegra - μBranca
Estimativa para a diferença: 1,50
IC de 95% para a diferença: (0,77; 2,23)
Teste t da diferença = 0: Valor-t = 4,04,
Valor-p = 0,000
```

(a) Interprete o intervalo de confiança encontrado. Você pode concluir que uma das médias populacionais é mais alta? Se pode, qual? Explique.

(b) Interprete o valor-p encontrado.

(c) Explique a conexão entre o resultado do teste de significância e o resultado do intervalo de confiança.

7.26 Um estudo[9] comparou características da personalidade entre filhos adultos de alcoólatras e um grupo de controle equiparado por idade e gênero. Para os 29 pares de mulheres, os autores relataram uma média de 24,8 para o bem-estar dos filhos dos alcoólatras e uma média de 29,0 para o grupo de controle. Eles relataram t = 2,67 para o teste comparando as médias. Supondo que esse é o resultado de uma análise de amostras dependentes, identifique o gl da estatística-teste t, determine e interprete o valor-p.

7.27 Um experimento de diferenças pareadas[10], tratando do tempo de reação a um ruído, sob duas condições, utilizou uma amostra de crianças de 9 meses e encontrou uma diferença média de 70,1 com um desvio padrão de 49,4 para a diferença. Na sua argumentação, os autores relataram uma estatística t de 4,9 tendo um valor-p < 0,01 para uma alternativa bilateral. Mostre como eles construíram a estatística t e confirme o valor-p.

7.28 Como parte do seu projeto de aula, uma estudante da Universidade da Flórida selecionou aleatoriamente 10 colegas estudantes para investigar suas atividades sociais mais comuns. Como parte do estudo, ela pediu aos estudantes para declarar quantas vezes eles fizeram cada uma das seguintes atividades durante o ano anterior: ir ao cinema, ir a um evento esportivo ou ir a uma festa. A Tabela 7.16 mostra os dados.

(a) Para comparar a frequência média de ida ao cinema com a de comparecimento a eventos esportivos usando inferência estatística, deveríamos tratar as amostras como independentes ou dependentes? Por quê?

(b) Para a análise em (a), o *software* mostra os seguintes resultados:

	N	Média	Desvio padrão	EP da média
cinema	10	13,000	13,174	4,166
esportes	10	9,000	8,380	2,650
Diferença	10	4,000	16,166	5,112

```
IC de 95% para a diferença média:
(-7,56, 15,56)
Teste t para a diferença média = 0
(versus ≠ 0)
Valor-t = 0,78 Valor-p = 0,454
```

Interprete o intervalo de 95% de confiança encontrado.

(c) Mostre como a estatística-teste apresentada na saída computacional foi obtida da outra informação dada. Relate o valor-p e interprete-o no contexto.

☑ **Tabela 7.16**

	Atividade		
Estudante	Cinema	Esportes	Festas
1	10	5	25
2	4	0	10
3	12	20	6
4	2	6	52
5	12	2	12
6	7	8	30
7	45	12	52
8	1	25	2
9	25	0	25
10	12	12	4

7.29 Considere o exercício anterior. Para comparar ida a festas e a eventos esportivos, o *software* mostra um intervalo de 95% de confiança de (−3,33; 28,93) e um valor-p de 0,106.

(a) Interprete o valor-p.

(b) Explique a conexão entre os resultados do teste e o intervalo de confiança.

7.30 Uma psicóloga clínica quer escolher entre duas terapias para tratar a depressão. De seis pacientes, ela seleciona aleatoriamente três para receber a terapia A e os outros três recebem a terapia B. Ela utiliza pequenas amostras por razões éticas, pois se o seu experimento indicar que uma terapia é superior, então ela será usada nos demais pacientes com os mesmos sintomas. Após um mês de tratamento, a melhora é medida pela mudança no escore de uma escala padronizada de gravidade da depressão. Os escores da melhora são 10, 20, 30 para os pacientes que recebem a terapia A e 30, 45, 45 para os pacientes que recebem a terapia B.

(a) Usando um método que supõe um desvio padrão comum para as duas terapias, mostre que o s combinado = 9,35 e o ep = 7,64.

(b) Quando os tamanhos da amostra são muito pequenos, pode valer a pena sacrificar alguma confiança para conseguir maior precisão.

Mostre que um intervalo de 90% de confiança para $(\mu_2 - \mu_1)$ é (3,7; 36,3). Interprete.

(c) Estime e resuma o tamanho do efeito.

7.31 Considere o exercício anterior. Para evitar tendenciosidade das amostras não sendo balanceadas com um n tão pequeno, a psicóloga delineia novamente o experimento. Ela forma três pares de sujeitos, tal que os pacientes equiparados em qualquer par dado sejam similares em saúde e *status* socioeconômico. Para cada par, ela seleciona aleatoriamente um sujeito para cada terapia. A Tabela 7.17 mostra os escores da melhora e a Tabela 7.18 mostra os resultados usando o SPSS para analisar os dados.

(a) Compare as médias (i) encontrando a diferença das médias amostrais para as duas terapias, (ii) encontrando a média da diferença dos escores. Compare.

(b) Verifique o desvio padrão das diferenças e o erro padrão para a diferença média.

(c) Verifique o intervalo de confiança mostrado para a diferença média populacional. Interprete.

(d) Verifique a estatística-teste, o gl e o valor-p para comparar as médias. Interprete.

☑ **Tabela 7.17**

Par	Terapia A	Terapia B
1	10	30
2	20	45
3	30	45

7.32 Um estudo[11] sobre bulimia entre universitárias considerou o efeito do abuso sexual na infância em vários componentes de uma Escala do Ambiente Familiar. Para uma medida da coesão da família, a média amostral para os estudantes bulímicos era de 2,0 para as 13 estudantes que sofreram abuso sexual e de 4,8 para as 17 estudantes que não sofreram abuso sexual. A Tabela 7.19 mostra os resultados do *software* de uma comparação de médias de duas amostras.

☑ **Tabela 7.19**

Variável: COESÃO

ABUSO	N	Média	DP	Erro Padrão
sim	13	2,0	2,1	0,58
não	17	4,8	3,2	0,78

Variâncias	T	GL	Valor-p
Desiguais	2,89	27,5	0,007
Iguais	2,73	28	0,011

☑ **Tabela 7.18**

Teste t para amostras emparelhadas

Variável	Número de Pares	Média	DP	EP da média
	3			
TERAPIA A		20,000	10,000	5,774
TERAPIA B		40,000	8,660	5,000

Diferenças Pareadas

Média	DP	EP da média	Valor-t	gl	Sig. bilateral
20,000	5,00	2,887	6,93	2	0,020

IC de 95% (7,58; 32,42)

(a) Supondo os desvios padrão populacionais iguais, construa um intervalo de 95% de confiança para a diferença na coesão média familiar para estudantes que sofreram abuso sexual e das que não sofreram. Interprete.

(b) Explique como interpretar os resultados dos testes de significância desta saída.

7.33 Para o levantamento de dados dos estudantes descritos no Exercício 1.11 (página 25), as respostas sobre a ideologia política tinham uma média de 3,18 e um desvio padrão de 1,72 para os 51 estudantes não vegetarianos e uma média de 2,22 e desvio padrão de 0,67 para os 9 estudantes vegetarianos. Quando usamos um *software* para comparar as médias com um teste de significância, obtemos:

Variâncias	T	GL	Valor-p
Desiguais	2,915	30,9	0,0066
Iguais	1,636	58,0	0,1073

Explique por que os resultados dos dois testes diferem tanto e dê a sua própria conclusão sobre se as médias populacionais são iguais.

7.34 Em 2006, a PSG perguntou sobre o número de horas por semana gastas na *web* ("WWWTIME"). As 1569 mulheres apresentaram uma média de 4,9 com um desvio padrão de 8,6. Os 1196 homens apresentaram uma média de 6,2 com um desvio padrão de 9,9. Use esses resultados para fazer uma inferência comparando mulheres e homens em WWWTIME na população, supondo desvios padrão populacionais iguais.

7.35 Dois novos cursos foram propostos para ajudar os estudantes que sofrem de fobia severa a matemática, tendo um escore de pelo menos 8 em uma medida de fobia de matemática que está entre 0 e 10 (com base em respostas a 10 questões.) Uma amostra de 10 estudantes foi aleatoriamente alocada para os dois cursos. Após o curso, a queda no escore de fobia da matemática foi registrada. Os valores amostrais foram:

Curso A: 0, 2, 2, 3, 3

Curso B: 3, 6, 6, 7, 8

(a) Faça uma comparação inferencial das médias supondo desvios padrão populacionais iguais. Interprete os seus resultados.

(b) Usando *software*, relate e interprete o valor-p para o teste bilateral de Wilcoxon.

(c) Encontre e interprete o tamanho do efeito $(\bar{y}_B - \bar{y}_A)/s$.

(d) Estime e interprete o tamanho do efeito $P(y_B > y_A)$.

7.36 Uma PSG perguntou às pessoas se elas acreditavam no paraíso ou no inferno. Dos 1120 sujeitos entrevistados, 833 acreditavam em ambos, 160 não acreditavam em nenhum, 125 acreditavam no paraíso, mas não no inferno e 2 acreditavam no inferno, mas não no paraíso.

(a) Exiba os dados em uma tabela de contingência, cruzando a crença no paraíso (*sim, não*) com a crença no inferno (*sim, não*).

(b) Estime a proporção populacional de quem acredita no paraíso e de quem acredita no inferno.

(c) Mostre todas as etapas do teste McNemar para comparar as proporções populacionais e interprete.

(d) Construa um intervalo de 95% de confiança para comparar as proporções populacionais e interprete.

7.37 Uma PSG perguntou às pessoas a opinião sobre o gasto do governo com a saúde e com policiamento. Em cada situação o gasto deve aumentar ou diminuir? A Tabela 7.20 mostra as opiniões.

(a) Encontre a proporção amostral a favor do aumento do gasto para cada item.

(b) Teste se as proporções populacionais são iguais. Determine o valor-p e interprete.

(c) Construa um intervalo de 95% de confiança para a diferença das proporções. Interprete.

☑ **Tabela 7.20**

Gasto com saúde	Gasto com policiamento	
	Aumentar	Diminuir
Aumentar	292	25
Diminuir	14	9

7.38 Um estudo[12] usou dados do Estudo Longitudinal do Envelhecimento para investigar como a saúde e as características sociais de pessoas idosas são influenciadas pela convivência com os seus filhos. Considere a Tabela 7.21, que mostra se um idoso mora com um filho em um dado momento e, então novamente, quatro anos mais tarde. O autor suspeitava que, à medida que as pessoas envelhecessem e a saúde deteriorava, eles teriam maior probabilidade de viver com um filho. Os dados confirmam essa suspeita? Justifique a sua resposta com uma análise inferencial.

☑ **Tabela 7.21**

Primeiro momento	Quatro anos mais tarde	
	Sim	Não
Sim	423	138
Não	217	2690

7.39 Um estudo[13] investigou a orientação sexual de adultos que foram criados em famílias lésbicas. Vinte e cinco crianças de mães lésbicas e um grupo de controle de 20 crianças de mães heterossexuais foram vistos com a idade de 10 anos e, novamente, com a idade aproximada de 24 anos. Com a idade mais avançada, eles foram entrevistados sobre sua identidade sexual, com respostas possíveis *Bissexual/Lésbica/Homossexual* ou *Heterossexual*. A Tabela 7.22 mostra os resultados em um formato de saída do SAS para a realização do teste exato de Fisher.

(a) Por que o teste exato de Fisher é usado para comparar esses grupos?

(b) Determine e interprete o valor-p para a alternativa de que a proporção da população identificada como *bissexual/lésbica/homossexual* é maior para aqueles com mães lésbicas.

7.40 Considere o exercício anterior. Aos jovens adultos foi perguntado, também, se eles já tinham tido um relacionamento sexual com pessoas do mesmo sexo. A Tabela 7.23 mostra os resultados. Use *software* para testar se esta probabilidade é maior para aqueles criados por mães lésbicas. Interprete.

☑ **Tabela 7.22**

```
                  IDENTIDADE

Mãe           | B/L/H  | HETERO |  Total
--------------+--------+--------+
Lésbica       |    2   |   23   |   25
--------------+--------+--------+
Heterossexual |    0   |   20   |   20
--------------+--------+--------+
Total              2       43       45

ESTATÍSTICAS PARA A TABELA ORIENTAÇÃO SEXUAL
DA MÃE versus DOS FILHOS
Estatística                        Prob
---------------------------------------------
Teste Exato de Fisher (Esquerda) 1,000
                      (Direita)  0,303
                      (Bilateral)0,495
```


☑ **Tabela 7.23**

Mãe	Relacionamento com o mesmo sexo	
	Sim	Não
Lésbica	6	19
Heterossexual	0	20

Conceitos e aplicações

7.41 Para o arquivo de dados do Exercício 1.11 da página 25, compare a ideologia política dos estudantes identificados com o partido Democrata com aqueles identificados com o Republicano:

(a) usando resumos gráficos e numéricos.

(b) usando métodos estatísticos inferenciais. Interprete.

7.42 Usando *software* com o conjunto de dados dos estudantes (Exercício 1.11, na página 25), construa um intervalo de confiança e conduza um teste:

(a) Para comparar as opiniões de homens e mulheres sobre a legalização do aborto. Interprete.

(b) Para comparar o tempo médio semanal gasto assistindo à televisão ao gasto praticando esportes e outros exercícios físicos.

7.43 Para o arquivo de dados criado no Exercício 1.12 (página 26), com as variáveis escolhidas pelo seu professor, formule uma questão de pesquisa e conduza análises estatísticas inferenciais. Prepare um relatório que resuma suas descobertas. Neste relatório, use, também, métodos gráficos e numéricos para descrever os dados e, se necessário, verificar as suposições que você fez para a sua análise.

7.44 O Exercício 3.6 (página 80), mostrou dados, dos países industrializadas, sobre a emissão de dióxido de carbono, um responsável importante do aquecimento global. Existe diferença de emissões entre países europeus e não europeus? Realize uma investigação para responder a essa pergunta.

7.45 Proponha uma hipótese nula e uma alternativa sobre o relacionamento entre o tempo gasto na internet ("WWWHR" na PSG) e um previsor binário disponível na PSG que você acredite estar relacionado com o uso da internet. Usando os dados mais recentes da PSG dessas variáveis disponíveis em sda.berkerly.edu/GSS, realize um teste. Prepare um breve relatório resumindo as suas análises. (*Nota*: o *site* da PSG faz a comparação das médias de grupos, clicando em [*Comparison of means*].)

7.46 Passe os olhos por um ou dois jornais diários como o *The New York Times* (cópia impressa ou *online*). Copie um artigo sobre um estudo que comparou dois grupos. Prepare um breve relatório que responda as seguintes perguntas:

(a) Qual foi o propósito do estudo?

(b) Identifique a variável explicativa e a resposta.

(c) Você pode dizer se a análise estatística usou (1) amostras independentes ou amostras dependentes ou (2) uma comparação das proporções ou uma comparação das médias?

7.47 Um estudo recente[14] considerou se um tempo maior, gasto por adolescentes, assistindo à televisão estava associado a uma probabilidade maior de cometer atos agressivos ao longo dos anos. Os pesquisadores amostraram aleatoriamente 707 famílias em dois condados ao norte do estado de Nova Iorque e fizeram observações subsequentes ao longo de 17 anos. Eles observaram se um adolescente amostrado cometeu, mais tarde, algum ato agressivo contra outra pessoa, de acordo com um relatório pessoal daquela pessoa ou de sua mãe. Dos 88 casos com menos do que uma hora diária assistindo à televisão, 5 haviam cometido atos agressivos. Dos 619 casos com pelo menos uma hora diária assistindo à televisão, 154 haviam cometido atos agressivos. Analise esses dados e resuma sua análise em um breve relatório.

7.48 Quando perguntado pela PSG sobre o número de pessoas com as quais o sujeito tinha discutido assuntos importantes ao longo dos últimos seis meses (variável "NUMGIVEN"), a resposta 0 foi dada por 8,9% dos 1531 respondentes em 1985 e por 24,6% dos 1482 respondentes em 2004. Analise esses dados inferencialmente e interprete.

7.49 Um estudo[15] comparou o uso de substâncias, deliquência, bem-estar psicológico e apoio social entre vários tipos de famílias, para uma amostra urbana de adolescentes afro-americanos masculinos. A amostra continha 108 sujeitos de domicílios de mães solteiras e 44 de domicílios com os pais biológicos. Os jovens responderam a uma bateria de perguntas que forneceu uma medida visível do apoio dos pais. Esta medida tinha médias amostrais de 46 ($s = 9$) para os domicílios de mães solteiras e 42 ($s = 10$) para os domicílios com ambos os pais biológicos. Considere a conclusão: "O suporte médio dos pais era 4 unidades mais alto para os domicílios de mães solteiras. Se as médias verdadeiras fossem iguais, uma diferença deste tamanho poderia ser esperada somente 2% das vezes. Para amostras deste tamanho, poderíamos esperar que em 95% das vezes essa diferença fique aquém de 3,4 do seu valor verdadeiro".

(a) Explique como esta conclusão se refere aos resultados de (i) um intervalo de confiança, (ii) um teste.

(b) Descreva como você explicaria os resultados do estudo a alguém que não estudou estatística inferencial.

7.50 Os resultados na Tabela 7.24 são de um estudo[16] de atração física e bem-estar subjetivo. Uma amostra de estudantes universitários foi classificada por um painel (lista de jurados) de acordo com sua atratividade. A tabela apresenta o número de encontros nos últimos três meses para estudantes classificados no quartil superior ou inferior da atratividade. Analise esses dados e interprete.

7.51 Um relatório (04/12/2002) do Pew Research Center sobre "O que o mundo pensa em 2002" relatou que "o público norte-americano está visivelmente em desacordo com o resto do mundo sobre a visão do papel dos Estados Unidos no mundo e o impacto global das ações norte-americanas". As conclusões foram baseadas em pesquisas em vários países. No Paquistão, em 2002, o percentual de sujeitos entrevistados que tinham uma visão favorável dos Estados Unidos era de 10%, e o percentual que achou que difundir as ideias e costumes norte-americanos era bom foi de 2% ($n = 2032$).

(a) Você tem informação o suficiente para fazer uma comparação inferencial das proporções? Se tiver, faça. Se não, o que mais você precisa saber?

(b) Para um levantamento de dados separado em 2000, o percentual estimado dos que tinham uma visão favorável dos Estados Unidos era de 23%. Para comparar inferencialmente os percentuais em 2000 e 2002, o que mais você precisa saber?

7.52 Um artigo da *Time Magazine* intitulado "Disparidade de gênero no Wal-Mart" (5 de julho de 2004) declarou que, em 2001, as mulheres gerentes do Wal-Mart recebiam $14500 por ano menos, em média, do que seus colegas do sexo masculino. Se você tivesse, também, os erros padrão do salário médio anual para homens e mulheres gerentes do Wal-Mart, você teria informação suficiente para determinar se esta é uma diferença "estatisticamente significativa"? Explique.

☑ Tabela 7.24

Atratividade	Número de encontros (Homens) Média	Desvio padrão	n	Número de encontros (Mulheres) Média	Desvio padrão	n
Mais	9,7	10,0	35	17,8	14,2	33
Menos	9,9	12,6	36	10,4	16,6	27

7.53 O International Adult Literacy Survey (Levantamento de Dados Internacional de Alfabetização Adulta) (www.nifl.gov/nifl/facts/IALS.html) foi um estudo realizado em 22 países nos quais amostras nacionalmente representativas de adultos foram entrevistadas e testadas, em casa, com o uso de um mesmo teste de alfabetização que tinha escores variando de 0 a 500. Para aqueles com idade entre 16 e 25 anos, alguns escores médios de escrita foram: Reino Unido 273,5; Nova Zelândia 276,8; Irlanda 277,7; Estados Unidos 277,9; Dinamarca 283,4; Austrália 283,6; Canadá 286,9; Holanda 293,5; Noruega 300,4; Suécia 312,1. O *site* não fornece os tamanhos da amostra ou os desvios padrão. Suponha que cada tamanho da amostra era de 250 e cada desvio padrão era 50. Quão distantes devem estar as médias amostrais antes de você se sentir confiante de que existe uma diferença real entre as médias populacionais? Explique seu raciocínio dando sua conclusão para a diferença entre o Canadá e os Estados Unidos.

7.54 A Tabela 7.25 compara dois hospitais quanto às admissões de pacientes com pneumonia severa. Embora o *status* do paciente seja uma variável ordinal, dois pesquisadores que analisaram os dados a trataram como intervalar. O primeiro pesquisador atribuiu os escores (0, 5, 10) às três categorias. O segundo pesquisador, acreditando que a categoria do meio está mais próxima da terceira categoria do que da primeira, utilizou os escores (0, 9, 10). Cada pesquisador calculou as médias para as duas instituições e identificou a instituição com a média mais alta como a que tinha mais sucesso no tratamento dos seus pacientes. Encontre as duas médias para o sistema de escores usado pelo (a) primeiro pesquisador, (b) segundo pesquisador. Interprete. (Observe que a conclusão depende da escala de escores utilizada. Portanto, se você usar métodos para variáveis quantitativas com dados ordinais, tome cuidado com a escala dos escores.)

☑ Tabela 7.25

	Status do paciente Morreu no hospital	Liberado após longa permanência	Liberado após breve permanência
Hospital A	1	29	0
Hospital B	8	8	14

7.55 Do Exemplo 6.4 (página 177) do Capítulo 6, para o grupo da terapia cognitivo-comportamental a mudança média no peso de 3,0 libras foi significativamente diferente de zero. Contudo, o Exemplo 7.7 (página 226) mostrou que ela não é significativamente diferente da mudança média para o grupo de controle, embora aquele grupo tivesse uma mudança média amostral negativa. Como você explica este paradoxo? (*Dica:* das Seções 7.1 e 7.3, verifique como o valor do *ep* para estimar a diferença entre as duas médias se compara ao valor do *ep* para estimar uma única média?)

7.56 Um levantamento da Harris Poll de 2201 norte-americanos em 2003 indicou que 51% acreditavam em fantasmas e 31% acreditavam em astrologia.

(a) É válido comparar as proporções usando métodos inferenciais para amostras independentes? Explique.

(b) Você tem informação o suficiente para compará-las usando métodos inferenciais para amostras dependentes? Explique.

7.57 Um grupo de seis candidatos para três posições de gerência inclui três mulheres e três homens. A Tabela 7.26 mostra os resultados.

(a) Designe as três mulheres por F_1, F_2, F_3 e os três homens por M_1, M_2, M_3. Identifique as 20 amostras distintas do tamanho três que podem ser escolhidas destes seis indivíduos.

(b) Considere $\hat{\pi}_1$ a proporção amostral de homens selecionados e $\hat{\pi}_2$ a de mulheres. Para a Tabela 7.26, $\hat{\pi}_1 - \hat{\pi}_2 = (2/3) - (1/3) = 1/3$. Das 20 amostras possíveis, mostre que 10 têm $\hat{\pi}_1 - \hat{\pi}_2 \geq 1/3$. Na verdade, este é o raciocínio que fornece o valor-p unilateral para o teste exato de Fisher.

(c) Encontre o valor-p se todos os três selecionados são homens. Interprete.

☑ **Tabela 7.26**

	Escolhido para a posição	
Gênero	**Sim**	**Não**
Masculino	2	1
Feminino	1	2

7.58 Descreva uma situação na qual seria mais sensível comparar as médias usando variáveis dependentes do que variáveis independentes.

7.59 Uma notícia da Associated Press (1° de fevereiro de 2007) sobre um levantamento de dados da Universidade de Chicago de 1600 pessoas com idade entre 15 a 25 anos, em várias cidades do Meio-Oeste norte-americano, indicou que 58% de jovens negros, 45% de jovens hispânicos e 23% de jovens brancos disseram escutar *rap* todo o dia.

(a) Verdadeiro ou falso: se um intervalo de 95% de confiança comparando as proporções da população para jovens hispânicos e brancos era de (0,18; 0,26), então podemos inferir que pelo menos 18%, mas não menos do que 26% de jovens brancos da população branca correspondente escuta *rap* diariamente.

(b) O estudo relatou que 66% das mulheres negras e 57% dos homens negros concordaram que os vídeos de *rap* retratam as mulheres negras de forma ruim e ofensiva. Verdadeiro ou falso: pelo fato de que ambos os grupos tinham a mesma raça, os métodos inferenciais comparando-os devem assumir amostras dependentes em vez de independentes.

7.60 Verdadeiro ou falso? Se um intervalo de 95% de confiança para $(\mu_2 - \mu_1)$ contém somente números positivos, então podemos concluir que ambos μ_1 e μ_2 são positivos.

7.61 Verdadeiro ou falso? Se você conhece o erro padrão da média amostral para cada uma de duas amostras independentes, você pode descobrir o erro padrão da diferença entre as médias amostrais mesmo se você não souber os tamanhos das amostras.

Nos Exercícios 7.62 a 7.64, selecione a(s) resposta(s) correta(s). Mais de uma resposta pode estar correta.

7.62 Um intervalo de 99% de confiança para a diferença $\pi_2 - \pi_1$ entre as proporções de homens e mulheres da Califórnia que são alcoólatras é igual a (0,02; 0,09).

(a) Estamos 99% confiantes de que a verdadeira proporção está entre 0,02 e 0,09.

(b) Estamos 99% confiantes de que a verdadeira proporção de homens da Califórnia que são alcoólatras está entre 0,02 e 0,09 acima da verdadeira proporção de mulheres alcoólatras da Califórnia.

(c) Nesse nível de confiança, existe evidência insuficiente para inferir que as proporções populacionais são diferentes.

(d) Estamos 99% confiantes de que a minoria dos residentes da Califórnia são alcoólatras.

(e) Visto que os intervalos de confiança não contêm 0, é impossível que $\pi_1 = \pi_2$.

7.63 Para comparar o rendimento médio anual dos hispânicos (μ_1) e dos brancos (μ_2), com empregos na construção civil, construímos um intervalo de 95% de confiança para $\mu_2 - \mu_1$.

(a) Se o intervalo de confiança é (3000, 6000), então, nesse nível de confiança, concluímos que a renda média da população é maior para brancos do que para hispânicos.

(b) Se o intervalo de confiança é (−1000, 3000), então ao nível $\alpha = 0,05$ correspondente ao teste H_0: $\mu_1 = \mu_2$ contra H_a: $\mu_1 \neq \mu_2$ rejeitamos H_0.

(c) Se o intervalo de confiança é (−1000, 3000), então é plausível que $\mu_1 = \mu_2$.

(d) Se o intervalo de confiança é (−1000, 3000), então estamos 95% confiantes de que a renda média anual dos brancos está entre \$1000 e \$3000 acima do rendimento médio anual dos hispânicos.

7.64 O teste de Wilcoxon difere dos procedimentos paramétricos (para médias) no sentido de que:

(a) Ele se aplica diretamente às variáveis ordinais assim como a variáveis resposta intervalares.

(b) Não é necessário supor que a distribuição da população seja normal.

(c) A amostragem aleatória não é necessária.

***7.65** Um teste consiste em 100 questões do tipo verdadeiro ou falso. João não estudou, assim em cada questão ele tenta adivinhar a resposta correta.

(a) Encontre a probabilidade de que ele acerte pelo menos 70 e assim passe no exame. (*Dica:* use a distribuição amostral para a proporção de respostas corretas.)

(b) Joana estudou um pouco e tem uma chance de 0,60 de acertar cada questão. Encontre a probabilidade de que seu escore, não obstante, seja menor do que o do João. (*Dica:* use a distribuição amostral da diferença das proporções amostrais.)

(c) Como as respostas para (a) e (b) dependem do número de questões? Explique.

***7.66** Considere y_{i1} a representação da observação para o sujeito i no tempo 1, y_{i2} a observação para o sujeito i no tempo 2 e $y_i = y_{i2} - y_{i1}$.

(a) Considerando $\bar{y}_1$, $\bar{y}_2$ e $\bar{y}_d$ a representação das médias destas observações, mostre que $\bar{y}_d = \bar{y}_2 - \bar{y}_1$.

(b) A diferença mediana (isto é, a mediana dos valores y_i) é igual à diferença entre as medianas dos valores de y_{i1} e y_{i2}? Mostre que isto é verdade ou dê um contraexemplo para mostrar que é falso.

☑ NOTAS

1 www.statistics.gov.uk.

2 BENSON, H. et al. *American Heart Journal*, v. 151, p. 934-52, 2006.

3 Os dados são cortesia de David Strayer, Universidade de Utah. Veja STRAYER, D., JOHNSTON, W. *Psych. Science*, v. 21, p. 462-66, 2001.

4 CHEN, S., CHEN, W. *IEEE Transactions on Speech and Audio Processing*, v. 3, p. 141-45, 1995.

5 JIN, H. et al. *Affective Disorders*, v. 94, p. 269-75, 2006.

6 *Journal of American College Health*, v. 50, p. 203-17, 2002.

7 KORAN et al. *Amer. J. Psychiatry*, v. 163, p. 1806, 2006.

8 DIFRANZA, J. et al. *Archives of Pediatric and Adolescent Medicine*, v. 156, p. 397-403, 2002.

9 BAKER, D., STEPHENSON, L. *Journal of Clinical Psychology*, v. 51, p. 694, 1995.

10 MORGAN, J., SAFFRAN, J. *Child Development*, v. 66, p. 911-36, 1995.

11 KERN, J., HASTINGS, T. *J. Clinical Psychology*, v. 51, p. 499, 1995.

12 SILVERSTEIN, M. *Demography*, v. 32, p. 35, 1995.

13 COLOMBOK, S., TASKER, F. *Development Psychology*, v. 32, p. 3-11, 1996.

14 JOHNSON, J. G. et al. *Science*, v. 295, p. 2468-71, 2002.

15 ZIMMERMAN, M. et al. *Child Development*, v. 66, p. 1598-613, 1995.

16 DIENER, E. et al. *Journal of Personality and Social Psychology*, v. 69, p. 120-9, 1995.