ser visto, somente os dos Trajetos 1 e 3 e os dos Trajetos 2 e 3 excedem 2,50 e são, portanto, significantes. Para resumir toda essa informação, traçamos uma reta por baixo de todos conjuntos de tratamentos para os quais as diferenças entre as médias não são significantes, obtendo, assim, para o nosso exemplo

Trajeto 1	Trajeto 2	Trajeto 3	Trajeto 4
25,8	26,8	28,2	30,0

Como estamos interessados em minimizar o tempo dirigindo para o trabalho, isso nos diz que os Trajetos 1, 2 e 4, *como um grupo*, são preferíveis ao Trajeto 3, e que os Trajetos 3 e 4 *como um grupo* são menos indicados que os outros dois. Para ir mais adiante, podemos precisar considerar outros fatores; talvez a beleza do cenário ao longo do caminho.

EXERCÍCIOS

15.6 Realiza-se um experimento para determinar qual dentre três marcas de bola de golfe, *A*, *B* ou *C*, atinge maior distância ao ser lançada. Critique o experimento se:

(a) um golfista profissional joga todas as bolas de marca *A*, um outro joga todas as da marca *B* e um terceiro joga todas as bolas da marca *C*;

(b) todas as bolas da marca *A* são jogadas primeiro, as da marca *B* em seguida e, por último, as da marca *C*.

15.7 Uma botânica deseja comparar três gêneros de bulbos de tulipa de flores vermelhas, brancas e amarelas, respectivamente. Ela dispõe de quatro bulbos de cada tipo e planta-os num canteiro com a seguinte disposição, onde V, B e A denotam as três cores:

V	V	V	V
B	B	B	B
A	A	A	A

Quando as plantas alcançam a maturidade, ela mede sua altura e faz uma análise de variância. Critique esse experimento e indique como poderia ser melhorado.

15.8 Para comparar três dietas de emagrecimento, cinco dentre 15 pessoas são designadas aleatoriamente para cada dieta. Após duas semanas seguindo a dieta, faz-se uma análise de variância de um critério em suas perdas de peso, a fim de testar a hipótese nula de que as três dietas são igualmente eficazes. Alegou-se que esse processo não pode oferecer uma conclusão válida, porque as cinco pessoas que inicialmente pesavam mais poderiam receber a mesma dieta. Verifique que a probabilidade de isso ocorrer por acaso é da ordem de 0,001.

15.9 Com referência ao exercício precedente, suponha que cinco das 15 pessoas sejam designadas aleatoriamente para cada uma das três dietas, descobrindo-se, posteriormente, que as cinco pessoas que inicialmente pesavam mais receberam todas a mesma dieta. Ainda caberia fazer uma análise de variância de um critério?

15.10 Refaça a parte (b) do Exercício 15.1 aplicando uma análise de variância, usando as fórmulas de cálculo para obter as necessárias somas de quadrados. Compare os valores de *F* assim obtidos com os da parte (a) do Exercício 15.1

15.11 Use um aplicativo computacional apropriado ou uma calculadora gráfica para refazer o Exercício 15.1.

15.12 Refaça a parte (b) do Exercício 15.4 aplicando uma análise de variância, usando as fórmulas de cálculo para obter as necessárias somas de quadrados. Compare os valores de *F* assim obtidos com os da parte (a) do Exercício 15.4.

15.13 Use um aplicativo computacional apropriado ou uma calculadora gráfica para refazer o Exercício 15.4.

15.14 Os números de erros cometidos em cinco ocasiões por quatro digitadores, ao digitarem um relatório técnico são os seguintes:

Digitador 1:	10	13	9	11	12
Digitador 2:	11	13	8	16	12
Digitador 3:	10	15	13	11	15
Digitador 4:	15	7	11	12	9

Supondo que as suposições necessárias possam ser atendidas, faça uma análise de variância e decida, ao nível 0,05 de significância, se as diferenças entre as quatro médias amostrais podem ser atribuídas ao acaso.

15.15 Os dados abaixo mostram os rendimentos (em *bushels* por hectare) de sementes de soja plantadas em fileiras a cada 5 cm, em lotes essencialmente análogos, com as fileiras afastadas 50, 60, 70 e 80 cm uma da outra:

50 cm	60 cm	70 cm	80 cm
23,1	21,7	21,9	19,8
22,8	23,0	21,3	20,4
23,2	22,4	21,6	19,3
23,4	21,1	20,2	18,5
23,6	21,9	21,6	19,1
21,7	23,4	23,8	21,9

Supondo que esses dados constituam amostras aleatórias de quatro populações normais com o mesmo desvio-padrão, faça uma análise de variância para testar, ao nível 0,01 de significância, se as diferenças entre as quatro médias amostrais podem ser atribuídas ao acaso.

15.16 Use um aplicativo computacional apropriado ou uma calculadora gráfica para refazer o Exercício 15.15.

15.17 Uma grande firma de propaganda utiliza muitas máquinas reprográficas, várias de cada um de quatro modelos diferentes. Durante os últimos seis meses, o chefe do escritório registrou, para cada máquina, o número médio de minutos por semana em que esteve parada devido a reparos, resultando os seguintes dados:

Modelo G:	56	61	68	42	82	70	
Modelo H:	74	77	92	63	54		
Modelo K:	25	36	29	56	44	48	38
Modelo M:	78	105	89	112	61		

Considerando que as suposições necessárias podem ser atendidas, faça uma análise de variância para decidir se as diferenças entre as médias das quatro amostras podem ser atribuídas ao acaso. Use $\alpha = 0{,}01$. (*Sugestão*: use que os totais das quatro amostras são 379, 360, 276 e 445, que o total geral é 1.460, e que $\sum\sum x^2 = 104.500$.)

15.18 Usadas com três lubrificantes diferentes, certo grupo de peças de máquina acusa as seguintes perdas de peso (em miligramas) causados por atrito:

Lubrificante X:	10	13	12	10	14	8	12	13			
Lubrificante Y:	9	8	12	9	8	11	7	6	8	11	9
Lubrificante Z:	6	7	7	5	9	8	4	10			

Supondo que esses dados constituam amostras aleatórias de três populações normais com o mesmo desvio-padrão, faça uma análise de variância para decidir se as diferenças entre as três médias amostrais podem ser atribuídas ao acaso. Use o nível 0,01 de significância.

15.19 Para estudar o desempenho de um novo projeto de barco a motor, foi marcado o tempo que o barco levou para percorrer um certo trajeto do mar sob várias condições de vento e água. Supondo que possamos atender as condições necessárias, use os dados seguintes (em minutos) para testar, ao nível 0,05 de significância, se as diferenças entre as três médias amostrais são significantes:

Mar calmo:	26	19	16	22	
Mar moderado:	25	27	25	20	18
Mar agitado:	23	25	28	31	26

15.20 Use um aplicativo computacional apropriado ou uma calculadora gráfica para refazer o
(a) Exercício 15.18;
(b) Exercício 15.19.

15.21 Os valores a seguir são as percentagens da safra do ano anterior para macieiras sujeitas a oito diferentes esquemas de pulverização.

Esquema						
A	130	98	128	106	139	121
B	142	133	122	131	132	141
C	114	141	95	123	118	140
D	77	99	84	76	70	75
E	109	86	113	101	103	112
F	148	143	111	142	131	100
G	149	129	134	108	119	126
H	92	129	111	103	107	125

Supondo que as condições necessárias possam ser atendidas, use um aplicativo computacional apropriado para conduzir uma análise variância com $\alpha = 0{,}05$.

15.22 No Exemplo 15.1, fizemos uma análise de variância para os dados exibidos à página 363, em que as médias para as três localidades ao longo do rio Mississipi foram 40,5, 39,0 e 31,50. Use o método do intervalo estudentizado para fazer um teste de comparações múltiplas ao nível 0,01 de significância e discuta os resultados supondo que a baixa contaminação por cálcio seja desejável.

15.23 Como uma continuação do Exercício 15.15, use o método do intervalo estudentizado para fazer um teste de comparação múltipla ao nível 0,01 de significância e interprete os resultados.

15.24 Como uma continuação do Exercício 15.21, use o método do intervalo estudentizado para fazer um teste de comparação múltipla ao nível 0,05 de significância e interprete os resultados.

15.25 Uma análise de variância e um teste de comparações múltiplas subseqüente do desempenho de quatro agentes imobiliários deram o seguinte resultado:

Bruno	Júlia	Nestor	Susana

onde Susana teve a média de vendas mais alta. Interprete os resultados.

15.26 Uma análise de variância e um teste de comparações múltiplas subseqüente do conteúdo de gordura de cinco refeições congeladas deram o seguinte resultado:

A C B F D E

onde A teve o maior conteúdo de gordura e E o menor. Interprete esses resultados, sabendo que as cinco refeições constam de uma lista de recomendações para dietas de baixa gordura.

15.27 Com referência à nota de rodapé à página 363, verifique que é igual a zero a soma α dos efeitos de tratamentos.

15.28 Verifique simbolicamente que para uma análise de variância de um critério

(a) $\dfrac{SQ(Tr)}{k-1} = n \cdot s^2_{\bar{x}}$;

(b) $\dfrac{SQE}{k(n-1)} = \dfrac{1}{k} \cdot \sum_{i=1}^{k} s_i^2$, onde s_i^2 é a variância da i-ésima amostra.

15.5 PLANEJAMENTO DE EXPERIMENTOS: BLOQUEAMENTO

Para introduzir um outro conceito importante no planejamento de experimentos, suponhamos que seja aplicado um teste de compreensão de leitura a amostras aleatórias de alunos da oitava série de quatro escolas, com os resultados seguintes:

Escola A:	87	70	92
Escola B:	43	75	56
Escola C:	70	66	50
Escola D:	67	85	79

As médias dessas quatro amostras são 83, 58, 62 e 77 e, como as diferenças entre elas são muito grandes, poderia parecer razoável concluir que haja algumas diferenças reais entre os graus de compreensão de leitura dos alunos da oitava série das quatro escolas. Não é o que decorre, entretanto, de uma análise de variância de um critério. Temos:

Fonte de variação	*Graus de liberdade*	*Soma de quadrados*	*Quadrado médio*	*F*
Tratamentos	3	1.278	426	2,90
Erro	8	1.176	147	
Total	11	2.454		

e como $F = 2{,}90$ é menor do que 4,07, o valor de $F_{0,05}$ para 3 e 8 graus de liberdade, a hipótese nula (de que as médias populacionais são todas iguais) não pode ser rejeitada ao nível 0,05 de significância.

A razão disso é que há não só consideráveis diferenças entre as quatro médias, mas também diferenças muito grandes entre os valores dentro das amostras. Na primeira amostra eles vão de 70 a 92, na segunda amostra vão de 43 a 75, na terceira vão de 50 a 70 e na quarta amostra vão de 67 a 85. Pensando um pouco sobre a questão, pareceria razoável concluir que essas diferenças dentro das amostras possam ser causadas por diferenças de capacidade, um fator irrelevante (que poderíamos considerar um fator de "incomodação") que foi aleatorizado ao tomarmos uma amostra aleatória de alunos de oitava série de cada escola. Assim, as variações causadas por diferenças de capacidade foram incluídas no erro experimental; isso "inflacionou" a soma de quadrados de erros que figura no denominador da estatística F, e os resultados não foram significantes.

Para evitar tal situação, poderíamos manter fixo o fator irrelevante, mas isso raramente nos dará a informação desejada. Em nosso exemplo, poderíamos limitar o estudo a alunos da oitava série com nota média (NM) de 90 ou mais, mas então os resultados se aplicariam apenas a alunos da oitava série com NM de 90 ou mais. Uma outra possibilidade seria fazer a fonte conhecida de variabilidade (o fator irrelevante) variar deliberadamente sobre um intervalo tão amplo quanto necessário, fazendo-o de forma que a variabilidade causada possa ser medida e, assim, eliminada do erro experimental. Isso significa que devemos planejar o experimento de tal forma que possamos fazer uma **análise de variância de dois critérios**, em que a variabilidade total dos dados seja dividida em três componentes atribuídos, respectivamente, aos tratamentos (em nosso exemplo, as quatro escolas), ao fator irrelevante e ao erro experimental.

Como veremos mais adiante, isso pode ser conseguido, em nosso exemplo, selecionando aleatoriamente, em cada escola, um aluno de oitava série com NM baixa, um aluno com NM típica e um aluno com NM alta, supondo que "baixa", "típica" e "alta" sejam definidas de maneira rigorosa. Suponha, então, que procedamos dessa maneira, obtendo os resultados mostrados na seguinte tabela:

	NM Baixa	**NM Típica**	**NM Alta**
Escola A	71	92	89
Escola B	44	51	85
Escola C	50	64	72
Escola D	67	81	86

O que fizemos aqui é denominado **bloqueamento**, e os três níveis de NM são denominados **blocos**. Em geral, os blocos são os níveis em que mantemos fixo um fator irrelevante, de modo que possamos medir sua contribuição para a variabilidade total dos dados por meio de uma análise de variância de dois critérios. No esquema escolhido para nosso exemplo, estamos trabalhando com **blocos completos**, que são completos no sentido de que cada tratamento figura o mesmo número de vezes em cada bloco. Há um aluno de oitava série de cada escola em cada bloco.

Suponha, além disso, que a ordem na qual os estudantes são testados possa ter algum efeito sobre os resultados. Se a ordem é aleatorizada dentro de cada bloco (isto é, para cada nível de NM), esse esquema é denominado **planejamento em bloco aleatorizado**.

15.6 ANÁLISE DE VARIÂNCIA DE DOIS CRITÉRIOS

A análise de experimentos em que utilizamos bloqueamento para reduzir a soma de quadrados dos erros requer uma **análise de variância de dois critérios**. Nesse tipo de análise, referimo-nos às duas variáveis sob consideração como "tratamentos" e "blocos", embora esse tipo de análise também seja aplicado a **experimentos de dois fatores**, em que ambas as variáveis têm interesse material.

Antes de entrar em detalhes, frisamos que há essencialmente duas maneiras de analisar tais experimentos de dois fatores, conforme as variáveis sejam independentes ou apresentem uma **interação**. Suponha que um fabricante de pneus esteja experimentando diferentes tipos de banda de rodagem e constate que uma delas se adapta especialmente bem a estradas de terra, enquanto que outra é especialmente adequada a estradas pavimentadas. Nesse caso, dizemos que há uma interação entre as condições da estrada e os tipos de banda de rodagem. Por outro lado, se cada uma das bandas de rodagem é afetada igualmente pelas condições da estrada, diríamos que não há interação e que as duas variáveis (condições da estrada e banda de rodagem) são independentes. Esse último caso será começado a considerar na próxima seção, e um método que também seja conveniente para testar para interações será descrito na Seção 15.9.

15.7 ANÁLISE DE VARIÂNCIA DE DOIS CRITÉRIOS SEM INTERAÇÃO

Para formular as hipóteses a serem testadas no caso de duas variáveis, escrevemos μ_{ij} para representar a média populacional que corresponde ao i-ésimo tratamento e ao j-ésimo bloco. Em nosso exemplo anterior, μ_{ij} é a nota média da compreensão de leitura na i-ésima escola para alunos de oitava série com nível j de nota média. Expressamos isso por

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

Como no rodapé da página 363, μ é a média global (a média de todas as médias populacionais μ_{ij}) e os α_i são os efeitos de tratamento (cuja soma é zero). Correspondentemente, denominamos os β_j de **efeitos de bloco** (cuja soma também é zero) e escrevemos as duas hipóteses nulas que desejamos testar como

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \quad \text{e} \quad \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

A alternativa para a primeira hipótese nula (que, em nossa exemplificação, equivale à hipótese de que o grau médio de compreensão de leitura dos alunos da oitava série é o mesmo nas quatro escolas) é que os efeitos de tratamento α_i não são todos iguais a zero; a alternativa da segunda hipótese nula (que, em nossa exemplificação, equivale à hipótese de que a compreensão média de leitura dos alunos da oitava série é a mesma para todos os três níveis de NM) é que os efeitos de bloco β_j não são todos nulos.

Para testar a segunda hipótese nula, precisamos de uma grandeza, análoga à soma de quadrados de tratamentos, que meça a variação entre as médias de blocos (58, 72 e 83 para os dados à página 380). Assim, denotando por $T_{\cdot j}$ o total de todos os valores do j-ésimo bloco, substituímos esse valor no lugar de $T_{\cdot j}$ na fórmula de cálculo de $SQ(Tr)$ à página 371, somamos em relação a j no lugar de em relação a i, e permutamos n e k, obtendo, analogamente a $SQ(Tr)$, a **soma de quadrados de blocos**

FÓRMULA DE CÁLCULO PARA SOMA DE QUADRADOS DE BLOCOS

$$SQB = \frac{1}{k} \cdot \sum_{j=1}^{n} T_{\cdot j}^2 - \frac{1}{kn} \cdot T_{\cdot\cdot}^2$$

Numa análise de variância de dois critérios (sem interação), calculamos STQ e $SQ(Tr)$ de acordo com as fórmulas à página 371, SQB de acordo com a fórmula imediatamente acima, e então obtemos SQE por subtração. Como

$$STQ = SQ(Tr) + SQB + SQE$$

temos

SOMA DE QUADRADOS DE ERROS (ANÁLISE DE VARIÂNCIA DE DOIS CRITÉRIOS)

$$SQE = STQ - [SQ(Tr) + SQB]$$

Note que a soma de quadrados de erros para uma análise de variância de dois critérios não é igual à soma de quadrados de erros para a análise de variância de um critério feita sobre os mesmos dados, embora ambas sejam denotadas pelo mesmo símbolo SQE. De fato, estamos agora dividindo a soma de quadrados de erros para a análise de variância de um critério em dois termos: a soma de quadrados de blocos, SQB, e o resto, que é a nova soma de quadrados de erros, SQE.

Podemos agora construir a seguinte tabela de análise de variância para uma análise de variância de dois critérios (sem interação):

Fonte de variação	*Graus de liberdade*	*Soma de quadrados*	*Quadrado médio*	*F*
Tratamentos	$k-1$	$SQ(Tr)$	$QM(Tr)=\dfrac{SQ(Tr)}{k-1}$	$\dfrac{QM(Tr)}{QME}$
Blocos	$n-1$	SQB	$QMB=\dfrac{SQB}{n-1}$	$\dfrac{QMB}{QME}$
Erro	$(k-1)(n-1)$	SQE	$QME=\dfrac{SQE}{(k-1)(n-1)}$	
Total	$kn-1$	STQ		

Os quadrados médios são novamente as somas de quadrados divididas por seus respectivos graus de liberdade, e os dois valores F são os quadrados médios para tratamentos e para blocos divididos pelo quadrado médio para o erro. Igualmente, o número de graus de liberdade para blocos é $n-1$ (como no caso de tratamentos, com n substituído por k), e o número de graus de liberdade para o erro é obtido subtraindo os graus de liberdade para tratamentos e blocos de $kn-1$, o número total de graus de liberdade:

$$\begin{aligned}(kn-1)-(k-1)-(n-1) &= kn-k-n+1\\ &= (k-1)(n-1)\end{aligned}$$

Assim, no teste de significância para tratamentos, os graus de liberdade do numerador e do denominador de F são $k-1$ e $(k-1)(n-1)$ e, no teste de significância para blocos, os graus de liberdade do numerador e do denominador de F são $n-1$ e $(k-1)(n-1)$.

EXEMPLO 15.4 No exemplo que utilizamos para ilustrar a necessidade dos blocos, fornecemos os seguintes dados para comparar as notas de alunos da oitava série de quatro escolas num teste de compreensão de leitura utilizando, para isso, os blocos das médias baixa, típica e alta:

	NM Baixa	**NM Típica**	**NM Alta**
Escola A	71	92	89
Escola B	44	51	85
Escola C	50	64	72
Escola D	67	81	86

Supondo que os dados consistam em amostras aleatórias independentes de populações normais, todas com o mesmo desvio-padrão teste, ao nível 0,05 de significância, se as diferenças entre as médias obtidas para as quatro escolas (tratamentos) são significantes, e também se as diferenças entre as médias obtidas para os três níveis de NM (blocos) são significantes.

Solução **1.** H_0: $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0$

$\beta_1=\beta_2=\beta_3=0$

H_A: os efeitos de tratamento não são todos iguais a zero; os efeitos de bloco não são todos iguais a zero.

2. $\alpha = 0{,}05$ para ambos os testes.

3. Para os tratamentos, rejeitar a hipótese nula se $F \geq 4{,}76$, onde F deve ser determinado por uma análise de variância de dois critérios e 4,76 é o valor de $F_{0,05}$ para $k - 1 = 4 - 1 = 3$ e $(k-1)(n-1) = (4-1)(3-1) = 6$ graus de liberdade. Para os blocos, rejeitar a hipótese nula se $F \geq 5{,}14$, onde F deve ser determinado por uma análise de variância de dois critérios e 5,14 é o valor de $F_{0,05}$ para $n - 1 = 3 - 1 = 2$ e $(k-1)(n-1) = (4-1)(3-1) = 6$ graus de liberdade. Se alguma das duas hipóteses nulas não pode ser rejeitada, devemos aceitá-la ou reservar julgamento.

4. Substituindo $k = 4$, $n = 3$, $T_{1\cdot} = 252$, $T_{2\cdot} = 180$, $T_{3\cdot} = 186$, $T_{4\cdot} = 234$, $T_{\cdot 1} = 232$, $T_{\cdot 2} = 288$, $T_{\cdot 3} = 332$, $T_{\cdot\cdot} = 852$ e $\sum\sum x^2 = 63.414$ nas fórmulas de cálculo para as somas de quadrados, obtemos

$$\begin{aligned} STQ &= 63.414 - \frac{1}{12}(852)^2 \\ &= 63.414 - 60.492 \\ &= 2.922 \\ SQ(Tr) &= \frac{1}{3}(252^2 + 180^2 + 186^2 + 234^2) - 60.492 \\ &= 1.260 \\ SSB &= \frac{1}{4}(232^2 + 288^2 + 332^2) - 60.492 \\ &= 1.256 \end{aligned}$$

e

$$\begin{aligned} SQE &= 2.922 - (1.260 + 1.256) \\ &= 406 \end{aligned}$$

Como os graus de liberdade são $k - 1 = 4 - 1 = 3$, $n - 1 = 3 - 1 = 2$, $(k-1)(n-1) = (4-1)(3-1) = 6$ e $kn - 1 = 4 \cdot 3 - 1 = 11$, obtemos, então, $QM(Tr) = \frac{1.260}{3} = 420$, $QMB = \frac{1.256}{2} = 628$, $QME = \frac{406}{6} \approx 67{,}67$, $F \approx \frac{420}{67,67} \approx 6{,}21$ para tratamentos e $F \approx \frac{628}{67,67} \approx 9{,}28$ para blocos. Todos esses resultados estão resumidos na seguinte tabela de análise de variância:

Fonte de variação	*Graus de liberdade*	*Soma de quadrados*	*Quadrado médio*	*F*
Tratamentos	3	1.260	420	6,21
Blocos	2	1.256	628	9,28
Erro	6	406	67,67	
Total	11	2.922		

5. Como $F = 6{,}21$ excede 4,76, a hipótese nula para os tratamentos deve ser rejeitada; e como $F = 9{,}28$ excede 5,14, a hipótese nula para os blocos deve ser rejeitada. Em outras palavras, concluímos que o grau médio de compreensão de leitura de alunos da oitava série não é o mesmo para as quatro escolas, e que o grau médio de compreensão de leitura de alunos da

oitava série não é o mesmo para os três níveis de NM. Observe que, com bloqueamento, obtivemos diferenças significantes entre os graus médios de compreensão de leitura de alunos da oitava série nas quatro escolas, o que não ocorreu sem o bloqueamento.

Quando conferimos os cálculos na solução do Exemplo 15.4, verificamos que a versão mais recente de MINITAB utiliza a abordagem do valor *p*. Como pode ser visto na Figura 15.4, as colunas encabeçadas por DF (graus de liberdade, em inglês), SS (somas de quadrados, em inglês), MS (quadrados médios, em inglês) e *F* são as mesmas do que antes; mas a Figura 15.4 também mostra uma coluna de valores *p*, que levam diretamente a decisões sobre as hipóteses. Para tratamentos, o valor *p* é 0,029 e, como 0,029 é menor do que 0,05, a hipótese nula para tratamentos deve ser rejeitada. Para blocos, o valor *p* é 0,015 e, como 0,015 é menor do que 0,05, a hipótese nula para blocos deve ser rejeitada.

Conforme salientado anteriormente, também se pode aplicar a análise de variância de dois critérios à análise de um experimento de dois fatores, em que ambas as variáveis (fatores) têm importância material. Poderia ser utilizada, por exemplo, na análise dos dados seguintes, coletados num experimento projetado para testar se o alcance de vôo de um míssil (em quilômetros) é afetado pelas diferenças entre os lançadores e também pelas diferenças entre os tipos de combustível.

	Combustível 1	*Combustível 2*	*Combustível 3*	*Combustível 4*
Lançador X	45,9	57,6	52,2	41,7
Lançador Y	46,0	51,0	50,1	38,8
Lançador Z	45,7	56,9	55,3	48,1

Note que utilizamos um formato diferente para essa tabela, para distinguir entre experimentos de dois fatores com fatores irrelevantes aleatorizados ao longo de todo o experimento, e experimentos com fatores irrelevantes aleatorizados separadamente ao longo de cada bloco.

Também, quando se aplica uma análise de variância de dois critérios dessa maneira, costumamos denominar as duas variáveis **fator A** e **fator B** (em vez de tratamentos e blocos) e escrevemos *SQA* e *QMA* em vez de *SQ*(*Tr*) e *QM*(*Tr*); continuamos escrevendo *SBQ* e *QMB*, mas agora o B representa o fator B, em vez de blocos.

15.8 PLANEJAMENTO DE EXPERIMENTOS: REPLICAÇÃO

Na Seção 15.5, mostramos como é possível aumentar a quantidade de informação a ser obtida de um experimento por bloqueamento, isto é, pela eliminação do efeito de um fator irrelevante. Outra maneira de aumentar a informação a ser obtida de um experimento consiste em aumentar o volume dos dados. Assim é que, no exemplo da página 379, poderíamos aumentar o tamanho das amostras e aplicar o teste de compreensão de leitura a vinte alunos da oitava série de cada esco-

Análise de Variância de Dois Critérios: C3 contra C1, C2

```
Analysis of Variance for C3
Source     DF       SS       MS       F       P
C1          3   1260.0    420.0    6.21   0.029
C2          2   1256.0    628.0    9.28   0.015
Error       6    406.0     67.7
Total      11   2922.0
```

Figura 15.4 Impresso de computador para o Exemplo 15.4.

la, em lugar de três. Para planejamentos mais complicados, chega-se ao mesmo objetivo realizando o experimento todo mais de uma vez, o que é denominado **replicação**. Quanto ao exemplo da página 379, poderíamos fazer o experimento (selecionar e testar doze alunos da oitava série) numa semana e então replicar (repetir) todo o experimento na semana seguinte.

A replicação não apresenta dificuldades conceituais, mas apresenta dificuldades computacionais, que mencionamos aqui somente porque é requerido pelo nosso estudo na Seção 15.9. Além disso, se replicamos um experimento que exige uma análise de variância de dois critérios, poderá haver a necessidade de uma análise de variância de três critérios, pois a própria replicação poderá constituir uma fonte de variação dos dados. Isso poderia ocorrer em nosso exemplo relativo às notas nos testes de compreensão de leitura, digamos, se o clima na segunda semana fosse muito quente e úmido, dificultando a concentração dos estudantes.

15.9 ANÁLISE DE VARIÂNCIA DE DOIS CRITÉRIOS COM INTERAÇÃO

Quando o conceito de interação foi mencionado pela primeira vez, foi descrito um experimento em que um fabricante de pneus descobre que um tipo de banda de rodagem é especialmente bom para estradas de terra, enquanto um outro tipo é especialmente bom para estradas pavimentadas. Uma situação semelhante surge quando um fazendeiro descobre que uma variedade de milho se adapta melhor com um tipo de fertilizante enquanto uma outra variedade se adapta melhor a um outro fertilizante; ou quando é observado que uma pessoa comete menos erros com um tipo de processador de texto enquanto uma outra pessoa comete menos erros com outro tipo de processador de texto.

Para considerar um exemplo numérico, voltemos ao experimento de dois fatores à página 384, o que tratou do efeito de três lançadores de mísseis e quatro tipos de combustível sobre o alcance de vôo de certos mísseis. Analisando esses dados pelo método da Seção 15.7, repartimos *STQ*, que é uma medida da variação total entre os dados, em três componentes atribuídos, respectivamente, aos lançadores diferentes, aos combustíveis diferentes e ao erro (ou acaso). Se existirem interações, o que é bem possível, essas variações causadas estariam ocultas por que estariam incluídas como parte de *SQE*, que é a soma de quadrados de erros. Para isolar uma soma de quadrados que possa ser atribuída à interação, precisamos de uma outra maneira de medir a variação devida ao acaso, e faremos isso repetindo todo o experimento. Suponhamos, então, que assim, obtenhamos os dados mostrados na tabela seguinte:

	Combustível 1	*Combustível 2*	*Combustível 3*	*Combustível 4*
Lançador X	46,1	55,9	52,6	44,3
Lançador Y	46,3	52,1	51,4	39,6
Lançador Z	45,8	57,9	56,2	47,6

que denominamos Réplica 2 para distingui-la dos dados à página 384, que agora passamos a denominar Réplica 1. Combinando as duas réplicas numa tabela, obtemos

	Combustível 1	*Combustível 2*	*Combustível 3*	*Combustível 4*
Lançador X	45,9, 46,1	57,6, 55,9	52,2, 52,6	41,7, 44,3
Lançador Y	46,0, 46,3	51,0, 52,1	50,1, 51,4	38,8, 39,6
Lançador Z	45,7, 45,8	56,9, 57,9	55,3, 56,2	48,1, 47,6

onde o primeiro valor em cada célula provém da Réplica 1 e o segundo valor provém da Réplica 2.

Agora podemos representar as variações devidas ao acaso pela variação *dentro* das 12 células da tabela e, em geral, a nova soma de quadrados de erros passa a ser

$$SQE = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{h=1}^{r}(x_{ijh} - \bar{x}_{ij\cdot})^2$$

onde x_{ijh} é o valor correspondente ao i-ésimo tratamento, o j-ésimo bloco e a h-ésima réplica, e $\bar{x}_{ij\cdot}$ é a média dos valores das células correspondentes ao i-ésimo tratamento e ao j-ésimo bloco.

Substituindo os dois valores de cada célula da tabela imediatamente precedente pela sua média, obtemos

	Combustível 1	*Combustível 2*	*Combustível 3*	*Combustível 4*
Lançador X	46	56,75	52,4	43
Lançador Y	46,15	51,55	50,75	39,2
Lançador Z	45,75	57,40	55,75	47,85

e é assim que nossos dados ficariam depois de removidas as variações devidas ao acaso. Em outras palavras, a única variação que persiste é devida aos tratamentos, aos blocos e à interação, e se fizéssemos uma análise de variância de dois critérios como na Seção 15.7, obteríamos correspondentes tratamentos, blocos e **somas de quadrados de interação**, onde essas últimas seriam o que era a soma de quadrados de erros. Na verdade, tudo isso é feito só conceitualmente. Se realmente fizéssemos uma análise de variância de dois critérios com as médias no lugar dos dois valores em cada célula, veríamos que cada uma das somas de quadrados foi dividida por um fator de 2. Do mesmo modo, se tivéssemos r réplicas, cada soma de quadrados teria sido dividida por um fator de r.

Como nas Seções 15.3 e 15.7, existem fórmulas de calcular as várias somas de quadrados de uma análise de variância de dois critérios com interações. Contudo, como as contas necessárias são, para dizer o mínimo, de difícil manuseio, costuma-se fazer essas contas com o auxílio de um computador. Isso é precisamente o que faremos aqui, obtendo os diversos graus de liberdade, somas de quadrados, quadrados médios, valores de F, e valores p do impresso de MINITAB mostrado na Figura 15.5.

Na verdade, tudo de que precisamos são os valores p dados no impresso como 0,000, o que significa 0,000 arredondado até a terceira casa decimal. Como esses valores são todos menores do que 0,05, as hipóteses nulas para lançadores, combustíveis e interações lançadores/combustíveis devem ser todas rejeitadas. (Os valores p reais, obtidos por meio de uma calculadora HP STAT/MATH, são 0,00000023, 0,000000000017 e 0,0001.)

Análise de Variância de Dois Critérios: C3 contra C1, C2

```
Analysis of Variance for C3
Source          DF        SS        MS        F        P
C1               2    91.503    45.752    70.61    0.000
C2               3   570.825   190.275   293.67    0.000
Interaction      6    50.937     8.489    13.10    0.000
Error           12     7.775     0.648
Total           23   721.040
```

Figura 15.5 Impresso de computador para uma análise de variância de dois critérios com interação.

EXERCÍCIOS

15.29 Para comparar os intervalos de tempo que três canais de televisão destinam a comerciais, uma pesquisadora mede o tempo dedicado a comerciais em amostras aleatórias de 15 programas em cada canal. Para sua surpresa, ela constata que há tanta variação dentro das amostras – para um canal, os números variam de 8 a 35 minutos – que é praticamente impossível obter resultados significantes. Há alguma forma que permita à pesquisadora superar esse obstáculo?

15.30 Para comparar cinco processadores de palavras, *A*, *B*, *C*, *D* e *E*, quatro pessoas, 1, 2, 3 e 4, foram cronometradas preparando um certo relatório em cada uma das máquinas. Os resultados (em minutos) constam na seguinte tabela:

	1	*2*	*3*	*4*
A	49,1	48,2	52,3	57,0
B	47,5	40,9	44,6	49,5
C	76,2	46,8	50,1	55,3
D	50,7	43,4	47,0	52,6
E	55,8	48,3	82,6	57,8

Explique por que esses dados não deveriam ser analisados pelo método da Seção 15.7.

15.31 Os conteúdos de colesterol (em miligramas por pacote) obtidos por quatro laboratórios para pacotes de 150 gramas de três alimentos dietéticos muito semelhantes são os seguintes:

	Laboratório 1	*Laboratório 2*	*Laboratório 3*	*Laboratório 4*
Alimento A	3,7	2,8	3,1	3,4
Alimento B	3,1	2,6	2,7	3,0
Alimento C	3,5	3,4	3,0	3,3

Faça uma análise de variância de dois critérios, utilizando o nível de significância 0,01 para ambos os testes.

15.32 Quatro testes de conhecimento em ciências, diferentes mas supostamente equivalentes, foram dados a cada um de cinco estudantes, que obtiveram as seguintes notas:

	Estudante C	*Estudante D*	*Estudante E*	*Estudante F*	*Estudante G*
Teste 1	77	62	52	66	68
Teste 2	85	63	49	65	76
Teste 3	81	65	46	64	79
Teste 4	88	72	55	60	66

Faça uma análise de variância de dois critérios ao nível 0,01 de significância para ambos os testes.

15.33 Um técnico de laboratório mediu a resistência à ruptura de cinco tipos de fio de linho, utilizando quatro instrumentos diferentes de medida I_1, I_2, I_3 e I_4 e obteve os seguintes resultados (em gramas):

	I_1	I_2	I_3	I_4
Fio 1	20,9	20,4	19,9	21,9
Fio 2	25,0	26,2	27,0	24,8
Fio 3	25,5	23,1	21,5	24,4
Fio 4	24,8	21,2	23,5	25,7
Fio 5	19,6	21,2	22,1	21,1

Faça uma análise de variância de dois critérios, utilizando o nível de significância 0,05 para ambos os testes.

15.34 Os números de peças defeituosas produzidas por quatro operários trabalhando, em turnos, em três máquinas diferentes são os seguintes:

		Operário			
		B_1	B_2	B_3	B_4
Máquina	A_1	35	38	41	32
	A_2	31	40	38	31
	A_3	36	35	43	25

Faça uma análise de variância de dois critérios, utilizando o nível 0,05 de significância para ambos os testes.

15.35 Num experimento planejado para avaliar três detergentes, um laboratório lavou roupa três vezes em cada combinação de detergente e temperatura de água, obtendo os seguintes registros de limpeza da roupa:

	Detergente A	*Detergente B*	*Detergente C*
Água fria	45, 39, 46	43, 46, 41	55, 48, 53
Água morna	37, 32, 43	40, 37, 46	56, 51, 53
Água quente	42, 42, 46	44, 45, 38	46, 49, 42

Use o nível 0,01 de significância para testar para diferenças entre os detergentes, diferenças devidas à temperatura da água e diferenças devidas a interações.

15.36 Um serviço de testes de produtos de consumo quer comparar a qualidade de 24 bolos assados em sua cozinha com cada uma de quatro misturas diferentes, preparadas de acordo com três receitas diferentes (em que variam as quantidades de ingredientes frescos adicionados), sendo elaborados uma vez pelo Chefe X e outra vez pelo Chefe Y. Pede-se a um provador para dar uma nota de 1 a 100 para cada bolo, obtendo os seguintes resultados, onde em cada caso o primeiro número se refere ao bolo assado pelo Chefe X e o segundo número ao bolo assado pelo Chefe Y:

	Mistura A	*Mistura B*	*Mistura C*	*Mistura D*
Receita 1	66, 62	70, 68	74, 68	73, 67
Receita 2	68, 61	71, 73	74, 70	66, 61
Receita 3	75, 68	69, 71	67, 63	70, 66

Use o nível 0,05 de significância para testar para diferenças devidas às diferentes receitas, diferenças devidas às diferentes misturas e diferenças devidas a interações entre receitas e misturas.

15.10 PLANEJAMENTO DE EXPERIMENTOS: CONSIDERAÇÕES ADICIONAIS

Na Seção 15.5, vimos como o bloqueamento pode eliminar a variabilidade do erro experimental devida a um fator estranho e como, em princípio, podemos lidar da mesma maneira com duas ou mais fontes estranhas de variação. O único problema real é que isso pode aumentar o tamanho de um experimento além dos limites práticos. Suponha que, no exemplo da compreensão de leitura por alunos da oitava série, quiséssemos também eliminar qualquer variabilidade causada por diferenças de idade (12, 13 ou 14) e de sexo. Admitindo todas as combinações possíveis de NM, idade e sexo, deveremos utilizar $3 \cdot 3 \cdot 2 = 18$ blocos diferentes, e se deve haver um aluno de oitava série de cada escola em cada bloco diferente, teremos de selecionar e testar, ao todo, $18 \cdot 4 = 72$ alunos de oitava série. Se quiséssemos eliminar qualquer variabilidade que possa ser devida à origem étnica, para a qual poderíamos considerar cinco categorias, isso elevaria para $72 \cdot 5 = 360$ o número de alunos da oitava série necessários.

Nesta seção, vamos mostrar como, às vezes, podemos resolver problemas como esse, ao menos em parte, planejando os experimentos como **quadrados latinos**. Ao mesmo tempo, esperamos inculcar no leitor que é através de planejamento adequado que experimentos podem proporcionar uma riqueza de informação. Para dar um exemplo, suponha que uma organização brasileira de pesquisa de mercado deseja comparar quatro maneiras de embalar um lanche rápido, mas está preocupada com as diferenças regionais possíveis na popularidade do lanche e, também, com os efeitos de anunciar o lanche de diversas maneiras. Assim, a organização decide testar os diferentes tipos de embalagem em quatro regiões cardinais do país, no Norte, no Sul, no Leste e no Oeste do país e fazer a promoção por meio de descontos, sorteios, vale-brindes e vendas do tipo dois-por-um. Haveria, assim, $4 \cdot 4 = 16$ blocos (combinações de regiões e métodos de promoção), o que exigiria $16 \cdot 4 = 64$ áreas de mercado (cidades) para promover cada tipo de embalagem, uma vez dentro de cada bloco. Além disso, os mercados de teste deveriam estar separados uns dos outros, de modo que os métodos de promoção não interfiram entre si, e o Brasil simplesmente não têm 64 mercados de teste suficientemente separados. Contudo, é interessante observar que, com um planejamento adequado, 16 áreas de mercado (cidades) serão suficientes. A título de ilustração, consideremos o seguinte arranjo, denominado quadrado latino, em que as letras *A*, *B*, *C* e *D* representam os quatro tipos de embalagem:

	Descontos	*Sorteios*	*Vale-Brindes*	*Vendas 2 por 1*
Norte	*A*	*B*	*C*	*D*
Sul	*B*	*C*	*D*	*A*
Leste	*C*	*D*	*A*	*B*
Oeste	*D*	*A*	*B*	*C*

Em geral, um quadrado latino é um arranjo quadrado das letras *A*, *B*, *C*, *D*, . . ., do alfabeto latino, tal que cada letra ocorra uma, e só uma, vez em cada linha e em cada coluna.

O quadrado latino precedente, encarado como um planejamento experimental, exige que sejam dados descontos com a embalagem *A* numa cidade do Norte, com a embalagem *B* numa cidade do Sul, com a embalagem *C* numa cidade do Leste e com a embalagem *D* numa cidade do Oeste; que os sorteios sejam utilizados com a embalagem *B* numa cidade do Norte, com a embalagem *C* numa cidade do Sul, com a embalagem *D* numa cidade do Leste e com a emba-

lagem *A* numa cidade do Oeste, e assim por diante. Note que cada tipo de promoção é usado uma vez em cada região e uma vez com cada tipo de embalagem; cada tipo de embalagem é usado uma vez em cada região e uma vez com cada tipo de promoção; e cada região é usada uma vez com cada tipo de embalagem e uma vez com cada tipo de promoção. Como veremos, isso nos permite fazer uma análise de variância que conduz a testes de significância para todas três variáveis.

A análise de um quadrado latino $r \times c$ é muito semelhante a uma análise de variância de dois critérios. A soma de quadrados total e as somas de quadrados para linhas e colunas são calculadas da mesma maneira como foram calculadas anteriormente *STQ*, *SQ*(*Tr*) e *SQB*, mas devemos encontrar uma soma extra de quadrados que meça a variabilidade que é devida à variável representada pelas letras *A*, *B*, *C*, *D*, . . ., a saber, uma nova soma de quadrados para tratamentos. A fórmula para essa soma de quadrados é

SOMA DE QUADRADOS DE TRATAMENTOS PARA QUADRADOS LATINOS

$$SQ(Tr) = \frac{1}{r} \cdot (T_A^2 + T_B^2 + T_C^2 + \cdots) - \frac{1}{r^2} \cdot T_{..}^2$$

onde T_A é o total das observações correspondentes ao tratamento *A*, T_B é o total das observações correspondentes ao tratamento *B*, e assim por diante. Finalmente, a soma de quadrados de erros é obtida novamente por subtração:

SOMA DE QUADRADOS DE ERROS PARA QUADRADOS LATINOS

$$SQE = STQ - [SQL + SQC + SQ(Tr)]$$

onde *SQL* e *SQC* são as somas de quadrados para as linhas e para as colunas, respectivamente.

Podemos, agora, construir uma tabela de análise de variância para um quadrado latino $r \times r$. Os quadrados médios são novamente as somas de quadrados divididas pelos seus respectivos números de graus de liberdade, e os três valores *F* são os quadrados médios para linhas, colunas e tratamentos divididos pelo quadrado médio de erro. Os graus de liberdade para linhas, colunas e tratamentos são todos iguais a $r - 1$ e, por subtração, o número de graus de liberdade para o erro é

$$(r^2 - 1) - (r - 1) - (r - 1) - (r - 1) = r^2 - 3r + 2 = (r - 1)(r - 2)$$

Assim, para cada um dos três testes de significância, os graus de liberdade do numerador e do denominador de *F* são $r - 1$ e $(r - 1)(r - 2)$.

Fonte de variação	*Graus de liberdade*	*Soma de quadrados*	*Quadrado médio*	*F*
Linhas	$r-1$	SQL	$QMR = \frac{SQL}{r-1}$	$\frac{QMR}{QME}$
Colunas	$r-1$	SQC	$QMC = \frac{SQC}{r-1}$	$\frac{QMC}{QME}$
Tratamentos	$r-1$	$SQ(Tr)$	$QM(Tr) = \frac{SQ(Tr)}{r-1}$	$\frac{QM(Tr)}{QME}$
Erro	$(r-1)(r-2)$	SQE	$QME = \frac{SQE}{(r-1)(r-2)}$	
Total	r^2-1	STQ		

EXEMPLO 15.5 Suponhamos que, no exemplo do lanche rápido dado nesta seção, a organização de pesquisa de mercado obtenha os dados constantes da tabela abaixo, na qual as cifras representam as vendas de uma semana em milhares de unidades monetárias:

	Descontos	*Sorteios*	*Vale-Brindes*	*Vendas 2 por 1*
Norte	A 48	B 38	C 42	D 53
Sul	B 39	C 43	D 50	A 54
Leste	C 42	D 50	A 47	B 44
Oeste	D 46	A 48	B 46	C 52

Supondo que as suposições necessárias possam ser atendidas, analise esse quadrado latino ao nível 0,05 de significância para cada teste.

Solução

1. H_0: os efeitos de linha, coluna e tratamento (definidos como no rodapé da página 363 e na página 390) são todos iguais a zero.

H_A: os efeitos de tratamento não são todos iguais a zero.

2. $\alpha = 0{,}05$ para cada testes.

3. Para linhas, coluna ou tratamentos, rejeitar a hipótese nula se $F \geq 4{,}76$, onde os F são obtidas por meio de uma análise de variância, e 4,76 é o valor de $F_{0{,}05}$ para $r - 1 = 4 - 1 = 3$ e $(r-1)(r-2) = (4-1)(4-2) = 6$ graus de liberdade.

4. Substituindo $r = 4$, $T_{1.} = 181$, $T_{2.} = 186$, $T_{3.} = 183$, $T_{4.} = 192$, $T_{.1} = 175$, $T_{.2} = 179$, $T_{.3} = 185$,

$T_{.4} = 203$, $T_A = 197$, $T_B = 167$, $T_C = 179$, $T_D = 199$, $T... = 742$ e $\sum\sum x^2 = 34.756$ nas fórmulas de cálculo para as somas dos quadrados, obtemos

$$SQT = 34.756 - \frac{1}{16}(742)^2 = 34.756 - 34.410{,}25 = 345{,}75$$

$$SQL = \frac{1}{4}(181^2 + 186^2 + 183^2 + 192^2) - 34.410{,}25 = 17{,}25$$

$$SQC = \frac{1}{4}(175^2 + 179^2 + 185^2 + 203^2) - 34.410{,}25 = 114{,}75$$

$$SQ(Tr) = \frac{1}{4}(197^2 + 167^2 + 179^2 + 199^2) - 34.410{,}25 = 174{,}75$$

$$SQE = 345{,}75 - (17{,}25 + 114{,}75 + 174{,}75) = 39{,}00$$

O trabalho restante é apresentado na seguinte tabela de análise de variância:

Fonte de variação	*Graus de liberdade*	*Soma de quadrados*	*Quadrado médio*	*F*
Linhas (regiões)	3	17,25	$\frac{17{,}25}{3} = 5{,}75$	$\frac{5{,}75}{6{,}5} \approx 0{,}88$
Colunas (método promocional)	3	114,75	$\frac{114{,}75}{3} = 38{,}25$	$\frac{38{,}25}{6{,}5} \approx 5{,}88$
Tratamentos (embalagens)	3	174,75	$\frac{174{,}75}{3} = 58{,}25$	$\frac{58{,}25}{6{,}5} \approx 8{,}96$
Erro	6	39,00	$\frac{39{,}00}{6} = 6{,}5$	
Total	15	345,75		

5. Para as linhas, como $F = 0{,}88$ é menor do que 4,76, a hipótese nula não pode ser rejeitada; para as colunas, como $F = 5{,}88$ excede 4,76, a hipótese nula deve ser rejeitada; para os tratamentos, como $F = 8{,}96$ excede 4,76, a hipótese nula deve ser rejeitada. Em outras palavras, concluímos que foram as diferenças em promoção e embalagem, e não as diferentes regiões, que afetaram a venda do lanche rápido.

Há muitos outros planejamentos experimentais além dos que abordamos, que atendem a uma ampla diversidade de propósitos especiais. São largamente utilizados, por exemplo, os **planejamentos em bloco incompletos**, aplicáveis quando não é possível termos cada tratamento em cada bloco.

A necessidade de tal planejamento surge, por exemplo, quando desejamos comparar 13 marcas de pneus mas não podemos colocá-los todos num carro de teste ao mesmo tempo. Numerando os pneus de 1 a 13, podemos usar o seguinte planejamento experimental:

Teste	Pneu	Teste	Pneu
1	1 2 4 10	*8*	8 9 11 4
2	2 3 5 11	*9*	9 10 12 5
3	3 4 6 12	*10*	10 11 13 6
4	4 5 7 13	*11*	11 12 1 7
5	5 6 8 1	*12*	12 13 2 8
6	6 7 9 2	*13*	13 1 3 9
7	7 8 10 3		

Há aqui 13 repetições de testes, ou blocos, e como cada tipo de pneu aparece juntamente com outro tipo de pneu uma vez dentro do mesmo bloco, o planejamento é denominado **planejamento em bloco incompleto equilibrado**. É importante o fato de cada tipo de pneu aparecer juntamente com outro tipo de pneu uma vez dentro do mesmo bloco, pois facilita a análise estatística ao assegurar que temos a mesma quantidade de informação para comparar cada par de tipos de pneu. Em geral, a análise dos planejamentos em bloco incompletos é bastante complicada, e não a abordaremos aqui, pois nosso objetivo é apenas mostrar o que se pode fazer com o planejamento cuidadoso de um experimento.

EXERCÍCIOS

15.37 Um agrônomo deseja comparar a safra de 15 variedades de milho e, ao mesmo tempo, estudar os efeitos de quatro fertilizantes diferentes e de três métodos de irrigação. Quantos lotes de teste ele deve plantar, se cada variedade de milho deve crescer num lote de teste com cada combinação possível de fertilizantes e métodos de irrigação?

15.38 Suponha que queiramos comparar o número de peças defeituosas fabricadas por cinco operários trabalhando em quatro máquinas diferentes (1, 2, 3 e 4) em dois turnos diferentes (I e II).

(a) Considerando os operários como "tratamentos" diferentes, relacione os blocos (combinações de máquinas e turnos) que seriam necessários para que cada peça fosse fabricada em cada turno.

(b) Quantas observações seriam necessárias se cada operário deve trabalhar duas vezes em cada máquina em cada turno?

15.39 Um fabricante de produtos farmacêuticos deseja lançar no mercado um novo remédio contra resfriados, que na verdade é uma combinação de quatro medicamentos, e tenciona experimentar inicialmente com duas dosagens de cada medicamento. Se A_L e A_H denotam as dosagens baixa e alta, respectivamente, do medicamento A, B_L e B_H denotam as dosagens baixa e alta, respectivamente, do medicamento B, C_L e C_H denotam as dosagens baixa e alta, respectivamente, do medicamento C, e D_L e D_H denotam as dosagens baixa e alta, res-

pectivamente, do medicamento D, relacione as 16 combinações que devem ser testadas, se cada dosagem de cada medicamento deve ser usada uma vez em combinação com cada dosagem de cada um dos outros medicamentos.

15.40 Usando o fato de que cada uma das letras deve ocorrer uma e uma só vez em cada linha e em cada coluna, complete os seguintes quadrados latinos:

(a)

		A
	B	

(b)

	A		
			B
A	C		
		C	

(c)

	A	E		
		B		E
C			A	
D				
				D

15.41 Para comparar quatro marcas diferentes de bolas de golfe, A, B, C e D, cada tipo foi lançado por cada um de quatro golfistas profissionais, P_1, P_2, P_3 e P_4, utilizando uma vez cada um de quatro tacos diferentes T_1, T_2, T_3 e T_4. As distâncias (em metros) do ponto de lançamento ao ponto em que as bolas pararam são dadas na tabela seguinte:

	T_1	T_2	T_3	T_4
P_1	D 231	B 215	A 261	C 199
P_2	C 234	A 300	B 280	D 266
P_3	A 301	C 208	D 247	B 255
P_4	B 253	D 258	C 210	A 290

Considerando que as suposições necessárias possam ser atendidas, use um computador para analisar esse quadrado latino, usando o nível 0,05 de significância para cada teste.

15.42 Os dados amostrais do quadrado latino 3 × 3 a seguir são as notas em História obtidas por nove alunos de faculdade de diferentes etnias e com vários interesses profissionais, que estudaram sob orientação dos professores A, B e C:

	Etnia		
	Latino	*Germânico*	*Eslavo*
Direito	A 75	B 86	C 69
Medicina	B 95	C 79	A 86
Engenharia	C 70	A 83	B 93

Considerando que as suposições necessárias possam ser atendidas, use um computador para analisar esse quadrado latino, usando o nível 0,05 de significância para cada teste.

15.43 De nove pessoas entrevistadas numa pesquisa, três são do Nordeste, três são do Sudeste e três são do Oeste. Quanto à profissão, três são professores, três são advogados e três são médicos, não havendo dois quaisquer da mesma profissão que provenham da mesma região. Além disso, três votam com o partido A, três com o partido B e três com o parido C, não havendo dois da mesma filiação política que exerçam a mesma profissão ou provenham da mesma região. Se um dos professores é do Nordeste e vota com o Partido C, outro professor é do Sudeste e vota com o partido B, e um dos advogados é do Sudeste e vota com o Partido A, qual é a filiação política do médico que é do Oeste? (*Sugestão*: construa um quadrado latino 3 × 3. Esse exercício é uma versão simplificada de um famoso problema proposto por R. A. Fisher em sua obra clássica *The Design of Experiments*.)

15.44 Para testar sua capacidade de tomar decisões sob pressão, nove dos mais graduados executivos de uma companhia devem ser entrevistados por cada um de quatro psicólogos. Como um psicólogo necessita de um dia inteiro para entrevistar três executivos, o esquema de entrevistas foi arranjado como segue, com os nove administradores denotados por *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H* e *I*:

Dia	Psicólogo	Executivos		
Março 2	I	*B*	*C*	?
Março 3	I	*E*	*F*	*G*
Março 4	I	*H*	*I*	*A*
Março 5	II	*C*	?	*H*
Março 6	II	*B*	*F*	*A*
Março 9	II	*D*	*E*	?
Março 10	III	*D*	*G*	*A*
Março 11	III	*C*	*F*	?
Março 12	III	*B*	*E*	*H*
Março 13	IV	*B*	?	*I*
Março 16	IV	*C*	?	*A*
Março 17	IV	*D*	*F*	*H*

Substitua os seis pontos de interrogação pelas letras apropriadas, considerando que cada um dos nove executivos deve ser entrevistado juntamente com cada um dos outros executivos uma e só uma vez no mesmo dia. Observe que isso tornará o arranjo um planejamento em bloco incompleto equilibrado, o que pode ser importante porque cada executivo é testado juntamente com outro, uma única vez sob condições idênticas.

15.45 Um jornal publica regularmente colunas de sete colaboradores, mas em cada edição tem espaço apenas para três deles. Complete a tabela a seguir, em que os colunistas são numerados de 1 a 7, de modo que o artigo de cada colunista apareça três vezes por semana e um artigo de cada colunista apareça em conjunto com um de cada um dos outros colunistas uma vez por semana.

Dia	*Colunista*		
Segunda	1	2	3
Terça	4		
Quarta	1	4	5
Quinta	2		
Sexta	1	6	7
Sábado	5		
Domingo	2	4	6

15.11 LISTA DE TERMOS-CHAVE (com indicação das páginas de suas definições)

Aleatorização, 363, 367
Análise de variância, 362, 368
Análise de variância de dois critérios, 380
Análise de variância de um critério, 363, 368
ANOVA, 362
Blocos, 380
Blocos completos, 380
Bloqueamento, 363, 380
Comparação múltipla, 363, 375
Distribuição F, 365
Efeitos de blocos, 381
Efeitos de tratamento, 363
Erro experimental, 368
Estatística F, 364
Estudentização, 375
Experimento controlado, 367
Experimento de dois fatores, 380
Experimentos de dois critérios, 363
Fatores, 384
Graus de liberdade do denominador, 365
Graus de liberdade do numerador, 365
Interação, 380
Intervalo estudentizado, 375
Média global, 363, 368
Planejamento completamente aleatorizado, 367
Planejamento de experimentos, 362
Planejamento em bloco aleatorizado, 380
Planejamento em bloco incompleto equilibrado, 393
Planejamento em bloco incompleto, 392
Quadrado latino, 389
Quadrado médio de erro, 370
Quadrado médio de tratamento, 369
Razão de variâncias, 365
Replicação, 385
Soma de quadrados de blocos, 381
Soma de quadrados de erros, 369
Soma de quadrados de tratamentos, 369
Soma de quadrados total, 368
Somas de quadrados de interação, 386
Tabela de análise de variância, 370
Total geral, 371
Tratamentos, 369

15.12 REFERÊNCIAS

Os seguintes são alguns dos muitos livros escritos sobre a análise de variância:

GUNTHER, W. C., *Analysis of Variance*, Upper Saddle River, N. J.: Prentice-Hall, Inc., 1964.

SNEDECOR, G. W., and COCHRAN, W. G., *Statistical Methods*, 6th ed. Ames, Iowa: Iowa State University Press, 1973.

Problemas relacionados ao planejamento de experimentos são abordados nos livros precedentes e em

ANDERSON, V. L., and MCLEAN, R. A., *Design of Experiments: A Realistic Approach*, New York: Marcel Dekker, Inc., 1974.

BOX, G. E. P., HUNTER, W. G., and HUNTER, J. S., *Statistics for Experimenters*, New York: John Wiley & Sons, Inc., 1978.

COCHRAN, W. G., and COX, G. M., *Experimental Design*, 2nd ed. New York: John Wiley & Sons, Inc., 1957.

FINNEY, D. J., *An Introduction to the Theory of Experimental Design*. Chicago: The University of Chicago Press, 1960.

FLEISS, J., *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, Inc., 1986.

HICKS, C. R., *Fundamental Concepts in the Design of Experiments*, 2nd ed. New York: Holt, Rinehart and Winston, 1973.

ROMANO, A., *Applied Statistics for Science and Industry*. Boston: Allyn and Bacon, Inc., 1977.

No livro mencionado, de W. G. COCHRAN e G. M. COX, encontra-se uma tabela de quadrados latinos para r = 3, 4, 5, . . ., e 12.

Alguns problemas de planejamento experimental são discutidos informalmente nos Capítulos 18 e 19 de

BROOK, R. J., ARNOLD, G. C., HASSARD, T. H., and PRINGLE, R. M., eds., *The Fascination of Statistics*. New York: Marcel Dekker, Inc., 1986.

O tópico de comparações múltiplas é tratado detalhadamente em

FEDERER, W. T., *Experimental Design, Theory and Application*. New York: Macmillan Publishing Co., Inc., 1955.

HOCHBERG, Y., and TAMHANE, A. *Multiple Comparison Procedures*. New York: John Wiley & Sons, Inc., 1987

16

REGRESSÃO

Em muitas pesquisas estatísticas, o objetivo principal é estabelecer relações que possibilitem prever uma ou mais variáveis em termos de outras. Assim é que se fazem estudos para prever as vendas futuras de um produto em função do seu preço, ou a perda de peso de uma pessoa em decorrência do número de semanas que se submete a uma dieta de 800 calorias-dia, ou as despesas de uma família com médico e remédios em função de sua renda, ou o consumo *per capita* de certos alimentos em função de seu valor nutritivo e do gasto com propaganda na televisão, e assim por diante.

Naturalmente, o ideal seria que pudéssemos prever uma quantidade exatamente em termos de outra, mas isso raramente é possível. Na maioria dos casos, devemos contentar-nos com a previsão de médias ou de valores esperados. Por exemplo, não podemos prever exatamente quanto ganhará um indivíduo específico formado em nível superior dez anos depois de sua formatura mas, com base em dados adequados, é possível prevermos o ganho médio de todos os graduados em nível superior dez anos depois de sua formatura. Analogamente, podemos prever a safra média de certa variedade de trigo em termos do índice pluviométrico de janeiro, e podemos prever a nota média esperada de um calouro do curso de Direito em função do seu QI. Esse problema da previsão do valor médio de uma variável em termos do valor conhecido de outra variável (ou dos valores conhecidos de outras variáveis) constitui o assim denominado problema da **regressão**. A origem desse termo remonta a Francis Galton (1822-1911), que o empregou pela primeira vez num estudo da relação entre as alturas de pais e filhos.

Nas Seções 16.1 e 16.2, apresentamos uma introdução geral ao ajuste de curvas e ao método mais utilizado, o **método dos mínimos quadrados**. Depois, na Seção 16.3, discutimos as questões referentes a inferências baseadas em ajuste de retas a dados emparelhados. Problemas nos quais as previsões se baseiam em mais de uma variável e problemas em que a relação entre duas variáveis não é linear são tratados nas Seções 16.4 e 16.5, que são opcionais.

16.1 AJUSTE DE CURVAS

Sempre que possível, procuramos expressar, ou aproximar, as relações entre grandezas conhecidas e grandezas que devem ser determinadas em termos de equações matemáticas. Isso tem tido muito sucesso nas ciências naturais, nas quais sabemos, por exemplo, que, a uma temperatura constante, a relação entre o volume y e a pressão x de um gás é dada pela fórmula

$$y = \frac{k}{x}$$

onde k é uma constante numérica. Mostra-se, também, que a relação entre o tamanho y de uma cultura de bactérias, e o tempo x durante o qual esteve exposta a certas condições ambientais, é dada por

$$y = a \cdot b^x$$

onde a e b são constantes numéricas. Mais recentemente, equações como essas têm sido usadas também para descrever relações no campo das ciências do comportamento, das ciências sociais e outros campos. Assim é que a primeira das equações precedentes costuma ser usada em Economia para descrever a relação entre preço e demanda, e a segunda tem sido usada para descrever o crescimento do vocabulário de uma pessoa ou a acumulação de riqueza.

Sempre que utilizamos dados observados para chegar a uma equação matemática que descreva a relação entre duas variáveis, o que constitui um processo conhecido como **ajuste de curvas**, precisamos encarar três tipos de problemas:

Devemos decidir que tipo de curva e, daí, que tipo de equação "de previsão" queremos utilizar.

Devemos encontrar a equação particular que é a melhor em algum sentido.

Devemos investigar certas questões relativas aos méritos da equação escolhida e de previsões feitas a partir dela.

O segundo desses problemas é abordado com algum detalhe na Seção 16.2, e o terceiro, na Seção 16.3.

O primeiro tipo de problema, em geral, é resolvido por inspeção direta dos dados. Esboçamos os dados em papel comum (aritmético) de gráfico ou, eventualmente, em papel de gráfico com escalas especiais (veja Seção 15.5) e decidimos visualmente o tipo de curva (uma reta, uma parábola,...) que melhor descreve o padrão geral dos dados. Há métodos que nos permitem fazer isso de maneira mais objetiva, mas são bastante avançados e não serão discutidos neste livro.

Para os nossos objetivos aqui, vamos nos concentrar principalmente em **equações lineares** a duas incógnitas. Essas equações são da forma

$$y = a + bx$$

onde a é o corte no eixo y (o valor de y para $x = 0$) e b é a inclinação da reta (a saber, a variação de y que acompanha um aumento de uma unidade em x).* As equações lineares são úteis e importantes não só porque muitas relações têm efetivamente essa forma, mas também porque, muitas vezes, constituem boas aproximações de relações que, de outro modo, seriam difíceis de descrever em termos matemáticos.

* Em outros ramos da Matemática, as equações lineares a duas incógnitas muitas vezes são escritas como $y = mx + b$, mas $y = a + bx$ tem a vantagem de prestar-se mais facilmente a generalizações, como, por exemplo, em $y = a + bx + cx^2$ ou em $y = a + b_1x_1 + b_2x_2$.

A expressão "equação linear" decorre do fato de que o gráfico de $y = a + bx$ é uma linha reta. Ou seja, todos pares de valores de x e y que satisfazem uma equação da forma $y = a + bx$ são pontos que estão sobre uma reta. Na prática, os valores de a e b costumam ser estimados com base em dados observados e, uma vez determinados, podemos substituir valores de x na equação e calcular os correspondentes valores que previmos para y.

A título de ilustração, suponha que tenhamos dados sobre a safra de trigo y de um município do Mato Grosso (em sacos por hectare), e sobre a precipitação pluviométrica x (em centímetros medidos de março a fevereiro) e que, ainda, pelo método da Seção 16.2, obtenhamos a equação de previsão

$$y = 0{,}23 + 4{,}42x$$

(veja Exercício 16.8). A Figura 16.1 exibe o gráfico correspondente, devendo-se observar que, para qualquer par de valores de x e y tais que $y = 0{,}23 + 4{,}42x$, obtemos um ponto (x, y) que cai na reta. Substituindo $x = 6$, por exemplo, vemos que, quando há uma precipitação anual de 6 centímetros, podemos esperar uma colheita de

$$y = 0{,}23 + 4{,}42 \cdot 6 = 26{,}75$$

sacos por hectare; da mesma forma, substituindo $x = 12$, vemos que, para uma precipitação anual de 12 centímetros, podemos esperar uma colheita de

$$y = 0{,}23 + 4{,}42 \cdot 12 = 53{,}27$$

sacos por hectare. Os pontos (6; 26,75) e (12; 53,27) estão sobre a reta da Figura 16.1, e isso vale para quaisquer outros pontos determinados da mesma maneira.

16.2 O MÉTODO DOS MÍNIMOS QUADRADOS

Uma vez que tenhamos decidido ajustar uma linha reta a um determinado conjunto de dados, encontramos o segundo tipo de problema, a saber, a determinação da equação da reta particular que, em certo sentido, constitui o melhor ajuste. Para ilustrar o que está em jogo, consideremos os seguintes dados amostrais obtidos num estudo da relação entre o tempo durante o qual uma pessoa esteve exposta a um alto nível de ruído e a amplitude da freqüência sonora à qual seus ouvidos respondem. Aqui x é o tempo (arredondado para a semana mais próxima) que uma pessoa mora na proximidade de um aeroporto movimentado, diretamente na trajetória dos aviões que decolam e pousam, e y é o seu alcance auditivo (em milhares de ciclos por segundo):

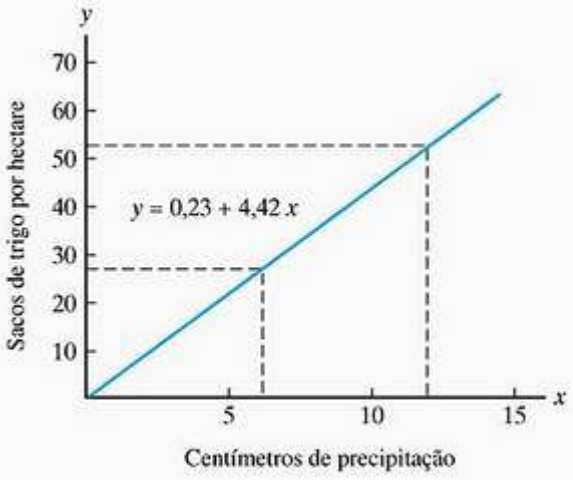

Figura 16.1 Gráfico de uma equação linear.

Número de semanas x	*Alcance auditivo* y
47	15,1
56	14,1
116	13,2
178	12,7
19	14,6
75	13,8
160	11,9
31	14,8
12	15,3
164	12,6
43	14,7
74	14,0

Esses doze **pontos de dados** (x, y) estão esboçados na Figura 16.2 no que se denomina um **diagrama de dispersão**. Isso foi feito com a ajuda de um computador, mas teria sido fácil fazê-lo à mão. Como pode ser visto, os pontos não caem todos sobre uma reta, mas o padrão geral da relação é descrito satisfatoriamente como sendo linear. Pelo menos, não há um desvio acentuado da linearidade, e por isso nos sentimos justificados na decisão de que uma linha reta é uma descrição adequada da relação subjacente.

Chegamos agora ao problema de encontrar a equação da reta que, em certo sentido, constitui o melhor ajuste aos dados e que, esperamos, virá a dar as melhores previsões possíveis de y a partir de x. Do ponto de vista lógico, não há uma limitação para o número de retas que podem ser traçadas numa folha de papel de gráfico. Algumas dessas retas ajustam tão mal os dados que podemos simplesmente ignorá-las, mas muitas outras parecem constituir ajustes mais ou menos bons, e o problema é encontrar justamente a reta que melhor se ajuste aos dados de alguma forma bem definida. Se todos os pontos se situam sobre uma reta, não existe problema, mas isso é um caso extremo, raramente encontrado na prática. Em geral, devemos contentar-nos com uma reta que tenha certas propriedades desejáveis, mas não necessariamente perfeita.

O critério que, hoje em dia, é usado quase que exclusivamente para definir uma reta de "melhor" ajuste, remonta à primeira metade do século XIX e ao trabalho do matemático francês Adrien Legendre; é conhecido como o **método dos mínimos quadrados**. Da maneira em que será utilizado aqui, esse método requer que a reta que ajustamos aos dados tenha a propriedade de que seja mínima a soma dos quadrados das distâncias verticais dos pontos à reta.

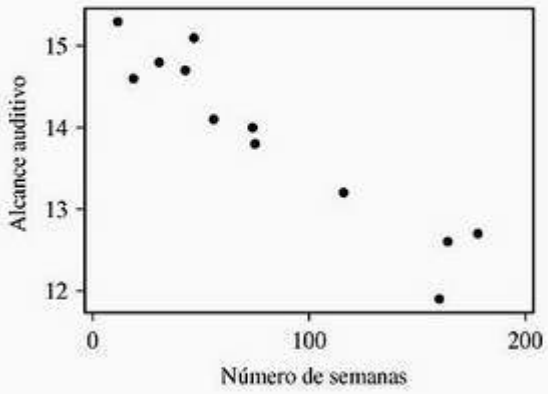

Figura 16.2
Impresso de computador para os dados auditivos.

Para explicar por que isso é feito, consideremos os dados a seguir, que poderiam representar os números de respostas certas, x e y, dadas por quatro estudantes, em duas partes de um teste de múltipla escolha:

x	y
4	6
9	10
1	2
6	2

Na Figura 16.3, esboçamos os pontos de dados correspondentes e traçamos duas retas por esses pontos para descrever o padrão geral.

Se utilizarmos a reta horizontal do diagrama à esquerda para "prever" y para os valores dados de x, obteremos $y = 5$ em cada caso, e os erros dessa "previsão" são $6 - 5 = 1$, $10 - 5 = 5$, $2 - 5 = -3$ e $2 - 5 = -3$. Na Figura 16.3, esses são os desvios verticais dos pontos de dados até a reta.

A soma desses erros é $1 + 5 + (-3) + (-3) = 0$, mas isso não é indicativo do tamanho desses erros, e nos encontramos numa situação semelhante à da página 86, que nos levou à definição do desvio-padrão. Elevando os erros ao quadrado, tal como elevamos ao quadrado os desvios da média, à página 86, vemos que a soma dos quadrados dos erros é $1^2 + 5^2 + (-3)^2 + (-3)^2 = 44$.

Consideremos, agora, a reta do diagrama à direita, traçada de modo a passar pelos pontos (1, 2) e (9, 10); vê-se facilmente que sua equação é $y = 1 + x$. Visualmente, essa reta parece ajustar-se muito melhor aos dados do que a reta horizontal do diagrama à esquerda e, se a utilizarmos para prever y para os valores dados de x, obteremos $1 + 4 = 5$, $1 + 9 = 10$, $1 + 1 = 2$ e $1 + 6 = 7$. Os erros dessas "previsões," que, na figura à direita, também são as distâncias verticais dos pontos de dados até a reta, são $6 - 5 = 1$, $10 - 10 = 0$, $2 - 2 = 0$ e $2 - 7 = -5$.

A soma desses erros é $1 + 0 + 0 + (-5) = -4$, que é numericamente maior do que a soma dos erros que obtivemos em relação à outra reta da Figura 16.3, mas isso não tem importância. A soma dos quadrados dos erros é agora $1^2 + 0^2 + 0^2 + (-5)^2 = 26$, e isso é muito menos do que o valor 44 obtido antes. Nesse sentido, a reta à direita proporciona um ajuste muito melhor aos dados do que a reta horizontal à esquerda.

Podemos ir um pouco mais além e procurar determinar a equação da reta para a qual a soma dos quadrados dos erros (a soma dos quadrados dos desvios verticais dos pontos de dados da reta) é um mínimo. No Exercício 16.11, pede-se ao leitor verificar que a equação de uma tal reta é $y = \frac{15}{17} + \frac{14}{17}x$ para o nosso exemplo. Essa reta é denominada a **reta de mínimos quadrados**.

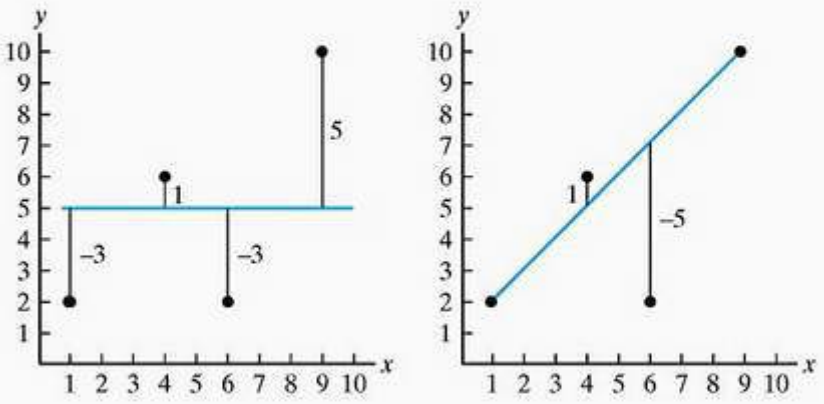

Figura 16.3
Duas retas ajustadas aos quatro pontos de dados.

Para mostrar como a equação de uma tal reta é efetivamente obtida para um dado conjunto de **pontos de dados**, consideremos n pares de números (x_1, y_1), (x_2, y_2), . . ., e (x_n, y_n), que podem representar, por exemplo, o impulso e a velocidade de n foguetes, a altura e o peso de n pessoas, a velocidade de leitura e o grau de compreensão de n estudantes, o custo de resgatar n navios afundados e o valor dos tesouros descobertos, ou a idade e custo de consertar n automóveis. Escrevendo a equação da reta como $\hat{y} = a + bx$, onde o símbolo $\hat{y}$ ("ípsilon chapéu") serve para distinguir entre os valores observados y e o valores correspondentes $\hat{y}$ na reta, o critério dos mínimos quadrados exige que minimizemos a soma dos quadrados das diferenças entre os y e os $\hat{y}$ (ver Figura 16.4). Isso significa que devemos encontrar os valores numéricos das constantes a e b que figuram na equação $\hat{y} = a + bx$, para os quais

$$\sum (y - \hat{y})^2 = \sum [y - (a + bx)]^2$$

tem o menor valor possível. Como a determinação das expressões de a e b que minimizam $\sum (y - \hat{y})^2$ é bastante trabalhosa ou exige recursos do Cálculo, limitamo-nos a enunciar o resultado, que a e b são dados pelas soluções, em relação a a e b, do seguinte sistema de duas equações lineares:

$$\sum y = na + b\left(\sum x\right)$$
$$\sum xy = a\left(\sum x\right) + b\left(\sum x^2\right)$$

Nessas equações, denominadas **equações normais**, n é o número de pares de observações, $\sum x$ e $\sum y$ são as somas dos valores observados x e y, $\sum x^2$ é a soma dos quadrados dos valores de x e $\sum xy$ é a soma dos produtos determinados multiplicando cada x pelo y correspondente.

EXEMPLO 16.1 Sabendo que para os dados de alcance auditivo à página 401 temos $n = 12$, $\sum x = 975$, $\sum x^2 = 117.397$, $\sum y = 166{,}8$, $\sum y^2 = 2.331{,}54$ e $\sum xy = 12.884{,}4$, obtenha as equações normais que determinam uma reta de mínimos quadrados.

Solução Substituindo $n = 12$ e quatro das cinco somas nas expressões das equações normais, obtemos

$$166{,}8 = 12a + 975b$$
$$12.884{,}4 = 975a + 117.397b$$

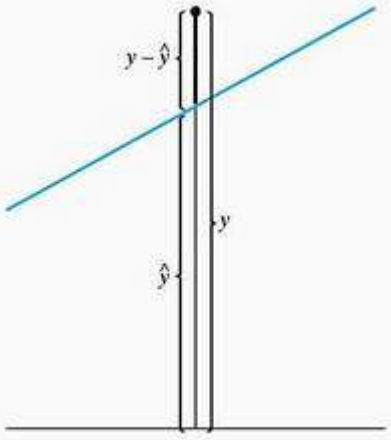

Figura 16.4
A diferença entre y e $\hat{y}$.

(Observe que não utilizamos $\sum y^2$ nesse exemplo, mas fornecemos esse valor agora, junto com os demais, para uso futuro.)

O leitor com alguma experiência na resolução de sistemas de equações lineares em Álgebra Linear elementar, pode continuar o exemplo acima e resolver essas duas equações em a e b usando ou o **método de eliminação gaussiana** ou então o método que utiliza **determinantes**. Alternativamente, podemos resolver as duas equações normais simbolicamente em a e b e então substituir os valores de n e dos diversos somatórios nas fórmulas resultantes. Dentre as várias maneiras de escrever essas fórmulas, talvez a mais conveniente seja o formato em que utilizamos as quantidades

$$S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 \quad \text{e} \quad S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

como quantidades intermediárias e então escrevemos as fórmulas para calcular a e b como

SOLUÇÕES DE EQUAÇÕES NORMAIS

$$b = \frac{S_{xy}}{S_{xx}}$$

$$a = \frac{\sum y - b\left(\sum x\right)}{n}$$

em que primeiro calculamos b e então substituímos esse valor na fórmula para a. (Observe que esse é o mesmo S_{xx} que utilizamos à página 89 na fórmula para calcular e desvio-padrão amostral.)

EXEMPLO 16.2 Utilize essas fórmulas para resolver as equações normais e encontrar a e b para o exemplo do alcance auditivo.

Solução Primeiro substituímos $n = 12$ e os somatórios necessários dados no Exemplo 16.1 nas fórmulas para S_{xx} e S_{xy}, obtendo

$$S_{xx} = 117.397 - \frac{1}{12}(975)^2 = 38.178{,}25$$

e

$$S_{xy} = 12.884{,}4 - \frac{1}{12}(975)(166{,}8) = -668{,}1$$

Então $b = \dfrac{-668{,}1}{38.178{,}25} \approx -0{,}0175$ e $a = \dfrac{166{,}8 - (-0{,}0175)(975)}{12} \approx 15{,}3$, ambos arredondados até o terceiro dígito significativo, e a equação da reta de mínimos quadrados pode ser escrita como

$$\hat{y} = 15{,}3 - 0{,}0175x$$

O que fizemos aqui pode muito bem ser descrito como um simples exercício de Aritmética, pois raramente, ou nunca, utilizamos tantos detalhes na determinação de uma reta de mínimos quadrados. Hoje em dia, os somatórios necessários podem ser obtidos até com as mais primitivas calculadoras, e os valores de a e b podem ser obtidos com qualquer tipo de programa estatístico. Na verdade, a parte mais sutil de toda essa operação é a digitação dos dados e, se necessário, fazer correções, a menos que utilizemos um computador ou uma calculadora gráfica, nos quais os dados podem ser dispostos e editados.

Observe também que quando b é negativo, como no Exemplo 16.2, a reta de mínimos quadrados tem uma *inclinação negativa* indo da esquerda para a direita. Em outras palavras, a relação entre x e y é tal que y decresce quando x cresce, como pode ser observado na Figura 16.2. Por

outro lado, quando b é positivo, a reta de mínimos quadrados tem uma *inclinação positiva* indo da esquerda para a direita, ou seja, que y cresce quando x cresce. Finalmente, quando b é igual a zero, a reta de mínimos quadrados é horizontal e o valor de x não é útil na estimação ou previsão do valor de y.

EXEMPLO 16.3 Use uma calculadora gráfica para refazer o Exemplo 16.2 sem utilizar os somatórios dados no Exemplo 16.1.

Solução A Figura 16.5 mostra os dados originais digitados numa calculadora gráfica. A janela da calculadora é muito pequena para mostrar todos os dados, mas o resto dos dados foi obtido dragando. Em seguida, o comando **STAT CALC 8** fornece os resultados mostrados na Figura 16.6. Arredondando até o terceiro dígito significativo, como antes, obtemos $a = 15{,}3$ e $b = -0{,}0175$, e a equação da reta de mínimos quadrados é, evidentemente, a mesma

$$\hat{y} = 15{,}3 - 0{,}0175x$$

Se tivéssemos utilizado um computador nesse exemplo, o MINITAB teria fornecido o impresso mostrado na Figura 16.7. A equação da reta de mínimos quadrados, denominada **equação de regressão** (o que será explicado mais tarde), novamente é $y = 15{,}3 - 0{,}0175$ e os coeficientes a e b são dados na coluna encabeçada por "Coef" como 15,3218 e – 0,017499. Alguns dos detalhes adicionais do impresso serão utilizados mais adiante.

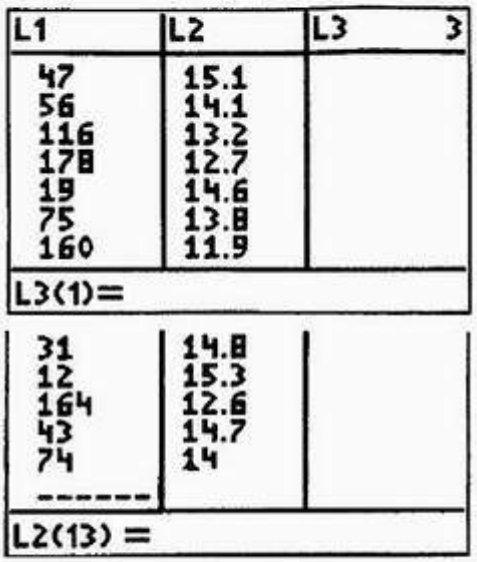

Figura 16.5
Dados do Exemplo 16.3.

Figura 16.6
Solução do Exemplo 16.3.

Análise de Regressão: *y* contra *x*

```
The regression equation is
y = 15.3 - 0.0175 x

Predictor       Coef          SE Coef          T          P
Constant     15.3218          0.1845       83.4      0.000
x          -0.017499        0.001865      -9.38      0.000

S = 0.3645          R-Sq = 89.8%     R-Sq(adj) = 88.8%

Analysis of Variance

Source           DF         SS         MS          F          P
Regression        1     11.691     11.691      88.00      0.000
Residual Error   10      1.329      0.133
Total            11     13.020
```

Figura 16.7 Impresso de MINITAB para o Exemplo 16.3.

EXEMPLO 16.4 Use a equação de mínimos quadrados obtida no Exemplo 16.2 ou no Exemplo 16.3 para estimar o alcance auditivo de uma pessoa que foi exposta ao ruído de aeroporto (conforme descrito à página 400) durante

(a) um ano;

(b) dois anos.

Solução **(a)** Substituindo $x = 52$ em $\hat{y} = 15{,}3 - 0{,}0175x$, obtemos $\hat{y} = 15{,}3 - 0{,}0175(52) = 14{,}4$ mil ciclos por segundo arredondando até o terceiro dígito significativo.

(b) Substituindo $x = 104$ nesta equação, obtemos $\hat{y} = 13{,}5$ mil ciclos por segundo, arredondado até o terceiro dígito significativo.

Quando fazemos uma estimativa como essa, ou uma previsão, na realidade não podemos esperar sempre obter precisamente a resposta correta. Com referência ao nosso exemplo, não seria nada razoável esperar que cada pessoa que tivesse sido exposta ao ruído de aeroporto durante um dado intervalo de tempo apresentasse precisamente o mesmo alcance auditivo. Para tornar nossas previsões significativas com base em retas de mínimos quadrados, devemos considerar como médias, ou valores esperados, os valores de $\hat{y}$ obtidos mediante substituição de valores dados de x. Interpretadas dessa maneira, as retas de mínimos quadrados são denominadas **retas de regressão** ou, melhor ainda, **retas de regressão estimadas**, já que os valores de a e b são estimados com base em dados amostrais e, portanto, pode-se esperar que variem de amostra para amostra. Na Seção 16.3, discutiremos questões relativas à validade dessas estimativas.

Nas discussões desta seção, consideramos apenas o problema de ajustar uma reta a pares de dados. Mais geralmente, o método dos mínimos quadrados também pode ser utilizado para ajustar outros tipos de curvas e para deduzir equações de previsão a mais de duas incógnitas. O problema de ajustar curvas além da reta pelo método dos mínimos quadrados será abordado sucintamente na Seção 16.5 e, na Seção 16.4, daremos alguns exemplos de previsão de equações a mais de duas incógnitas. Ambas seções estão marcadas como opcionais.

EXERCÍCIOS

16.1 Um cachorro com seis horas de treinamento de obediência cometeu cinco erros numa exposição canina, um cachorro com doze horas de treinamento de obediência cometeu seis erros, e um cachorro com dezoito horas de treinamento de obediência cometeu apenas um erro. Denotando por x o número de horas de treinamento de obediência e por y o número de erros cometidos, qual das duas retas

$$y = 10 - \frac{1}{2}x \quad \text{ou} \quad y = 8 - \frac{1}{3}x$$

fornece um ajuste melhor aos três pontos de dados, (6, 5), (12, 6) e (18, 1), no sentido de mínimos quadrados?

16.2 Com referência ao Exercício 16.1, use um computador ou uma calculadora gráfica para conferir se a reta de melhor ajuste é uma reta de mínimos quadrados.

16.3 Para verificar se um conservante de alimentos largamente utilizado contribui para a hiperatividade de crianças em idade pré-escolar, um nutricionista escolheu uma amostra aleatória de dez crianças de quatro anos reconhecidas como bastante hiperativas de várias escolinhas e observou seu comportamento 45 minutos depois de terem ingerido quantidades controladas de comida contendo o conservante. Na tabela a seguir, x é a quantidade de comida consumida contendo o conservante (em gramas) e y é uma medição subjetiva de hiperatividade (numa escala de 1 a 20) baseada na agitação da criança e na interação com outras crianças:

x	y
36	6
82	14
45	5
49	13
21	5
24	8
58	14
73	11
85	18
52	6

(a) Esboce um diagrama de dispersão para decidir se uma reta pode descrever de modo razoável o comportamento geral dos dados.
(b) Use uma régua para traçar uma reta que, visualmente, deveria estar próxima de uma reta de mínimos quadrados.
(c) Use a reta da parte (b) para estimar a medida de hiperatividade de uma dessas crianças que ingeriu 65 gramas de comida com o conservante 45 minutos antes.

16.4 Com referência ao Exercício 16.3, use um aplicativo apropriado ou uma calculadora gráfica para verificar que a equação de mínimos quadrados para estimar y em termos de x é dada por $\hat{y} = 1{,}5 + 0{,}16x$, arredondados até o segundo dígito significativo. Também use essa equação para estimar a medida de hiperatividade de uma dessas crianças que ingeriu 65 gramas de comida com o conservante 45 minutos antes e compare o resultado com o da parte (c) do Exercício 16.3.

16.5 Com referência ao Exercício 16.3, em que $\sum x = 525$, $\sum y = 100$, $\sum x^2 = 32.085$ e $\sum xy = 5.980$, monte as duas equações normais e resolva-as usando o método da eliminação gaussiana ou dos determinantes.

16.6 A tabela a seguir mostra durante quantas semanas seis pessoas estão trabalhando num posto de inspeção de automóveis e quantos carros cada uma inspecionou entre o meio dia e as 14 horas, em determinado dia:

Número de semanas trabalhadas x	*Número de carros inspecionados* y
2	13
7	21
9	23
1	14
5	15
12	21

Sabendo que $\sum x = 36$, $\sum y = 107$, $\sum x^2 = 304$ e $\sum xy = 721$, use as fórmulas dadas à página 404 para calcular a e b e, assim, obter a equação da reta de mínimos quadrados.

16.7 Use o resultado do Exercício 16.6 para estimar quantos automóveis pode-se esperar que uma pessoa inspecione durante o mesmo período de duas horas se ela está trabalhando na posto de inspeção há oito semanas.

16.8 Verifique que a equação do exemplo à página 399/400 pode ser obtida pelo ajuste de uma reta de mínimos quadrados aos seguintes dados:

Precipitação pluviométrica (em centímetros)	*Safra de trigo (em sacos por hectare)*
12,9	62,5
7,2	28,7
11,3	52,2
18,6	80,6
8,8	41,6
10,3	44,5
15,9	71,3
13,1	54,4

16.9 Os dados abaixo referem-se ao resíduo de cloro numa piscina, em vários momentos após ter sido tratada com produtos químicos:

Número de horas	*Resíduo de cloro (partes por milhão)*
0	2,2
2	1,8
4	1,5
6	1,4
8	1,1
10	1,1
12	0,9

em que a leitura a 0 hora foi feita imediatamente após completado o tratamento químico.

(a) Use as fórmulas de cálculo da página 404 para ajustar uma reta de mínimos quadrados que nos permita prever o resíduo de cloro em termos do número de horas após a piscina ter sido tratada com produtos químicos.

(b) Use a equação da reta de mínimos quadrados obtida na parte (a) para estimar o resíduo de cloro na piscina cinco horas após esta ter sido tratada com produtos químicos.
(c) Suponha que seja revelado que os dados desse exercício tenham sido obtidos durante um dia muito quente. Explique por que os resultados das partes (a) e (b) podem ser bastante enganosos.

16.10 Use um aplicativo apropriado ou uma calculadora gráfica para refazer a parte (a) do Exercício 16.9.

16.11 Com referência aos quatro pontos de dados à página 402, que eram (4, 6), (9, 10), (1, 2) e (6, 2), verifique que a equação de mínimos quadrados é

$$\hat{y} = \frac{15}{17} + \frac{14}{17}x$$

Também calcule a soma dos quadrados dos desvios verticais dos quatro pontos a essa reta e compare o resultado com 44 e 26, as somas de quadrados correspondentes obtidas para as duas retas mostradas na Figura 16.3.

16.12 A matéria-prima usada na fabricação de uma fibra sintética é armazenada num local sem controle de umidade. Durante 12 dias, mediu-se a umidade relativa no local de armazenamento e o conteúdo de umidade de uma amostra da matéria-prima (ambos em percentagens), obtendo os seguintes resultados:

Umidade x	*Conteúdo de umidade* y
46	12
53	14
37	11
42	13
34	10
29	8
60	17
44	12
41	10
48	15
33	9
40	13

(a) Esboce um diagrama de dispersão para verificar que o relacionamento global entre essas duas variáveis é muito bem descrito por uma reta.
(b) Sabendo que $\sum x = 507$, $\sum y = 144$, $\sum x^2 = 22.625$ e $\sum xy = 6.314$, monte as duas equações normais.
(c) Resolva as duas equações normais usando o método da eliminação gaussiana ou dos determinantes.

16.13 Com referência ao Exercício 16.12, use os somatórios dados na parte (b) e as fórmulas de cálculo à página 404 para encontrar a equação da reta de mínimos quadrados.

16.14 Use um aplicativo apropriado ou uma calculadora gráfica para encontrar a equação da reta de mínimos quadrados para os dados de unidade relativa e o conteúdo de umidade do Exercício 16.12.

16.15 Use a equação obtida nos Exercícios 16.12, 16.13 ou 16.14, para estimar o conteúdo de umidade quando a umidade relativa é de 38%.

16.16 Suponha que, no Exercício 16.12, quiséssemos estimar qual umidade relativa daria um conteúdo de umidade de 10%. Poderíamos fazer $\hat{y} = 10$ na equação obtida em qualquer um dos Exercícios 16.12, 16.13, ou 16.14, e resolver para x, mas isso não daria uma estimativa no sentido de mínimos quadrados. Para obter uma estimativa de mínimos quadrados da umidade relativa em termos do conteúdo de umidade, devemos denotar o conteúdo de umidade por x e a umidade relativa por y e, então, ajustar a esses dados uma reta de mínimos quadrados. Use um aplicativo apropriado ou uma calculadora gráfica para encontrar uma tal reta de mínimos quadrados e use-a para estimar a umidade relativa que dará um conteúdo de umidade de 10%.

16.17 Quando os x são igualmente espaçados (isto é, quando as diferenças entre valores sucessivos de x são todas iguais), podemos simplificar enormemente encontrar a equação de uma reta de mínimos quadrados codificando os x mediante atribuição dos valores $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$ quando n é ímpar ou $\ldots, -5, -3, -1, 1, 3, 5, \ldots$, quando n é par. Com essa codificação, e denotando os x codificados por u, a soma dos x codificados é zero e as fórmulas para calcular a e b da página 404 se tornam

$$a = \frac{\sum y}{n} \quad \text{e} \quad b = \frac{\sum uy}{\sum u^2}$$

Naturalmente, a equação da reta de mínimos quadrados resultante expressa y em termos de u e devemos levar isso em conta quando utilizarmos a equação para estimativas ou previsões.

(a) Durante seus cinco primeiros anos de operação, a receita bruta das vendas de uma companhia foi 1,4; 2,1; 2,6; 3,5; e 3,7 milhões de unidades monetárias. Ajuste uma reta de mínimos quadrados e, admitindo que a tendência permaneça, faça uma previsão da receita bruta da companhia em seu sexto ano de operação.

(b) Ao final de oito anos sucessivos, uma indústria teve 1,0; 1,7; 2,3; 3,1; 3,5; 3,4; 3,9; e 4,7 milhões de unidades monetárias investidas em instalações e equipamentos. Ajuste uma reta de mínimos quadrados e, admitindo que a tendência permaneça, faça uma previsão do investimento da companhia em instalações e equipamentos ao final do décimo ano.

***16.18** Verifique que resolvendo *simbolicamente* as equações normais usando determinantes, obtemos as fórmulas alternativas seguintes para calcular a e b:

$$a = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

***16.19** Use as fórmulas de cálculo do Exercício 16.18 para refazer o

(a) Exercício 16.6;

(b) Exercício 16.13.

16.3 ANÁLISE DE REGRESSÃO

No Exemplo 16.4, utilizamos uma reta de mínimos quadrados para estimar, ou prever, o alcance auditivo de uma pessoa exposta ao ruído de aeroporto durante dois anos como sendo 13,5 mil ciclos por segundo. Mesmo que interpretemos corretamente a reta de mínimos quadrados como

uma reta de regressão (isto é, que consideremos as estimativas baseadas nessa reta como médias ou valores esperados), ainda há perguntas que precisam ser respondidas. Por exemplo,

Qual é a precisão dos valores obtidos para a e b na equação de mínimos quadrados $\hat{y} = 15{,}3 - 0{,}0175x$?

Qual a precisão da estimativa $\hat{y} = 13{,}5$ mil ciclos por segundo do alcance auditivo médio de pessoas expostas ao ruído de aeroporto durante dois anos?

Afinal de contas, $a = 15{,}3$ e $b = -0{,}0175$, bem como $\hat{y} = 13{,}5$, são apenas estimativas baseadas em dados amostrais e, se basearmos nossos cálculos numa amostra diferente, o método de mínimos quadrados provavelmente daria valores diferentes para a e b, e um valor diferente de $\hat{y}$ para $x = 104$. Assim, para fazer previsões, poderíamos perguntar:

É possível estabelecer um intervalo para o qual possamos afirmar, com algum grau de confiança, que contém o alcance auditivo de uma pessoa expostas ao ruído de aeroporto durante dois anos?

Quanto à primeira dessas questões, dissemos que $a = 15{,}3$ e $b = -0{,}0175$ são "apenas estimativas baseadas em dados amostrais" e isso implica a existência dos correspondentes valores reais, denotados, em geral, por α e β e denominados os verdadeiros **coeficientes de regressão**. Conseqüentemente, há também uma verdadeira reta de regressão $\mu_{y|x} = \alpha + \beta x$, onde $\mu_{y|x}$ é a verdadeira média de y para um dado valor de x. Para distinguir entre os a e α e entre b e β, nos referimos a a e b como os **coeficientes de regressão estimados**. Muitas vezes os denotamos por $\hat{\alpha}$ e $\hat{\beta}$, em vez de a e b.

Para esclarecer o conceito de uma verdadeira reta de regressão, consideremos a Figura 16.8, em que esboçamos as distribuições de y para diversos valores de x. Com referência ao nosso exemplo numérico, essas curvas são as distribuições dos alcances auditivos de pessoas que ficaram expostas a ruído de aeroporto durante uma, duas e três semanas e, para completar a figura, podemos visualizar curvas análogas para todos os outros valores de x dentro do alcance dos valores sob consideração. Note que as médias de todas as distribuições da Figura 16.8 estão sobre a verdadeira reta de regressão α.

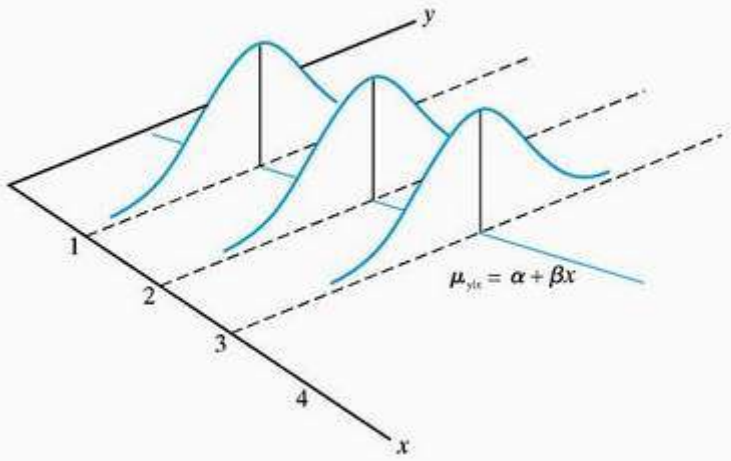

Figura 16.8
Distribuições de y para valores dados de x.

Em **análise de regressão linear**, admitimos que os x sejam constantes, não valores de variáveis aleatórias, e que, para cada valor de x, a variável a ser prevista, y, tenha uma distribuição (como na Figura 16.8) cuja média é $\alpha + \beta x$. Em **análise da regressão normal**, admitimos, ainda, que essas distribuições sejam todas normais e com o mesmo desvio-padrão σ.

Com base nessas suposições, pode ser mostrado que os coeficientes de regressão estimados a e b, obtidos pelo método dos mínimos quadrados, são valores de variáveis aleatórias de distribuições normais com médias α e β e desvios-padrão

$$\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} \quad \text{e} \quad \frac{\sigma}{\sqrt{S_{xx}}}$$

Entretanto, os coeficientes de regressão estimados, a e b, não são estatisticamente independentes. Note que ambas as fórmulas do erro-padrão exigem que estimemos σ, o desvio-padrão comum das distribuições normais ilustradas na Figura 16.8. Caso contrário, como supomos os x constantes, não há problema na determinação de $\bar{x}$ e S_{xx}. A estimativa de σ que vamos utilizar é denominada **erro-padrão da estimativa** e é denotado por s_e. Sua fórmula é

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

onde, novamente, os y são os valores observados de y e os $\hat{y}$ são os valores correspondentes na reta de mínimos quadrados. Note que s_e^2 é a soma dos quadrados dos desvios verticais dos pontos em relação à reta (a saber, a quantidade minimizada pelo método dos mínimos quadrados) dividida por $n-2$.

A fórmula precedente define s_e, mas, na prática, calculamos seu valor por meio da fórmula de cálculo

ERRO-PADRÃO DA ESTIMATIVA

$$s_e = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}}$$

onde

$$S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

analogamente à fórmula para S_{xx}, à página 404.

EXEMPLO 16.5 Calcule s_e para a reta de mínimos quadrados que ajustamos aos dados à página 401.

Solução Como $n = 12$ e já mostramos que $S_{xy} = -668{,}1$, o único outro valor que necessitamos é S_{yy}. Como $\sum y = 166{,}8$ e $\sum y^2 = 2.331{,}54$ foram dados no Exemplo 16.1, segue que

$$S_{yy} = 2.331{,}54 - \frac{1}{12}(166{,}8)^2 = 13{,}02$$

e que, portanto,

$$s_e = \sqrt{\frac{13{,}02 - (-0{,}0175)(-668{,}1)}{10}}$$
$$\approx 0{,}3645$$

Na verdade, esse trabalho todo não é realmente necessário; o resultado é dado no impresso de computador da Figura 16.7, onde diz $s = 0{,}3645$. Também uma calculadora gráfica poderia ter fornecido $s = 0{,}3644981554$, mas não exibimos esse detalhe na Figura 16.6.

Se admitirmos todas as suposições da análise de regressão normal, de que os x são constantes e que os y são valores de variáveis aleatórias de distribuições normais com as médias $\mu_{y|x} = \alpha + \beta x$ e com o mesmo desvio-padrão σ, então as inferências sobre os coeficientes de regressão α e β podem ser baseadas nas estatísticas

ESTATÍSTICAS PARA INFERÊNCIAS SOBRE COEFICIENTES DE REGRESSÃO

$$t = \frac{a - \alpha}{s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

$$t = \frac{b - \beta}{s_e/\sqrt{S_{xx}}}$$

cujas distribuições amostrais são distribuições t com $n - 2$ graus de liberdade. Note que os valores nos denominadores são estimativas dos erros-padrão correspondentes, com s_e substituindo σ.

O exemplo a seguir mostra como testar hipóteses sobre qualquer um dos coeficientes de regressão α e β.

EXEMPLO 16.6 Suponhamos que se tenha afirmado que o alcance auditivo de uma pessoa decresceu 0,02 mil ciclos por segundo para cada semana que a pessoa tenha morado na proximidade de um aeroporto, diretamente na trajetória dos aviões que decolam e pousam, e que os dados da página 401 tenham sido obtidos com o objetivo de testar essa afirmação ao nível 0,05 de significância.

Solução No que segue, devemos admitir que todas as suposições subjacentes à análise de regressão normal tenham sido satisfeitas.

1. $H_0 : \beta = -0{,}02$
   $H_A : \beta \neq -0{,}02$
2. $\alpha = 0{,}05$
3. Rejeitar a hipótese nula se $t \leq -2{,}228$ ou $t \geq 2{,}228$, onde

$$t = \frac{b - \beta}{s_e/\sqrt{S_{xx}}}$$

   e 2,228 é o valor de $t_{0{,}025}$ para $12 - 2 = 10$ graus de liberdade; caso contrário, aceitar a hipótese nula ou reservar julgamento.
4. Como já sabemos pelos Exemplos 16.1, 16.2 e 16.5 que $S_{xx} = 38.178{,}25$, $b = -0{,}0175$ e $s_e = 0{,}3645$, substituindo esses valores, junto com $\beta = -0{,}02$, fornece

$$t = \frac{-0{,}0175 - (-0{,}02)}{0{,}3645/\sqrt{38.178{,}25}} \approx 1{,}340$$

5. Como $t = 1{,}340$ cai no intervalo de –2,228 a 2,228, a hipótese nula não pode ser rejeitada; não há evidência real para refutar a alegação.

Novamente, poderíamos ter poupado trabalho recorrendo ao impresso de computador da Figura 16.7. Na coluna encabeçada por SE Coef é dado que o erro-padrão estimado de b, que é o

valor que figura no denominador da estatística t, é 0,001865, de forma que podemos escrever diretamente

$$t = \frac{-0{,}0175 - (-0{,}02)}{0{,}001865} = 1{,}340$$

Os testes referentes ao coeficiente de regressão α são feitos de maneira idêntica, com a diferença apenas que utilizamos, em vez da segunda, a primeira das duas estatísticas t. Na maioria das aplicações práticas, entretanto, o coeficiente de regressão α não tem muita importância — é apenas o corte no eixo y, ou seja, o valor de y correspondente a $x = 0$. Em muitos casos não tem qualquer significado real.

Para construir intervalos de confiança para os coeficientes de regressão α e β, substituímos o termo médio de $-t_{\alpha/2} < t < t_{\alpha/2}$ pela estatística t apropriada da página 413. Então, mediante um cálculo algébrico relativamente simples, chegamos às fórmulas

LIMITES DE CONFIANÇA PARA COEFICIENTES DE REGRESSÃO

$$a \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

e

$$b \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{S_{xx}}}$$

onde o grau de confiança é $(1 - \alpha)100\%$ e $t_{\alpha/2}$ é a entrada na Tabela II para $n - 2$ graus de liberdade.

EXEMPLO 16.7 Os dados a seguir mostram os tempos médios semanais, em horas, que seis estudantes dedicaram aos seus trabalhos para casa e os índices de pontuação para as disciplinas que fizeram naquele semestre:

Horas gastas em deveres de casa x	*Índice de pontuação* y
15	2,0
28	2,7
13	1,3
20	1,9
4	0,9
10	1,7

Admitindo que todas as suposições subjacentes à análise de regressão normal tenham sido satisfeitas, construa um intervalo de 99% de confiança para β, a quantidade pela qual um estudante da população amostrada poderia aumentar seu índice de pontuação estudando uma hora extra por semana.

Solução Utilizando o impresso de computador mostrado na Figura 16.9, verificamos que $b = 0{,}06860$ e que a estimativa do erro-padrão de b, pelo qual devemos multiplicar $t_{\alpha/2}$, é 0,01467. Como $t_{0{,}005} = 4{,}604$ para $6 - 2 = 4$ graus de liberdade, obtemos $0{,}0686 \pm 4{,}604(0{,}01467)$ e, portanto,

$$0{,}0011 < \beta < 0{,}1361$$

Esse intervalo de confiança é bastante amplo, e isso se deve a dois fatores — ao tamanho muito pequeno da amostra e à variação relativamente grande medida por s_e, ou seja, a variação entre os índices de pontuação de estudantes sujeitos à mesma quantidade de temas de casa.

Análise de Regressão: *y* contra *x*

```
The regression equation is
C2 = 0.721 + 0.0686 x

Predictor       Coef          SE Coef          T          P
Constant      0.7209           0.2464       2.93      0.043
x            0.06860          0.01467       4.68      0.009

S = 0.2720          R-Sq = 84.5%     R-Sq(adj) = 80.7%
```

Figura 16.9 Impresso de MINITAB para o Exemplo 16.7.

Para responder à segunda questão formulada na página 411, relativa à estimativa, ou previsão, do valor médio de y para um dado valor de x, utilizamos um método semelhante ao que acabamos de discutir. Com as mesmas suposições de antes, baseamos nosso argumento numa outra estatística t, chegando aos seguintes limites de $(1 - \alpha)100\%$ de confiança para $\mu_{y|x_0}$, a média de y quando $x = x_0$:

LIMITES DE CONFIANÇA PARA A MÉDIA DE *y* QUANDO $x = x_0$

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Como anteriormente, o número de graus de liberdade é $n - 2$ e os valores correspondentes de $t_{\alpha/2}$ podem ser lidos na Tabela II.

EXEMPLO 16.8 Reportando-nos novamente aos dados da página 401, suponha que queiramos estimar o alcance auditivo de pessoas que tenham morado nas proximidades de um aeroporto, diretamente na trajetória dos aviões que decolam. Construa um intervalo de 95% de confiança.

Solução Admitindo que todas as suposições subjacentes à análise de regressão normal tenham sido satisfeitas, substituímos $n = 12$, $x_0 = 104$ semanas, $\sum x = 975$ (do Exemplo 16.1) e, portanto, $\bar{x} = 975/12 = 81{,}25$, $S_{xx} = 38.178{,}25$ (do Exemplo 16.2), $a + bx_0 = 13{,}5$ (do Exemplo 16.4), $s_e = 0{,}3645$ (do Exemplo 16.5) e $t_{0{,}025} = 2{,}228$ para $12 - 2 = 10$ graus de liberdade na fórmula precedente do intervalo de confiança, obtendo

$$13{,}5 \pm 2{,}228(0{,}3645)\sqrt{\frac{1}{12} + \frac{(104 - 81{,}25)^2}{38.178{,}25}}$$

e portanto

$$13{,}25 < \mu_{y|x_0} < 13{,}75$$

milhares de ciclos por segundo quando $x = 104$ semanas. (Se tivéssemos usado $a = 15{,}32$, em vez de $a = 15{,}3$ no Exemplo 16.4, teríamos obtido $a + bx_0 = 13{,}50$, em vez de 13,48, que arredondamos para 13,5. Assim, o resultado teria sido o mesmo.)

A terceira questão proposta na página 411 difere das outras duas. Ela não se refere à estimativa de um parâmetro populacional, mas sim à previsão de uma única observação futura. Os extremos de um intervalo para o qual possamos afirmar com um certo grau de confiança que conterá tal observação são chamados **limites de previsão**, e o cálculo desses limites responderá a ter-

ceira questão. Baseando nosso argumento em mais uma outra estatística t, chegamos aos seguintes limites de previsão de $(1-\alpha)100\%$ para um valor de y quando $x = x_0$:

LIMITES DE PREVISÃO

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Novamente, o número de graus de liberdade é $n-2$ e o valor correspondente de $t_{\alpha/2}$ pode ser lido da Tabela II.

Observe que a única diferença entre esses limites de previsão e os limites de confiança para $\mu_{y|x0}$ dados anteriormente é que adicionamos 1 ao radicando. Assim, deixamos a cargo do leitor verificar, no Exercício 16.24, que, para o exemplo do alcance auditivo e $x_0 = 104$, os limites de 95% de previsão são 12,65 e 14,35. Não deveria ser surpresa esse intervalo ser muito mais amplo do que o obtido no Exemplo 16.8. Enquanto que os limites de previsão se aplicam a previsões para uma pessoa, os limites de confiança obtidos no Exemplo 16.8 se aplicam à média de todas pessoas que tenham morado durante dois anos nas proximidades do aeroporto, diretamente na trajetória dos aviões que decolam.

Tenha em mente que todos esses métodos se baseiam nas suposições bastante restritivas da análise de regressão normal. Além disso, se fundamentarmos mais de uma inferência nos mesmos dados, encontraremos problemas relativos aos níveis de significância e/ou aos graus de liberdade. As variáveis aleatórias em que se baseiam os vários processos certamente não são independentes.

EXERCÍCIOS

16.20 Suponha que os dados do Exercício 16.3 satisfaçam as suposições requeridas pela análise de regressão normal.

(a) Se o trabalho no Exercício 16.4 foi feito com um computador, use a informação fornecida pelo aplicativo para testar a hipótese nula $\beta = 0{,}15$ contra a hipótese alternativa $\beta \neq 0{,}15$ ao nível 0,05 de significância.

(b) Para testes relativos ao coeficiente de regressão β, a calculadora gráfica TI-83 fornece o valor de t somente para testes da hipótese nula $\beta = 0$. Como a diferença está somente no numerador, o valor de t para testes da hipótese nula $\beta = \beta_0$ pode ser obtida multiplicando o valor de t dado pela calculadora por

$$\frac{b - \beta_0}{b}$$

Se o trabalho no Exercício 16.4 foi feito com uma calculadora gráfica, use esse método de calcular t para testar a hipótese nula $\beta = 0{,}15$ contra a hipótese alternativa $\beta \neq 0{,}15$ ao nível 0,05 de significância.

16.21 Suponha que os dados do Exercício 16.6 satisfaçam as suposições requeridas pela análise de regressão normal.

(a) Use as somas dadas naquele exercício, $\sum y^2 = 2.001$ e o resultado que $b = 0{,}898$, para calcular o valor de s_e.

(b) Use a informação fornecida na parte (a), bem como seu resultado, para testar a hipótese nula $\beta = 1{,}5$ contra a hipótese alternativa $\beta < 1{,}5$ ao nível 0,05 de significância.

16.22 Use um computador ou uma calculadora gráfica para refazer ambas partes do Exercício 16.21. Se for usada uma calculadora gráfica, siga a sugestão dada na parte (b) do Exercício 16.20.

16.23 Supondo que os dados do Exercício 16.8 satisfaçam as suposições requeridas pela análise de regressão normal, use o resultado daquele exercício e um computador ou uma calcula-

dora gráfica para testar a hipótese nula $\beta = 3{,}5$ contra a hipótese alternativa $\beta > 3{,}5$ ao nível 0,01 de significância. Se for usada uma calculadora gráfica, siga a sugestão dada na parte (b) do Exercício 16.20.

16.24 Com referência aos dados à página 401 e os cálculos no Exemplo 16.8, mostre que para $x_0 = 104$, os limites de 95% de previsão para o alcance auditivo são 12,65 e 14,35 mil ciclos por segundo.

16.25 Com referência ao Exercício 16.9, use um computador ou uma calculadora gráfica para testar a hipótese nula $\beta = -0{,}15$ contra a hipótese alternativa $\beta \neq -0{,}15$ ao nível 0,01 de significância. Deve ser suposto, evidentemente, que os dados do Exercício 16.9 satisfaçam as suposições requeridas pela análise de regressão normal. Também, se for usada uma calculadora gráfica, siga a sugestão dada na parte (b) do Exercício 16.20.

16.26 Com referência ao exercício precedente e com as mesmas suposições, construa um intervalo de 95% de confiança para a redução do resíduo de cloro por hora.

16.27 Suponha que os dados do Exercício 16.12 satisfaçam as suposições requeridas pela análise de regressão normal.

(a) Use as somas dadas naquele exercício, $\sum y^2 = 1.802$ e o resultado que $b = 0{,}272$, para calcular o valor de s_e.

(b) Use a informação fornecida na parte (a), bem como seu resultado, para testar a hipótese nula $\beta = 0{,}40$ contra a hipótese alternativa $\beta < 0{,}40$ ao nível 0,05 de significância.

16.28 Use um computador ou uma calculadora gráfica para refazer ambas partes do Exercício 16.27. Se for usada uma calculadora gráfica, siga a sugestão dada na parte (b) do Exercício 16.20.

16.29 Supondo que os dados do Exercício 16.12 satisfaçam as suposições requeridas pela análise de regressão normal, use um computador ou uma calculadora gráfica para determinar um intervalo de 95% de confiança para o conteúdo médio de umidade quando a umidade relativa é de 50%.

16.30 A tabela a seguir dá os valores, em milhares de unidades monetárias, de avaliação e os preços de venda de oito casas, que constituem uma amostra aleatória de todas as casas vendidas recentemente numa zona rural:

Valor de avaliação	*Preço de venda*
70,3	114,4
102,0	169,3
62,5	106,2
74,8	125,0
57,9	99,8
81,6	132,1
110,4	174,2
88,0	143,5

Supondo que esses dados satisfaçam as suposições requeridas pela análise de regressão normal, use um computador ou uma calculadora gráfica para encontrar

(a) um intervalo de 95% de confiança para o preço de venda médio de uma casa nessa área rural que está avaliada em 90.000 unidades monetárias;

(b) os limites de 95% de previsão para uma casa nessa área rural que foi avaliada em 90.00 unidades monetárias.

16.31 Supondo que os dados do Exercício 16.3 satisfaçam as suposições requeridas pela análise de regressão normal, use um computador ou uma calculadora gráfica para determinar
(a) um intervalo de 99% de confiança para a medida de hiperatividade de uma criança de quatro anos de uma das escolinhas 45 minutos depois de ter ingerido 60 gramas de comida com o conservante;
(b) os limites de 99% de previsão para a medida de hiperatividade de uma dessas crianças que ingeriu 60 gramas de comida com o conservante 45 minutos antes.

16.32 Supondo que os dados do Exercício 16.8 satisfaçam as suposições requeridas pela análise de regressão normal, use um computador ou uma calculadora gráfica para determinar
(a) um intervalo de 98% de confiança para a safra média de trigo quando há uma precipitação pluviométrica de 10 centímetros;
(b) os limites de 98% de previsão para a safra de trigo quando há uma precipitação pluviométrica de 10 centímetros.

16.33 Supondo que os dados do Exercício 16.7 satisfaçam as suposições requeridas pela análise de regressão normal, use um computador ou uma calculadora gráfica para determinar
(a) um intervalo de 95% de confiança para o índice de pontuação de estudantes que dedicaram aos seus trabalhos para casa uma média de cinco horas semanais durante o semestre;
(b) os limites de 95% de previsão para o índice de pontuação de um estudante que dedicou aos seus trabalhos para casa uma média de cinco horas semanais durante o semestre.

16.4 REGRESSÃO MÚLTIPLA

Embora existam muitos problemas em que uma variável pode ser prevista com bastante precisão em termos de outra, é razoável esperar que as previsões devam melhorar levando em conta informações relevantes adicionais. Por exemplo, deveríamos poder fazer melhores previsões sobre o desempenho de professores recém-contratados se levarmos em consideração não somente sua formação, mas também seu tempo de experiência e sua personalidade. Poderemos também fazer melhor previsão do sucesso de um novo livro se considerarmos não só a qualidade do trabalho, mas também o potencial de procura e a concorrência.

Muitas fórmulas matemáticas podem servir para expressar relações entre mais do que duas variáveis, mas as mais comumente usadas em Estatística (em parte por questões de conveniência) são equações lineares da forma

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

Aqui, y é a variável a ser prevista, $x_1, x_2, \ldots$, e x_k são as k variáveis conhecidas, sobre as quais se basearão as previsões, e $b_0, b_1, b_2, \ldots$, e b_k são constantes numéricas a serem determinadas com base nos dados observados.

Para ilustrar, consideremos a seguinte equação, obtida num estudo sobre a demanda por diferentes tipos de carnes:

$$\hat{y} = 3{,}489 - 0{,}090x_1 + 0{,}064x_2 + 0{,}019x_3$$

Aqui, y denota o consumo total de carne bovina e lombo suíno inspecionados: pelas autoridades sanitárias, em milhões de quilogramas, x_1 denota o preço de varejo da carne de gado em centavos

* Esta seção está assinalada como opcional porque os cálculos, embora possivelmente executáveis numa calculadora para problemas muito simples, geralmente exigem aplicativos computacionais especiais.

por quilograma, x_2 denota o preço de varejo do lombo em centavos por quilograma e x_3 é a renda familiar medida pelo índice de certa folha de pagamento. Com essa equação, podemos prever o consumo total de carne de gado e de lombo inspecionados pelas autoridades sanitárias correspondente a valores especificados de x_1, x_2 e x_3.

O problema da determinação de uma equação linear a mais de duas variáveis que melhor descreva determinado conjunto de dados consiste em encontrar valores numéricos de b_0, b_1, b_2, . . ., e b_k. Geralmente, isso é feito pelo método dos mínimos quadrados; isto é, minimiza-se a soma de quadrados $\sum (y - \hat{y})^2$, onde, como anteriormente, os y são os valores observados e os $\hat{y}$ são os valores calculados por meio da equação linear. Em princípio, o problema da determinação de b_0, b_1, b_2, . . ., e b_k é o mesmo que o do caso de duas variáveis, mas as soluções manuais podem ser muito trabalhosas porque o método dos mínimos quadrados exige a resolução de tantas equações normais quantas são as constantes desconhecidas b_0, b_1, b_2, . . ., e b_k. Por exemplo, quando há duas variáveis independentes, x_1 e x_2, e queremos ajustar a equação

$$y = b_0 + b_1 x_1 + b_2 x_2$$

devemos resolver as três equações normais

EQUAÇÕES NORMAIS (DUAS VARIÁVEIS INDEPENDENTES)

$$\sum y = n \cdot b_0 + b_1 \left(\sum x_1\right) + b_2 \left(\sum x_2\right)$$

$$\sum x_1 y = b_0 \left(\sum x_1\right) + b_1 \left(\sum x_1^2\right) + b_2 \left(\sum x_1 x_2\right)$$

$$\sum x_2 y = b_0 \left(\sum x_2\right) + b_1 \left(\sum x_1 x_2\right) + b_2 \left(\sum x_2^2\right)$$

Aqui, $\sum x_1 y$ é a soma dos produtos obtidos multiplicando cada valor dado de x_1 pelo valor correspondente de y, $\sum x_1 x_2$ é a soma dos produtos obtidos multiplicando cada valor dado de x_1 pelo correspondente valor de x_2, e assim por diante.

EXEMPLO 16.9 Os dados a seguir mostram o número de quartos, o número de banheiros e os preços (em unidades monetárias) pelos quais oito casas unifamiliares de um certo bairro foram vendidas recentemente:

Número de quartos x_1	*Número de banheiros* x_2	*Preço* y
3	2	143.800
2	1	109.300
4	3	158.800
2	1	109.200
3	2	154.700
2	2	114.900
5	3	188.400
4	2	142.900

Encontre uma equação linear que permita prever o preço de venda médio de uma casa unifamiliar no bairro dado, em termos do número de quartos e do número de banheiros.

Solução As quantidades necessárias para substituir nas três equações normais são $n = 8$, $\sum x_1 = 25$, $\sum x_2 = 16$, $\sum y = 1.122.000$, $\sum x_1^2 = 87$, $\sum x_1 x_2 = 55$, $\sum x_2^2 = 36$, $\sum x_1 y = 3.711.100$ e $\sum x_2 y = 2.372.700$, resultando

$$1.122.000 = 8b_0 + 25b_1 + 16b_2$$
$$3.711.100 = 25b_0 + 87b_1 + 55b_2$$
$$2.372.700 = 16b_0 + 55b_1 + 36b_2$$

Poderíamos resolver essas equações pelo método da eliminação ou utilizando determinantes, mas em vista dos cálculos bastante extensos, hoje em dia tal tarefa é deixada para os computadores. Assim, recorramos ao impresso de computador da Figura 16.10, onde, na coluna encabeçada por "Coef", vemos que $b_0 = 65.430$, $b_1 = 16.752$ e $b_2 = 11.235$. Na linha imediatamente acima dos coeficientes, vemos que a equação de mínimos quadrados é

$$\hat{y} = 65.430 + 16.752\,x_1 + 11.235\,x_2$$

Isso nos diz que (no bairro dado e na época em que foi feito o estudo), cada quarto adicional acrescentava, em média, 16.752 unidades monetárias e cada banheiro adicional acrescentava 11.235 unidades monetárias ao preço de venda de uma casa.

EXEMPLO 16.10 Com base no resultado do Exemplo 16.9, determine o preço de venda médio de uma casa com três quartos e dois banheiros (no bairro dado na época em que foi feito o estudo).

Solução Substituindo $x_1 = 3$ e $x_2 = 2$ na equação de mínimos quadrados obtida no Exemplo 16.9, obtemos

$$\hat{y} = 65.430 + 16.752(3) + 11.235(2)$$
$$= 138.156$$

ou 138.200 unidades monetárias, aproximadamente.

Figura 16.10 Impresso de MINITAB para o Exemplo 16.9

Análise de Regressão: y contra x1, x2

```
The regression equation is
y = 65430 + 16752  x1 + 11235  x2

Predictor       Coef          SE Coef          T          P
Constant       65430            12134       5.39      0.003
x1             16752             6636       2.52      0.053
x2             11235             9885       1.14      0.307
```

EXERCÍCIOS

***16.34** A seguir estão os dados relativos às idades e aos rendimentos (em unidades monetárias) de uma amostra aleatória de cinco executivos de uma grande companhia multinacional, juntamente com o número de anos de estudos de pós-graduação de cada um:

Idade x_1	*Anos de pós-graduação* x_2	*Renda* y
38	4	181.700
46	0	173.300
39	5	189.500
43	2	179.800
32	4	169.900
52	7	212.500

(a) Use o aplicativo computacional adequado para ajustar uma equação da forma $y = b_0 + b_1x_1 + b_2x_2$ aos dados fornecidos.
(b) Use a equação obtida na parte (a) para estimar a renda média de um executivo da companhia multinacional de 39 anos de idade que fez três anos de pós-graduação uma universidade.

***16.35** Os dados a seguir foram coletados para determinar a relação entre duas variáveis de processamento e a dureza de certo tipo de aço:

Dureza (Rockwell 30 T) y	*Conteúdo de cobre (percentagem)* x_1	*Temperatura de temperar (graus Fahrenheit)* x_2
78,9	0,02	1.000
55,2	0,02	1.200
80,9	0,10	1.000
57,4	0,10	1.200
85,3	0,18	1.000
60,7	0,18	1.200

(a) Use aplicativo computacional adequado para ajustar uma equação da forma $y = b_0 + b_1x_1 + b_2x_2$ aos dados fornecidos.
(b) Use a equação obtida na parte (a) para estimar a dureza do aço quando seu conteúdo de cobre é 0,14% e a temperatura em que é temperado é de 1.100 graus Fahrenheit.

***16.36** Quando os x_1 e/ou os x_2 estão igualmente espaçados, o cálculo dos coeficientes de regressão pode ser consideravelmente simplificado usando o tipo de codificação descrito no Exercício 16.17. Refaça o Exercício 16.35 sem usar computador após codificar como –1, 0 e 1 os três valores de x_1, e como –1 e 1 os dois valores de x_2. (Note que, codificado, o conteúdo de 0,14% de cobre passa a ser 0,50 e a temperatura de 1.100 graus Fahrenheit em que é temperado passa a ser 0.)

***16.37** Os dados a seguir representam as eficácias percentuais de um analgésico e a quantidade (em miligramas) de três medicamentos presentes em cada cápsula:

Medicamento A x_1	*Medicamento B* x_2	*Medicamento C* x_3	*Eficácia percentual* y
15	20	10	47
15	20	20	54
15	30	10	58
15	30	20	66
30	20	10	59
30	20	20	67
30	30	10	71
30	30	20	83
45	20	10	72
45	20	20	82
45	30	10	85
45	30	20	94

(a) Use aplicativo computacional adequado para ajustar uma equação da forma $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$ aos dados fornecidos.

(b) Use a equação obtida na parte (a) para estimar a eficácia percentual média de cápsulas contendo 12,5 miligramas do Medicamento A, 25 miligramas do medicamento B e 15 miligramas do Medicamento C.

***16.38** Refaça o Exercício 16.37 sem usar computador após codificar como – 1, 0 e 1 os três valores de x_1, como –1 e 1 os dois valores de x_2 e como –1 e 1 os dois valores de x_3.

*16.5 REGRESSÃO NÃO-LINEAR

Quando o padrão de um conjunto de dados se afasta consideravelmente de uma reta, precisamos considerar ajustar algum outro tipo de curva. Nesta seção, descreveremos primeiro dois casos em que a relação entre x e y não é linear, mas nos quais, mesmo assim, é possível aplicar o método da Seção 16.2. Em seguida, daremos um exemplo de **ajuste de uma curva polinomial**, ajustando uma parábola.

Em geral, esboçamos pares de dados em vários tipos de papel de gráfico para ver se há escalas em que os pontos se aproximem de uma reta. Naturalmente, quando isso ocorre no papel de gráfico comum, procedemos como na Seção 16.2. Se isso ocorre quando usamos o **papel de gráfico semilog** (com subdivisões iguais para x e uma escala logarítmica para y, conforme indica a Figura 16.11), isso indica que uma **curva exponencial** dará um bom ajuste. A equação de uma tal curva é

$$y = a \cdot b^x$$

ou, em forma logarítmica,

$$\log y = \log a + x(\log b)$$

onde "log" representa o logaritmo de base 10. (Na verdade, poderíamos usar qualquer base, inclusive o número irracional e, caso em que a equação costuma ser escrita como $y = a \cdot e^{bx}$ ou, em formato logarítmico, como $\ln y = \ln a + bx$.)

Observe que, se representamos $\log a$ por A, $\log b$ por B e $\log y$ por Y, a equação original em formato logarítmico será escrita como $Y = A + Bx$, que é a equação usual de uma reta. Assim, para ajustar uma curva exponencial a um dado conjunto de pares de dados, simplesmente aplicamos o método da Seção 16.2 aos pares de dados (x, Y).

EXEMPLO 16.11 Os dados referentes aos lucros líquidos (em milhares de unidades monetárias) de uma companhia durante os seis primeiros anos de operação são os seguintes:

Ano	*Lucro líquido*
1	112
2	149
3	238
4	354
5	580
6	867

Na Figura 16.11, esboçamos esses dados em papel de gráfico comum no lado esquerdo e em papel de gráfico semilog (com escala logarítmica para y) no lado direito. Como pode ser visto, o padrão global está visivelmente "linearizado" na figura da direita, e isso sugere que deveríamos ajustar uma curva exponencial.

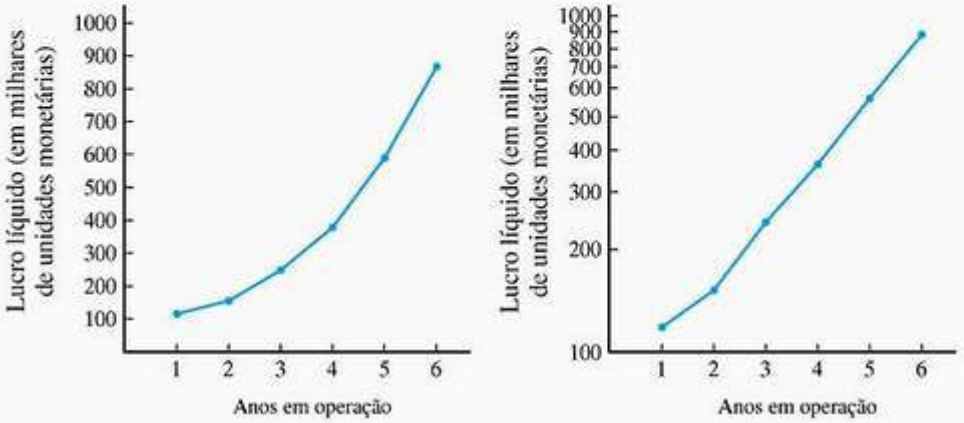

Figura 16.11
Dados esboçados em papel de gráfico comum e semilog.

Solução Obtendo os logaritmos de y com uma calculadora ou, talvez, de uma tabela de logaritmos, obtemos

x	y	$Y = \log y$
1	112	2,0492
2	149	2,1732
3	238	2,3766
4	354	2,5490
5	580	2,7634
6	867	2,9380

Assim, para esses dados, obtemos $n = 6$, $\sum x = 21$, $\sum x^2 = 91$, $\sum Y = 14{,}8494$ e $\sum xY = 55{,}1664$ e, portanto, $S_{xx} = 91 - \frac{1}{6}(21)^2 = 17{,}5$ e $S_{xy} = 55{,}1664 - \frac{1}{6}(21)(14{,}8494) = 3{,}1935$. Finalmente, a substituição nas duas fórmulas à página 404 fornece

$$B = \frac{3{,}1935}{17{,}5} \approx 0{,}1825$$

$$A = \frac{14{,}8494 - 0{,}1825(21)}{6} \approx 1{,}8362$$

e a equação que descreve a relação é

$$\hat{Y} = 1{,}8362 + 0{,}1825x$$

Como 1,8362 e 0,1825 são as estimativas correspondentes a $\log a$ e $\log b$, vemos, tomando antilogaritmos, que $a = 68{,}58$ e $b = 1{,}52$. Assim, a equação da curva exponencial que melhor descreve a relação entre o lucro líquido da companhia e o número de anos de operação é

$$\hat{y} = 68{,}58(1.52)^x$$

onde $\hat{y}$ é dado em milhares de unidades monetárias.

Embora os cálculos no Exemplo 16.11 tenham sido bastante fáceis, poderíamos, é claro, ter usado um computador. Digitando os valores de x e de Y em colunas c1 e c2, obtemos o impresso mostrado na Figura 16.12. Como pode ser visto, os valores que calculamos para A e B aparecem na coluna encabeçada por "Coef".

Para obter a equação exponencial em seu formato final, teria sido mais fácil ainda usar uma calculadora gráfica. Depois de digitar os x e y originais, o comando **STAT CALC ExpReg** fornece os dados mostrados na Figura 16.13. Como pode ser visto, as constantes a e b, arredondadas até a segunda casa decimal, são idênticas às da equação exponencial dada no fim do Exemplo 16.11.

Análise de Regressão: Y = log y contra x

```
The regression equation is
Y = log y = 1.84 + 0.182  x

Predictor       Coef          SE Coef          T          P
Constant     1.83620         0.02243      81.85      0.000
x            0.182486        0.005760     31.68      0.000

S = 0.02410          R-Sq = 99.6%     R-Sq(adj) = 99.5%
```

Figura 16.12 Impresso de computador para o Exemplo 16.11.

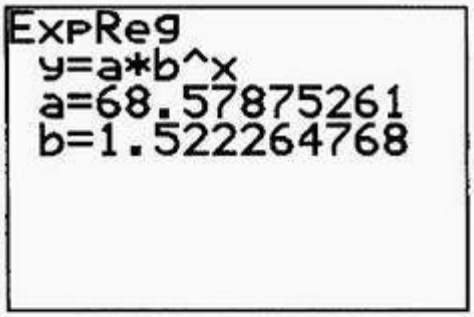

Figura 16.13 Valores de a e b reproduzidos da janela de uma calculadora científica TI-83.

Uma vez ajustada uma curva exponencial a um conjunto de pares de dados, podemos prever um valor futuro de y por substituição em sua equação do valor correspondente de x. Contudo, em geral é muito mais conveniente substituir x na forma logarítmica da equação, ou seja, em

$$\log \hat{y} = \log a + x(\log b)$$

EXEMPLO 16.12 Com referência ao Exemplo 16.11, preveja o lucro líquido da companhia em seu oitavo ano de operação.

Solução Substituindo $x = 8$ na forma logarítmica da equação para a curva exponencial, obtemos

$$\log \hat{y} = 1{,}8362 + 8(0{,}1825)$$
$$= 3{,}2962$$

e, portanto, $\hat{y} = 1.980$, ou 1.980.000 unidades monetárias.

Se os pontos de dados caem perto de uma reta quando esboçados em **papel de gráfico log-log** (com escalas logarítmicas para x e y), isso indica que uma equação da forma

$$y = a \cdot x^b$$

fornecerá um bom ajuste. Em forma logarítmica, a equação de uma tal **função potência** é

$$\log y = \log a + b(\log x)$$

que é uma equação linear em $\log x$ e $\log y$. (Representando $\log a$, $\log x$ e $\log y$ por A, X e Y, respectivamente, a equação é dada por

$$Y = A + bX$$

que é a equação usual de uma reta.) Para ajustar uma curva potência, podemos, portanto, aplicar o método da Seção 16.2 ao problema descrito por $Y = A + bX$. O trabalho necessário para ajustar

uma função potência é análogo ao que tivemos no Exemplo 16.11, e não o ilustraremos aqui por meio de exemplo. Contudo, nos Exercícios 16.47 e 16.49, o leitor encontrará problemas nos quais o método pode ser aplicado.

Quando os valores de y primeiro crescem e depois decrescem, ou primeiro decrescem e depois crescem, muitas vezes um bom ajuste é dado pela **parábola** de equação

$$y = a + bx + cx^2$$

Essa equação também pode ser escrita como

$$y = b_0 + b_1x + b_2x^2$$

para ficar de acordo com a notação da Seção 16.4. Assim, pode ser visto que podemos considerar parábolas como equações lineares nas duas incógnitas $x_1 = x$ e $x_2 = x^2$ e que ajustar uma parábola a um conjunto de pares de dados não constitui novidade — basta aplicarmos o método da Seção 16.4. Se realmente quiséssemos usar as equações normais da página 403 com $x_1 = x$ e $x_2 = x^2$, isso requereria a determinação de $\sum x$, $\sum x^2$, $\sum x^3$, $\sum x^4$, $\sum y$, $\sum xy$, e $\sum x^2y$, e subseqüente resolução simultânea de três equações lineares. Como pode ser imaginado, isso exigiria uma grande quantidade de contas e raramente é feito sem o uso de aplicativos apropriados. Nos dois exemplos a seguir, vamos primeiro ilustrar o ajuste de uma parábola utilizando um computador e depois repetir o problema usando uma calculadora gráfica.

EXEMPLO 16.13 O tempo de secagem (em horas) de um verniz e a quantidade (em gramas) de certo aditivo químico são os seguintes:

Quantidade de aditivo x	*Tempo de secagem* y
1	7,2
2	6,7
3	4,7
4	3,7
5	4,7
6	4,2
7	5,2
8	5,7

(a) Ajuste uma parábola que, conforme sugere a Figura 16.14, é o tipo correto de curva para ajustar os dados fornecidos.

(b) Use o resultado da parte (a) para prever o tempo de secagem do verniz quando se adicionam 6,5 gramas do aditivo químico.

Solução (a) Usando o impresso de MINITAB mostrado na Figura 16.15, verificamos que $b_0 = 9{,}2446$, $b_1 = -2{,}0149$ e $b_2 = 0{,}19940$ (na coluna encabeçada por "Coef").

Arredondando até a segunda casa decimal, podemos, portanto, escrever a equação da parábola como

$$\hat{y} = 9{,}24 - 2{,}01x + 0{,}20x^2$$

(Observe que digitamos os valores de x na coluna c1, os valores de x^2 na coluna c2 e os valores de y na coluna c3.)

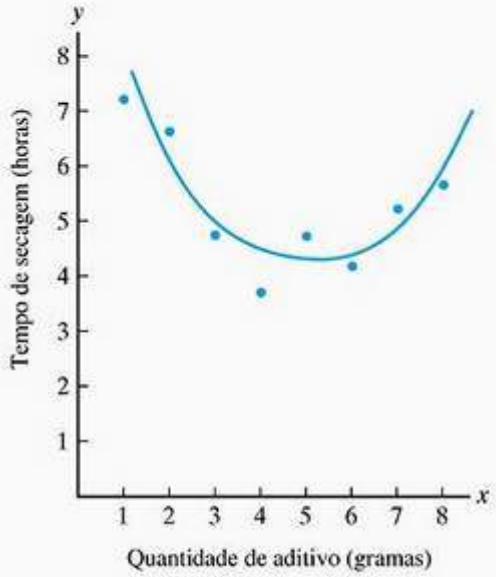

Figura 16.14
Diagrama de dispersão dos dados de secagem de verniz.

(b) Substituindo $x = 6{,}5$ na equação obtida na parte (a), resulta

$$\hat{y} = 9{,}24 - 2{,}01(6{,}5) + 0{,}20(6{,}5)^2$$
$$\approx 4{,}62 \text{ horas}$$

EXEMPLO 16.14 Refaça o Exemplo 16.13 utilizando uma calculadora gráfica.

Solução Para evitar confusão, chamamos a atenção para o fato de que a TI-83 utiliza a equação $y = ax^2 + bx + c$, com a e c permutados em relação à versão dada na página 425. Depois de digitarmos os x e os y, o comando **STAT CALC QuadReg** fornece o resultado que aparece na Figura 16.16. Arredondando para duas casas decimais, obtemos

$$\hat{y} = 0{,}20x^2 - 2{,}01x + 9{,}24$$

que é idêntico ao obtido anteriormente, exceto pela ordem das parcelas.

Na página 425, introduzimos as parábolas como curvas que se curvam uma única vez — isto é, seus valores primeiro crescem e depois decrescem, ou primeiro decrescem e depois cres-

Análise de Regressão: y contra x, x2

```
The regression equation is
y = 9.24 - 2.01 x + 0.199 x2

Predictor     Coef   SE Coef        T       p
Constant   9.2446    0.7645    12.09   0.000
x         -2.0149    0.3898    -5.17   0.004
x2        0.19940   0.04228     4.72   0.005

S = 0.5480    R-Sq = 85.3%  R-Sq(adj) = 79.4%
```

Figura 16.15
Impresso de computador para o ajuste de parábola.

Figura 16.16
Ajuste de parábola reproduzido na janela de uma calculadora gráfica TI-83.

cem. Para padrões que se curvam mais de uma vez, **equações polinomiais** de grau mais alto, tais como $y = a + bx + cx^2 + dx^3$ ou $y = a + bx + cx^2 + dx^3 + ex^4$, podem ser ajustadas pela técnica ilustrada no Exemplo 16.13. Na prática, costumamos trabalhar com partes dessas curvas, em particular partes de parábolas, quando há apenas um pequeno encurvamento no padrão que pretendemos descrever.

EXERCÍCIOS

***16.39** Os dados a seguir referem-se ao crescimento de uma colônia de bactérias num meio de cultura:

Dias desde a inoculação x	*Contagem de bactérias (milhares)* y
2	112
4	148
6	241
8	363
10	585

(a) Sabendo que $\sum x = 30$, $\sum x^2 = 220$, $\sum Y = 11{,}9286$ (onde $Y = \log y$) e $\sum xY = 75{,}2228$, monte as duas equações normais para o ajuste de uma curva exponencial, resolva-as em $\log a$ e $\log b$ pelo método da eliminação ou por determinantes, e escreva a equação da curva em formato logarítmico.
(b) Transforme a equação obtida na parte (a) para o formato $y = ab^x$.
(c) Use a equação obtida na parte (a) para estimar a contagem de bactérias cinco dias depois da inoculação.

***16.40** Use um computador ou uma calculadora gráfica para refazer o Exercício 16.39.

***16.41** Use um computador ou uma calculadora gráfica para ajustar uma curva exponencial aos dados seguintes, relativos à percentagem de pneus radiais fabricados por certa indústria e que ainda estão em condições de uso após terem rodado o número de quilômetros indicado:

Quilômetros rodados (em milhares) x	*Percentagem utilizável* y
1	97,2
2	91,8
5	82,5
10	64,4
20	41,0
30	29,9
40	17,6
50	11,3

***16.42** Transforme a equação obtida no Exercício 16.41 para o formato logarítmico e use-a para estimar a percentagem dos pneus que ainda estará utilizável após rodar 25.000 quilômetros.

***16.43** Use um computador ou uma calculadora gráfica para ajustar uma curva exponencial aos dados seguintes, relativos ao tempo de secagem x de amostras de teste de concreto e sua força de tensão y:

x *(horas)*	y *(pascals)*
1	3,54
2	8,92
3	27,5
4	78,8
5	225
6	639

***16.44** Supondo que a tendência exponencial continue, use a equação obtida no Exercício 16.43 (no formato logarítmico) para estimar a força de tensão de uma amostra de teste de concreto com tempo de secagem de oito horas.

***16.45** Um pedaço pequeno de um cacto raro e de crescimento lento foi enxertado num outro cacto dotado de raízes bem desenvolvidas, medindo-se sua altura anualmente, como mostra a tabela seguinte:

Anos depois do enxerto x	*Altura (milímetros)* y
1	22
2	25
3	29
4	34
5	38
6	44
7	51
8	59
9	68

Use um computador ou uma calculadora gráfica para ajustar uma curva exponencial.

*16.46 Os dados a seguir representam a quantidade x (em quilograma por metro quadrado) de fertilizante aplicado ao solo e a safra (em quilogramas por metro quadrado) de um certo alimento:

x	y
0,5	32,0
1,1	34,3
2,2	15,7
0,2	20,8
1,6	33,5
2,0	21,5

(a) Esboce um diagrama de dispersão para verificar que é razoável descrever o padrão global dos dados com uma parábola.
(b) Use um computador ou uma calculadora gráfica para ajustar uma parábola aos dados fornecidos.
(c) Use a equação obtida na parte (b) para estimar a safra quando se aplicam 1,5 quilogramas do fertilizante por metro quadrado.

*16.47 Os dados a seguir se referem à demanda y (em milhares de unidades) por um produto e seu preço x (em unidades monetárias) em cinco áreas comerciais bastante semelhantes:

Preço x	*Demanda* y
20	22
16	41
10	120
11	89
14	56

Use um computador ou uma calculadora gráfica para ajustar uma parábola a esses dados, e use-a para estimar a demanda quando o preço do produto é de 12 unidades monetárias.

16.6 LISTA DE TERMOS-CHAVE (com indicação das páginas de suas definições)

*Ajuste de curva polinomial, 422
Ajuste de curvas, 399
Análise de regressão, 410
Análise de regressão linear, 412
Análise de regressão normal, 412
Coeficientes de regressão, 411
Coeficientes de regressão estimados, 411
*Curva exponencial, 422
Determinantes, 404
Diagrama de dispersão, 401
Equação de regressão, 405
Equação linear, 399
*Equação polinomial, 427
Equações normais, 403
Erro-padrão da estimativa, 412
*Função potência, 424
Limites de previsão, 415
Método de eliminação, 404
Método dos mínimos quadrados, 398, 401
*Papel de gráfico log-log, 424
Papel de gráfico semilog, 422
*Parábola, 425
Pontos de dados, 401, 403
Regressão, 398
*Regressão múltipla, 418
Regressão não-linear, 422
Reta de mínimos quadrados, 402
Reta de regressão, 406
Reta de regressão estimada, 406

16.7 REFERÊNCIAS

Em livros de Análise Numérica e textos mais avançados de Estatística, podem ser encontrados métodos para decidir qual tipo de curva pode ser ajustado a um conjunto de pares de dados. Informações adicionais sobre o conteúdo deste capítulo podem ser encontradas em

CHATTERJEE, S., and PRICE, B., *Regression Analysis by Example*, 2nd ed. New York: John Wiley & Sons, Inc., 1991.

DANIEL, C., and WOOD, F., *Fitting Equations to Data*, 2nd ed. New York: John Wiley & Sons, Inc. 1980.

DRAPER, N. R., and SMITH, H., *Applied Regression Analysis*, 2nd ed., New York: John Wiley & Sons, Inc. 1981.

EZEKIEL, M., and FOX, K. A., *Methods of Correlation and Regression Analysis*, 3rd ed. New York: John Wiley & Sons, Inc., 1959.

WEISBERG, S., *Applied Linear Regression*, 2nd ed. New York: John Wiley & Sons, Inc., 1985.

WONNACOTT, T. H. and WONNACOTT, R. J., *Regression: A Second Course in Statistics*. New York: John Wiley & Sons, Inc., 1981.

17

CORRELAÇÃO

Tendo estudado como ajustar uma reta de mínimos quadrados a pares de dados, voltamo-nos, agora, para o problema de determinar quão bom é o ajuste de uma tal reta aos dados. Naturalmente, podemos ter alguma idéia disso observando um diagrama de dispersão que exiba a reta juntamente com os dados, mas para mostrar como podemos ser mais objetivos, vamos voltar aos dados que utilizamos para ilustrar o ajuste de uma reta de mínimos quadrados; mais precisamente, ao exemplo do alcance auditivo de pessoas expostas ao ruído de decolagem de aviões durante um certo período de tempo:

Número de semanas x	*Alcance auditivo* y
47	15,1
56	14,1
116	13,2
178	12,7
19	14,6
75	13,8
160	11,9
31	14,8
12	15,3
164	12,6
43	14,7
74	14,0

Como pode ser visto nessa tabela, há diferenças substanciais entre os y, o menor dos quais é 11,9 e o maior é 15,3. Contudo, vemos também que o alcance auditivo de 11,9 mil ciclos por segundo

foi o de uma pessoa que morou naquele local por 160 semanas, enquanto que o alcance auditivo de 15,3 mil ciclos por segundo foi o de uma pessoa que morou naquele local por apenas 12 semanas. Isso sugere que as diferenças no alcance auditivo podem muito bem ser devidas, pelo menos parcialmente, às diferenças de tempo em que as pessoas ficaram expostas ao ruído do aeroporto. Isso suscita a seguinte questão, que será respondida neste capítulo: da variação total entre os y, quanto pode ser atribuído à relação entre as duas variáveis x e y (ou seja, ao fato de que os valores observados de y corresponderem a diferentes valores de x) e quanto pode ser atribuído ao acaso?

Na Seção 17.1, introduzimos o coeficiente de correlação como uma medida da intensidade da relação linear entre duas variáveis, na Seção 17.2, vemos como interpretá-lo e, na Seção 17.3, abordamos problemas correlatos de inferência. Os problemas de correlação múltipla e parcial serão tratados sucintamente na Seção 17.4, que é opcional.

17.1 O COEFICIENTE DE CORRELAÇÃO

Para responder a questão levantada na abertura do capítulo, vamos ressaltar que estamos diante de uma análise da variância. A Figura 17.1 mostra o que queremos dizer. Como pode ser visto no diagrama, o desvio de um valor observado de y em relação à média de todos os y, $y - \bar{y}$, pode ser escrito como a soma de duas parcelas. Uma parcela é o desvio de $\hat{y}$ (o valor na reta correspondente a um valor observado de x) a partir da média de todos os y, $\hat{y} - \bar{y}$; a outra parcela é o desvio do valor observado de y a partir do valor correspondente na reta, $y - \hat{y}$. Simbolicamente, escrevemos

$$y - \bar{y} = (\hat{y} - \bar{y}) + (y - \hat{y})$$

para qualquer valor observado y e, elevando ao quadrado ambos os membros dessa identidade e somando sobre todos os n valores de y, verificamos que algumas simplificações algébricas levam a

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

A quantidade à esquerda mede a variação total dos y e é denominada a **soma de quadrados total**; note que $\sum (y - \bar{y})^2$ é simplesmente a variância dos y multiplicada por $n - 1$. O primeiro dos dois somatórios à direita, $\sum (\hat{y} - \bar{y})^2$, é denominado a **soma de quadrados de regressão** e mede a parcela da variação total dos y que pode ser atribuída à relação entre as duas variáveis x e y; de fato, se todos os pontos estão sobre a reta de mínimos quadrados, então $y = \hat{y}$ e a soma de quadrados de regressão é igual à soma de quadrados total. Na prática, isso ocorre raramente, se é que ocorre, e o fato de que os pontos não estão todos sobre uma reta de mínimos quadrados é uma indicação de que há outros fatores, além das diferenças entre os x, que afetam os valores de y. Costuma-se combinar todos esses outros fatores sob uma rubrica geral de "acaso". A variação de-

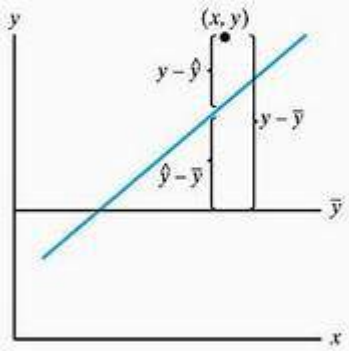

Figura 17.1
Ilustração para mostrar que $y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$.

vida ao acaso é, pois, medida pelos desvios dos pontos em relação à reta; especificamente, é medida por $\sum (y - \hat{y})^2$, denominado **soma de quadrados residual**, que é a segunda das duas parcelas em que dividimos a soma de quadrados total.

Para determinar essas somas de quadrados para o exemplo do alcance auditivo, poderíamos substituir os valores observados de y, ou seja, $\bar{y}$, e os valores de $\hat{y}$ obtidos pela substituição dos x em $\hat{y} = 15{,}3 - 0{,}0175x$ (ver página 404), mas há simplificações. Em primeiro lugar, para $\sum (y - \bar{y})^2$, temos a fórmula de cálculo

$$S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

e na página 412 mostramos que essa expressão é igual a 13,02 no nosso exemplo. Em segundo lugar, $\sum (y - \hat{y})^2$ é a quantidade que minimizamos pelo método dos mínimos quadrados e, dividido por $n - 2$, define s_e^2 à página 412. Assim, essa expressão é igual a $(n - 2)s_e^2$ e $(12 - 2)(0{,}3645)^2 \approx 1{,}329$ em nosso exemplo, para o qual foi mostrado no Exemplo 16.5 que $s_e = 0{,}3645$. Finalmente, subtraindo, a soma de regressão de quadrados é dada por

$$\sum(\hat{y} - \bar{y})^2 = \sum(y - \bar{y})^2 - \sum(y - \hat{y})^2$$

e, para nosso exemplo, obtemos 13,02 – 1,329 = 11,69 (arredondado até a segunda casa decimal).

É interessante observar que todas as somas de quadrados que calculamos aqui poderiam ter sido obtidas diretamente do impresso de computador da Figura 16.7, reproduzido na Figura 17.2. Sob o título geral "Analysis of Variance", na coluna encabeçada por SS, encontramos que a soma de quadrados total é 13,020, a soma de quadrados de erro (residual) é 1,329 e a soma de quadrados de regressão é 11,691. As pequenas diferenças entre os valores mostrados aqui e os calculados anteriormente devem-se ao arredondamento.

Estamos agora em condições de examinar as somas de quadrados. Comparando a soma de quadrados de regressão com a soma de quadrados total, vemos que

$$\frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{11{,}69}{13{,}02} \approx 0{,}898$$

é a proporção da variação total dos alcances auditivos que pode ser atribuída à relação com x, isto é, às diferenças entre a duração do tempo em que as 12 pessoas da amostra foram expostas ao ruído do aeroporto. Essa quantidade é denominada o **coeficiente de determinação** e é denotada por r^2. Observe que o coeficiente de determinação também é dado no impresso da Figura 17.2, onde diz, perto do meio, que R-sq = 89,8%.

Tomando a raiz quadrada do coeficiente de determinação, obtemos o **coeficiente de correlação**, denotado pela letra r. Seu sinal é escolhido de modo a igualar o do coeficiente de regressão estimado b e, para nosso exemplo, em que b é negativo, obtemos

$$r = -\sqrt{0{,}898} \approx -0{,}95$$

arredondado até a segunda casa decimal.

Segue que o coeficiente de correlação é positivo quando a reta de mínimos quadrados tem inclinação para cima, isto é, quando a relação entre x e y é tal que valores pequenos de y tendem a corresponder a valores pequenos de x e valores grandes de y tendem a corresponder a valores grandes de x. Do mesmo modo, o coeficiente de correlação é negativo quando a reta de mínimos quadrados tem inclinação para baixo, isto é, quando valores grandes de y tendem a corresponder a valores pequenos de x e valores pequenos de y tendem a corresponder a valores grandes de x. Exemplos de **correlação positiva** e de **correlação negativa** estão exibidos nos dois primeiros diagramas da Figura 17.3.

Análise de Regressão: y contra x

```
The regression equation is
y = 15.3 - 0.0175 x

Predictor        Coef     SE Coef        T        P
Constant      15.3218      0.1845    83.04    0.000
x          -0.017499    0.001865    -9.38    0.000

S = 0.3645    R-Sq = 89.8%  R-Sq(adj) = 88.8%

Analysis of Variance

Source            DF       SS       MS        F        p
Regression         1   11.691   11.691    88.00    0.000
Residual Error    10    1.329    0.133
Total             11   13.020
```

Figura 17.2
Cópia da Figura 16.7.

Como uma parte da variação dos y não pode exceder sua variação total, $\sum (y - \hat{y})^2$ não pode exceder $\sum (y - \overline{y})^2$ e, da fórmula que define r, decorre que o coeficiente de correlação deve situar-se no intervalo de -1 a $+1$. Se todos os pontos estão sobre uma linha reta, a soma de quadrados residual, $\sum (y - \hat{y})^2$, é zero,

$$\sum (\hat{y} - \overline{y})^2 = \sum (y - \overline{y})^2$$

e o valor resultante de r, que é -1 ou $+1$, indica um ajuste perfeito. Entretanto, se a dispersão dos pontos é tal que a reta de mínimos quadrados é uma reta horizontal coincidindo com $\overline{y}$ (ou seja, uma reta com inclinação nula que corta o eixo y em $a = \overline{y}$), então

$$\sum (y - \hat{y})^2 = \sum (y - \overline{y})^2 \quad \text{e} \quad r = 0$$

Nesse caso, nenhuma das variações dos y pode ser atribuída à sua relação com x, e o ajuste é tão pobre que o conhecimento de x em nada contribui para a previsão de y. O valor previsto de y é $\overline{y}$, independentemente de x. Um exemplo disso é mostrado no terceiro diagrama da Figura 17.3.

A fórmula que define r mostra claramente a natureza, ou essência, do coeficiente de correlação, mas, na prática, raramente é utilizada para determinar seu valor. Em vez disso, utilizamos a fórmula de cálculo

FÓRMULA DE CÁLCULO DO COEFICIENTE DE CORRELAÇÃO

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

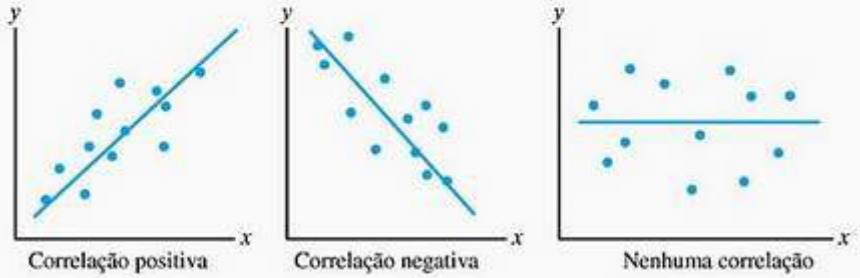

Figura 17.3
Tipos de correlação.

que tem a vantagem adicional de dar, automaticamente, o sinal de r. As quantidades necessárias para calcular r com essa fórmula foram definidas anteriormente, mas por razões de conveniência vamos lembrar o leitor de que

$$S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

$$S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

EXEMPLO 17.1 A seguir, temos as notas que 12 estudantes obtiveram em exames finais de Economia e Antropologia:

Economia	*Antropologia*
51	74
68	70
72	88
97	93
55	67
73	73
95	99
74	73
20	33
91	91
74	80
80	86

Use a fórmula de cálculo para calcular r.

Solução Calculando inicialmente os somatórios necessários, obtemos $\sum x = 850$, $\sum x^2 = 65.230$, $\sum y = 927$, $\sum y^2 = 74.883$ e $\sum xy = 69.453$. Em seguida, substituindo esses valores, juntamente com $n = 12$, nas fórmulas de S_{xx}, S_{yy} e S_{xy}, resulta

$$S_{xx} = 65.230 - \frac{1}{12}(850)^2 \approx 5.021{,}67$$

$$S_{yy} = 74.883 - \frac{1}{12}(927)^2 = 3.272{,}25$$

$$S_{xy} = 69.453 - \frac{1}{12}(850)(927) = 3.790{,}5$$

e

$$r = \frac{3.790{,}5}{\sqrt{(5.021{,}67)(3.272{,}25)}} \approx 0{,}935$$

A expressão S_{xy} no numerador da fórmula para r é, na verdade, uma fórmula de cálculo para $\sum (x - \bar{x})(y - \bar{y})$ que, dividida por n, é denominada o primeiro **momento produto**. Por essa razão, às vezes r é denominado também de **coeficiente de correlação do momento produto**. Note que, em $\sum (x - \bar{x})(y - \bar{y})$, somamos os produtos obtidos pela multiplicação dos desvios de cada x a partir de $\bar{x}$ pelos desvios de cada y a partir de $\bar{y}$. Dessa forma, medimos, literalmente, como os x e os y variam

em conjunto. Se sua relação é tal que valores grandes de x tendem a corresponder a valores grandes de y e valores pequenos de x a valores pequenos de y, então os desvios $x - \bar{x}$ e $y - \bar{y}$ tendem a ser ambos positiva ou ambos negativos, e a maioria dos produtos $(x - \bar{x})(y - \bar{y})$ será positiva. Por outro lado, se a relação é tal que valores grandes de x tendem a corresponder a valores pequenos de y e valores pequenos de x a valores grandes de y, então os desvios $x - \bar{x}$ e $y - \bar{y}$ tendem a ter sinais opostos, e a maioria dos produtos $(x - \bar{x})(y - \bar{y})$ será negativa. Por essa razão, $\sum (x - \bar{x})(y - \bar{y})$ dividido por $n - 1$ é denominada a **covariância amostral**.

Os coeficientes de correlação são calculados, às vezes, na análise de tabelas $r \times c$, desde que estejam ordenadas as categorias de linha, bem como as categorias de coluna. Esse é o tipo de alternativa à análise qui-quadrado que sugerimos ao final da Seção 14.4, onde destacamos que o ordenamento das categorias não era levado em conta no cálculo da estatística χ^2. Para usar r num problema como esse, substituímos as categorias ordenadas por conjuntos de números ordenados analogamente. Como já observamos à página 346, escolhemos números que geralmente são inteiros consecutivos, de preferência inteiros que simplifiquem a aritmética ao máximo, embora isso não seja necessário. Para três categorias, poderíamos usar 1, 2 e 3, ou então –1, 0 e 1; para quatro categorias, poderíamos usar 1, 2, 3 e 4, ou então –1, 0, 1 e 2, ou, quem sabe, –3, –1, 1 e 3. O cálculo de r como uma medida da intensidade da relação entre duas variáveis categóricas é ilustrado pelo exemplo seguinte.

EXEMPLO 17.2 No Exemplo 14.7, analisamos a seguinte tabela 3 × 3 para ver se há alguma relação entre as notas nos testes de qualificação de pessoas que fizeram certo programa de treinamento e seu desempenho subseqüente no emprego

		Desempenho			
		Fraco	*Razoável*	*Bom*	
	Abaixo da média	67	64	25	156
Notas no teste	*Média*	42	76	56	174
	Acima da média	10	23	37	70
		119	163	118	400

Rotule as notas obtidas no teste por $x = -1$, $x = 0$ e $x = 1$, as avaliações do desempenho por $y = -1$, $y = 0$ e $y = 1$, e calcule r.

Solução Rotulando as linhas e colunas conforme indicado, obtemos

		y			
		–1	*0*	*1*	
	–1	67	64	25	156
x	*0*	42	76	56	174
	1	10	23	37	70
		119	163	118	400

onde os totais de linhas nos dizem quantas vezes x é igual a -1, 0 e 1, e os totais de colunas nos dizem quantas vezes y é igual a -1, 0 e 1. Assim,

$$\sum x = 156(-1) + 174 \cdot 0 + 70 \cdot 1 = -86$$
$$\sum x^2 = 156(-1)^2 + 174 \cdot 0^2 + 70 \cdot 1^2 = 226$$
$$\sum y = 119(-1) + 163 \cdot 0 + 118 \cdot 1 = -1$$
$$\sum y^2 = 119(-1)^2 + 163 \cdot 0^2 + 118 \cdot 1^2 = 237$$

e, para $\sum xy$, precisamos somar os produtos obtidos multiplicando cada freqüência de célula pelos correspondentes valores de x e y. Omitindo todas células em que ou $x = 0$ ou $y = 0$, obtemos

$$\sum xy = 67(-1)(-1) + 25(-1)1 + 10 \cdot 1(-1) + 37 \cdot 1 \cdot 1$$
$$= 69$$

Então, substituindo nas formulas de S_{xx}, S_{yy}, e S_{xy}, resulta

$$S_{xx} = 226 - \frac{1}{400}(-86)^2 = 207{,}51$$
$$S_{yy} = 237 - \frac{1}{400}(-1)^2 = 237{,}00$$
$$S_{xy} = 69 - \frac{1}{400}(-86)(-1) = 68{,}78$$

todos arredondados até a terceiras casa decimal e, finalmente,

$$r = \frac{68{,}78}{\sqrt{(207{,}51)(237{,}00)}} \approx 0{,}31$$

17.2 A INTERPRETAÇÃO DE r

Quando r é igual a $+1$, -1 ou 0, não há problema quanto à interpretação do coeficiente de correlação. Como já indicamos, r é $+1$ ou -1 quando todos os pontos efetivamente estão sobre uma reta, e é zero quando o ajuste da reta de mínimos quadrados é tão pobre que o conhecimento de x em nada contribui para a previsão de y. De modo geral, a definição de r nos diz que $100r^2$ é a percentagem da variação total dos y que é explicada por sua relação com x, ou devida à relação. Isso só já é uma medida importante da relação entre duas variáveis; além disso, permite comparações válidas da intensidade de várias relações.

EXEMPLO 17.3 Se $r = 0{,}80$ num estudo e $r = 0{,}40$ num outro, estaria correto dizer que a correlação de 0,80 é duas vezes mais forte do que a correlação de 0,40?

Solução Não! Quando $r = 0{,}80$, então $100(0{,}80)^2 = 64\%$ da variação dos y é explicada pela relação com x, e quando $r = 0{,}40$, então apenas $100(0{,}40)^2 = 16\%$ da variação dos y é explicada pela relação com x. Assim, no sentido de "percentagem de variação explicada por", podemos dizer que a correlação de 0,80 é quatro vezes mais forte que a correlação de 0,40.

Da mesma forma, dizemos que uma relação em que $r = 0{,}60$ é nove vezes mais forte do que uma relação em que $r = 0{,}20$.

Há várias ciladas na interpretação do coeficiente de correlação. Em primeiro lugar, nem sempre lembramos que r mede apenas a intensidade de relações lineares; em segundo lugar, deveria ser lembrado que uma correlação forte (um valor de r próximo de +1 ou de –1) não implica necessariamente uma relação de causa e efeito.

Se r é calculado indiscriminadamente, por exemplo, para os três conjuntos de dados da Figura 17.4, obtemos $r = 0{,}75$ em cada caso, mas r só é uma medida significativa da intensidade da relação no primeiro caso. No segundo caso, há uma relação curvilínea muito forte entre as duas variáveis, e no terceiro caso, seis dos sete pontos estão efetivamente sobre uma reta, mas o sétimo ponto está tão distante que sugere a possibilidade de um erro grosseiro de mensuração ou no registro dos dados. Assim, antes de calcularmos r, deveríamos sempre esboçar os dados para ver se há razão para crer que a relação seja, de fato, linear.

O erro de interpretar um valor elevado de r (ou seja, um valor próximo de +1 ou de –1) como indicação de uma relação de causa e efeito é melhor explicado com alguns exemplos. Uma ilustração freqüentemente usada é a elevada correlação positiva entre as vendas anuais de chicletes e a incidência de crime nos EUA. Obviamente, não podemos concluir que se possa reduzir a taxa de criminalidade com a proibição da venda de chicletes; ambas as variáveis dependem do tamanho da população, e é essa relação mútua com uma terceira variável (tamanho da população) que produz a correlação positiva. Um outro exemplo é a forte correlação positiva que foi observada entre o número de cegonhas que fazem ninho em aldeias inglesas e o número de crianças nascidas nas mesmas aldeias. Deixamos à imaginação do leitor explicar por que poderia haver uma forte correlação neste caso, na ausência de qualquer relação de causa e efeito.

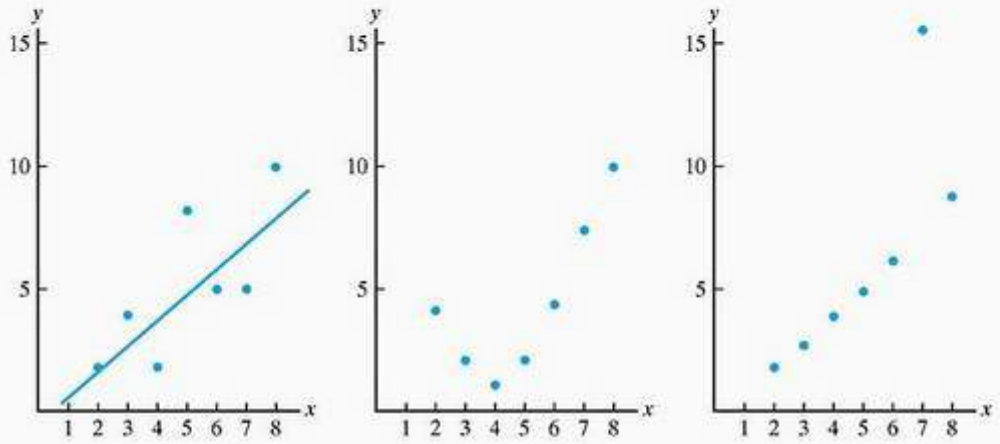

Figura 17.4
Três conjuntos de pares de dados para os quais $r = 0{,}75$.

EXERCÍCIOS

17.1 No Exemplo 16.7 fornecemos os dados seguintes relativos aos tempos médios semanais, em horas, que seis estudantes dedicaram aos seus trabalhos para casa e os índices de pontuação para as disciplinas que fizeram naquele semestre:

Horas gastas em deveres de casa x	*Índice de pontuação* y
15	2,0
28	2,7
13	1,3
20	1,9
4	0,9
10	1,7

Calcule r e compare o resultado com a raiz quadrada do valor de r^2 fornecido no impresso da Figura 16.9.

17.2 No Exercício 16.3 fornecemos os dados relativos às quantidades controladas de comida contendo um certo conservante que $n = 10$ crianças de quatro anos tinham consumido e a medição subjetiva de sua hiperatividade 45 minutos depois. Sabendo que $\sum x = 525$, $\sum x^2 = 32.085$, $\sum y = 100$, $\sum y^2 = 1.192$ e $\sum xy = 5.980$, calcule r.

17.3 Com referência ao Exercício 17.2, qual é a percentagem da variação total de y (medição de hiperatividade) que é devida pela relação com x (a quantidade de comida contendo o conservante que foi ingerida)?

17.4 Os tempos (em minutos) que $n = 12$ mecânicos levaram para montar uma máquina de manhã, x, e de tarde, y, são os seguintes:

x	y
12	14
11	11
9	14
13	11
10	12
11	15
12	12
14	13
10	16
9	10
11	10
12	14

Sabendo que $S_{xx} = 25{,}67$, $S_{xy} = -0{,}33$ e $S_{yy} = 42{,}67$, calcule r.

17.5 Use um computador ou uma calculadora gráfica para calcular r a partir dos dados originais fornecidos no Exercício 17.4.

17.6 Os dados a seguir foram obtidos num estudo da relação entre a resistência (ohms) e o tempo de falha (minutos) de certos resistores sobrecarregados:

Resistência	*Tempo de falha*
48	45
28	25
33	39
40	45
36	36
39	35
46	36
40	45
30	34
42	39
44	51
48	41
39	38
34	32
47	45

Use um computador ou uma calculadora gráfica para calcular r. Também determine qual percentagem da variação no tempo de falha é causada por diferenças na resistência.

17.7 Os dados a seguir registram a altitude (em pés) e a temperatura máxima média (em graus Fahrenheit) de oito cidades no Arizona, EUA, num certo feriado no fim de verão:

Altitude	*Temperatura máxima*
1.418	92
6.905	70
735	98
1.092	94
5.280	79
2.372	88
2.093	90
196	96

Use um computador ou uma calculadora gráfica para calcular r. Também determine qual percentagem da variação nas temperaturas máximas é causada por diferenças na altitude.

17.8 Depois de calcular r para um conjunto grande de pares de dados, uma estudante descobriu, para seu desalento, que a variável que deveria ter sido designada por x fora designada por y, e vice-versa. Há alguma razão para desalento?

17.9 Depois de calcular r para as alturas e os pesos de um grande número de pessoas, um estudante constatou que os pesos eram dados em libras e as alturas em polegadas. Ele pretendia obter r para os pesos em quilogramas e as alturas em centímetros. Sabendo que cada centímetro corresponde a 0,393 polegadas e cada quilograma corresponde a 2,2 libras, como deve ele corrigir seus cálculos?

17.10 Se calcularmos r para cada um dos conjuntos de dados a seguir, seria surpreendente se obtivermos $r = 1$ para (a) e $r = -1$ para (b)? Explique suas respostas.

(a)

x	y
6	9
14	11

(b)

x	y
12	5
8	15

17.11 Decida em cada caso a seguir se pode ser esperada uma correlação positiva, uma correlação negativa, ou nenhuma correlação:

(a) as idades de maridos e mulheres;
(b) a quantidade de borracha em pneus e a quilometragem por eles percorrida;
(c) o número de horas que os golfistas praticam e os pontos que obtêm;
(d) tamanho do sapato e QI;
(e) o peso da carga de caminhões e seu consumo de combustível.

17.12 Decida em cada caso a seguir se pode ser esperada uma correlação positiva, uma correlação negativa, ou nenhuma correlação:

(a) medida de pólen e vendas de remédios antialérgicos;
(b) renda e educação;
(c) número de dias ensolarados em Curitiba no mês de fevereiro e freqüência ao zoológico de Curitiba;
(d) número da camisa e senso de humor;
(e) número de pessoas que tomam remédio contra gripe e número de pessoas que contraem gripe.

17.13 Se $r = 0{,}41$ para um conjunto de pares de dados e $r = 0{,}29$ para um outro conjunto de pares de dados, compare as intensidades das duas relações.

17.14 Num estudo médico, obteve-se $r = 0{,}70$ para o peso de crianças de seis meses de idade e seu peso ao nascerem, e $r = 0{,}60$ para o peso de crianças de seis meses de idade e sua ingestão diária de alimentos. Dê um contra-exemplo para mostrar que não é válido concluir que o peso ao nascer e a ingestão diária de alimentos conjuntamente respondam por

$$(0{,}70)^2 100\% + (0{,}60)^2 100\% = 85\%$$

da variação do peso das crianças quando elas têm seis meses de idade. Você pode explicar por que essa conclusão não é válida?

17.15 Trabalhando com vários dados sócio-econômicos dos últimos anos, um pesquisador obteve $r = 0{,}9225$ para o número de diplomas de língua estrangeira conferidos por faculdades e universidades brasileiras e a extensão das rodovias federais brasileiras. Pode-se concluir que

$$(0{,}9225)^2 100\% \approx 85{,}1\%$$

da variação nos diplomas de língua estrangeira é devida às diferenças na propriedade de rodovias federais?

17.16 Uma estudante calculou a correlação entre a altura e o peso de um grupo grande de crianças da terceira série, obtendo um valor de $r = 0{,}32$. Ela não conseguiu decidir se ela deveria concluir ser alto faz com que a criança aumente de peso, ou se apresentar excesso de peso que faz com que a criança cresça mais. Ajude-a a sair desse dilema.

17.17 No Exemplo 17.2, ilustramos o uso do coeficiente de correlação na análise de uma tabela de contingência com categorias ordenadas. Use o mesmo procedimento para analisar a tabela seguinte, reproduzida do Exercício 14.29, em que foi analisada a relação entre a fidelidade e a seletividade de aparelhos de rádio utilizando o critério qui-quadrado:

		Fidelidade		
		Baixa	*Média*	*Alta*
Seletividade	*Baixa*	7	12	31
	Média	35	59	18
	Alta	15	13	0

17.18 Use o mesmo procedimento que no Exemplo 17.2 para analisar a tabela seguinte, reproduzida do Exercício 14.24, em que foi analisada a relação entre o padrão de vestuário de empregados de bancos e a velocidade de sua progressão profissional utilizando o critério qui-quadrado:

		Velocidade de progressão		
		Baixa	*Média*	*Alta*
Padrão de vestuário	*Muito bem vestido*	38	135	129
	Bem vestido	32	68	43
	Mal vestido	13	25	17

Observe que a velocidade da progressão funcional vai de baixa para alta, enquanto que o padrão de vestuário vai de alto para baixo.

17.3 ANÁLISE DE CORRELAÇÃO

Quando calculamos r com base em dados amostrais, podemos obter uma correlação positiva ou negativa bastante forte apenas por acaso, mesmo que não haja relação alguma entre as duas variáveis sob consideração.

Suponha, por exemplo, que tomemos dois dados, um vermelho e outro verde, e que os joguemos cinco vezes, obtendo os resultados seguintes:

Dado vermelho x	*Dado verde* y
4	5
2	2
4	6
2	1
6	4

Presumivelmente, não há relação alguma entre x e y, os números que aparecem nos dois dados. É difícil ver por que valores grandes de x devam corresponder a valores grandes de y e valores pequenos de x com valores pequenos de y mas, calculando r, obtemos o valor surpreendentemente alto de $r = 0{,}66$. Isso suscita a questão sobre se não há algo errado na suposição de não haver relação entre x e y e, para respondê-la, devemos verificar se o valor elevado de r pode ser atribuído ao acaso.

Quando calculamos um coeficiente de correlação a partir de dados amostrais, como no exemplo precedente, o valor de r obtido é apenas uma estimativa de um parâmetro correspondente, que é o **coeficiente de correlação populacional**, que denotamos por ρ (a letra grega rô). O que r mede numa amostra, ρ mede numa população.

Para fazer inferências sobre ρ com base em r, devemos fazer várias suposições sobre as distribuições das variáveis aleatórias cujos valores observamos. Na **análise de correlação normal** fazemos as mesmas suposições que na análise de regressão normal (ver página 412), com a exceção de que os x não são constantes, e sim valores de uma variável aleatória com distribuição normal.

Como a distribuição amostral de r é bastante complicada sob tais suposições, é prática comum basear as inferências sobre ρ na **transformação Z de Fisher**, que é uma mudança de escala de r para Z, dada por

$$Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r}$$

Aqui, ln denota o "logaritmo natural", ou seja, o logaritmo de base e, onde $e = 2{,}71828....$ Essa transformação é assim chamada em homenagem a R. A. Fisher, um estatístico proeminente que mostrou que, sob as suposições da análise de correlação normal e para qualquer valor de ρ, a distribuição de Z é aproximadamente normal com

$$\mu_Z = \frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho} \quad \text{e} \quad \sigma_Z = \frac{1}{\sqrt{n-3}}$$

Transformando Z em unidades padrão (ou seja, subtraindo μ_z e então dividindo por σ_z), concluímos que

ESTATÍSTICA PARA INFERÊNCIAS SOBRE ρ

$$z = (Z - \mu_Z)\sqrt{n-3}$$

tem aproximadamente a distribuição normal padrão. A aplicação dessa teoria é muito facilitada pela Tabela X no fim do livro, que dá os valores de Z correspondentes a $r = 0{,}00$; 0,01; 0,02; 0,03;... e 0,99. Observe que a tabela dá apenas valores positivos; se r é negativo, basta procurarmos por $-r$ e tomar o negativo do valor correspondente de Z. Observe também que a fórmula de μ_z é semelhante à de Z, com r substituído por ρ; portanto, a Tabela X pode ser usada para procurar valores de μ_z correspondentes a valores dados de ρ.

EXEMPLO 17.4 Ao nível 0,05 de significância, teste a hipótese nula de ausência de correlação (isto é, a hipótese nula $\rho = 0$) para a ilustração à página 442, em que jogamos um par de dados cinco vezes, obtendo $r = 0{,}66$.

Solução

1. $H_0 : \rho = 0$

   $H_A : \rho \neq 0$

2. $\alpha = 0{,}05$

3. Como $\mu_z = 0$ para $\rho = 0$, rejeitar a hipótese nula se $z \leq -1{,}96$ ou $z \geq 1{,}96$, onde

$$z = Z \cdot \sqrt{n-3}$$

   Caso contrário, concluir que o valor de r não é significativo.

4. Substituindo $n = 5$ e $Z = 0{,}793$, que é o valor de Z correspondente a $r = 0{,}66$, de acordo com a Tabela X, obtemos

$$z = 0{,}793\sqrt{5-3}$$
$$= 1{,}12$$

5. Como $z = 1{,}12$ cai entre $-1{,}96$ e 1,96, a hipótese nula não pode ser rejeitada. Em outras palavras, o valor de r obtido não é significativo, como devíamos esperar, evidentemente. Uma maneira alternativa de tratar esse tipo de problema (ou seja, de testar a hipótese nula $\rho = 0$) é dada no Exercício 17.22.

EXEMPLO 17.5 Com referência ao exemplo do alcance auditivo e ruído de aeroporto, no qual mostramos que $r = -0{,}95$ para $n = 12$, teste a hipótese nula $\rho = -0{,}80$ contra a hipótese alternativa $\rho < -0{,}80$, ao nível 0,01 de significância.

Solução

1. $H_0 : \rho = -0{,}80$

   $H_A : \rho < -0{,}80$

2. $\alpha = 0{,}01$

3. Rejeitar a hipótese nula se $z \geq -2{,}33$, onde

$$z = (Z - \mu_Z)\sqrt{n-3}$$

4. Substituindo $n = 12$ e $Z = -1{,}832$, que é o valor correspondente a $r = -0{,}95$, e $\mu_z = -1{,}099$, que corresponde a $\rho = -0{,}80$, obtemos

$$z = [-1{,}832 - (-1{,}099)]\sqrt{12-3}$$
$$\approx -2{,}20$$

5. Como $-2{,}20$ cai entre $-2{,}33$ e 2,33, a hipótese nula não pode ser rejeitada.

Para construir intervalos de confiança para ρ, primeiro construímos intervalos de confiança para μ_z e fazemos, então, a transformação para r e ρ por meio da Tabela X. Pode-se obter uma fórmula do intervalo de confiança para μ_z substituindo

$$z = (Z - \mu_Z)\sqrt{n-3}$$

como termo médio da desigualdade dupla $-z_{\alpha/2} < z < z_{\alpha/2}$ e, então, manipulando algebricamente os termos de modo que o termo do meio seja μ_z. Isso leva ao seguinte intervalo de $(1-\alpha)100\%$ de confiança para μ_z:

INTERVALO DE CONFIANÇA PARA μ_z

$$Z - \frac{z_{\alpha/2}}{\sqrt{n-3}} < \mu_Z < Z + \frac{z_{\alpha/2}}{\sqrt{n-3}}$$

EXEMPLO 17.6 Sabendo que $r = 0{,}62$ para as estimativas feitas por dois mecânicos para uma amostra aleatória de $n = 30$ consertos, construa um intervalo de 95% de confiança para o coeficiente de correlação populacional ρ.

Solução Na Tabela X obtemos $Z = 0{,}725$ correspondente a $r = 0{,}62$, e substituindo esse valor, juntamente com $n = 30$ e $z_{0{,}025} = 1{,}96$ na fórmula precedente do intervalo de confiança para μ_z, obtemos

$$0{,}725 - \frac{1{,}96}{\sqrt{27}} < \mu_Z < 0{,}725 + \frac{1{,}96}{\sqrt{27}}$$

ou

$$0{,}348 < \mu_Z < 1{,}102$$

Finalmente, procurando na Tabela X os valores de r que estão mais próximos de $Z = 0{,}348$ e $Z = 1{,}102$, obtemos o intervalo de 95 % de confiança

$$0{,}33 < \rho < 0{,}80$$

para a verdadeira intensidade da relação linear entre as estimativas de custo feitas pelos dois mecânicos.

EXERCÍCIOS

17.19 Considerando que as suposições para uma análise de correlação normal tenham sido satisfeitas, teste a hipótese nula $\rho = 0$ contra a hipótese alternativa $\rho \neq 0$ ao nível 0,05 de significância, sabendo que
(a) $n = 15$ e $r = 0{,}59$;
(b) $n = 20$ e $r = 0{,}41$;
(c) $n = 40$ e $r = 0{,}36$.

17.20 Considerando que as suposições para uma análise de correlação normal tenham sido satisfeitas, teste a hipótese nula $\rho = 0$ contra a hipótese alternativa $\rho \neq 0$ ao nível 0,01 de significância, sabendo que
(a) $n = 14$ e $r = 0{,}54$;
(b) $n = 22$ e $r = -0{,}61$;
(c) $n = 44$ e $r = 0{,}42$.

17.21 Considerando que as suposições para uma análise de correlação normal tenham sido satisfeitas, teste a hipótese nula $\rho = -0{,}50$ contra a hipótese alternativa $\rho > -0{,}50$ ao nível 0,01 de significância, sabendo que

(a) $n = 17$ e $r = -0{,}22$;
(b) $n = 34$ e $r = -0{,}43$.

17.22 Diante das suposições de uma análise de correlação normal, o teste da hipótese nula $\rho = 0$ também pode ser baseado na estatística

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

que tem a distribuição t com $n - 2$ graus de liberdade. Use essa estatística para testar em cada caso se o valor de r é significante ao nível 0,05 de significância:
(a) $n = 12$ e $r = 0{,}77$;
(b) $n = 16$ e $r = 0{,}49$.

17.23 Refaça o Exercício 17.22 trocando o nível de significância para 0,01.

17.24 A estatística t dada no Exercício 17.22 é idêntica à estatística t que testa $\beta = 0$ na análise de regressão linear, sendo seu valor fornecido pelos programas de regressão linear de aplicativos estatísticos e calculadoras gráficas. Isso é ilustrado pelo seguinte:
(a) Usando o programa de regressão linear de um pacote de aplicativos estatísticos ou uma calculadora gráfica para obter os valores de t e r para os dados do Exercício 16.8.
(b) Substitua o valor obtido para r na parte (a) e $n = 8$ na fórmula para t dada no Exercício 17.22.
(c) Compare os dois valores de t.

17.25 Num estudo da relação entre a taxa de mortalidade decorrente de câncer no pulmão e o consumo *per capita* de cigarros 20 anos antes, os dados de $n = 9$ países acusaram $r = 0{,}73$. Ao nível 0,05 de significância, teste a hipótese nula $\rho = 0{,}50$ contra a hipótese alternativa $\rho > 0{,}50$.

17.26 Num estudo da relação entre o calor proporcionado pela queima de (um metro cúbico) de madeira verde e de madeira seca ao ar livre, os dados de $n = 13$ tipos de madeira acusaram $r = 0{,}94$. Ao nível 0,01 de significância, teste a hipótese nula $\rho = 0{,}75$ contra a alternativa $\rho \neq 0{,}75$.

17.27 Considerando que as suposições para uma análise de correlação normal tenham sido satisfeitas, use transformação Z de Fisher para construir intervalos de 95% de confiança ρ, sabendo que
(a) $n = 15$ e $r = 0{,}80$;
(b) $n = 28$ e $r = -0{,}24$;
(c) $n = 63$ e $r = 0{,}55$.

17.28 Considerando que as suposições para uma análise de correlação normal tenham sido satisfeitas, use transformação Z de Fisher para construir intervalos de 99% de confiança ρ, sabendo que
(a) $n = 20$ e $r = -0{,}82$;
(b) $n = 25$ e $r = 0{,}34$;
(c) $n = 75$ e $r = 0{,}18$.

*17.4 CORRELAÇÕES MÚLTIPLA E PARCIAL

Na Seção 17.1, introduzimos o coeficiente de correlação como uma medida da aderência de uma reta de mínimos quadrados a um conjunto de pares de dados. Se as previsões devem ser feitas por meio de uma equação da forma

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

obtida pelo método dos mínimos quadrados, como na Seção 16.4, definimos o **coeficiente de correlação múltipla** da mesma forma pela qual definimos originalmente r. Tomamos a raiz quadrada da quantidade

$$\frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

que é a proporção da variação total dos y que pode ser atribuída à relação com os x. A única diferença é que, agora, calculamos $\hat{y}$ por meio da equação de regressão múltipla, em lugar da equação $\hat{y} = a + bx$.

Para exemplificar, voltamos ao Exemplo 16.9, em que baseamos uma equação de regressão linear múltipla no impresso de computador mostrado na Figura 16.10. Como foi indicado naquela ocasião, suprimimos uma parte daquele impresso, mas como agora vamos precisar daquela parte, reproduzimos o impresso completo na Figura 17.5.

EXEMPLO 17.7 Use a definição de coeficiente de correlação múltipla dado acima para determinar seu valor para os dados do Exemplo 16.9, que tratou com o número de quartos, o número de banheiros e os preços (em unidades monetárias) pelos quais oito casas unifamiliares de um certo bairro foram vendidas recentemente.

Solução Na análise de variância da Figura 17.5 (na coluna encabeçada por SS), vemos que a soma de quadrados de regressão é 4.877.608.452 e que a soma de quadrados total é 5.455.580.000. Assim, o coeficiente de correlação múltipla é igual à raiz quadrada de

$$\frac{4.877.608.452}{5.455.580.000} \approx 0{,}894$$

Análise de Regressão: y contra x1, x2

```
The regression equation is
y = 65430 + 16752 x1 + 11235 x2

Predictor    Coef   SE Coef        T      P
Constant    65430     12134     5.39  0.003
x1          16752      6636     2.52  0.053
x2          11235      9885     1.14  0.307

S = 10751   R-Sq = 89.4%   R-Sq(adj) = 85.2%

Analysis of Variance

Source          DF          SS          MS      F      P
Regression       2  4877608452  2438804226  21.10  0.004
Residual Error   5   577971548   115594310
Total            7  5455580000

Source          DF       Seq SS
x1               1  4728284225
x2               1   149324227
```

Figura 17.5 Impresso completo do exemplo de regressão múltipla.

Denotando esse valor por R, escrevemos $R \approx \sqrt{0,894} = 0,95$, arredondado até a segunda casa decimal, o que é considerado "essencialmente não-negativo", de acordo com o dicionário de termos estatísticos de Kendall e Buckland. Na prática, R^2 é usado mais freqüentemente do que R, e deveria ser observado que seu valor, que é denotado por R-Sq, é de fato dado como 89,4% no impresso da Figura 17.5.

Esse exemplo também serve para ilustrar que o acréscimo de mais variáveis independentes num estudo de correlação pode não ser suficientemente produtivo a ponto de justificar o trabalho extra. Como se pode mostrar que $r = 0,93$ para y e x_1 (número de quartos) sozinhos, vê-se que se ganha muito pouco considerando também x_2 (número de banheiros). A situação é bastante diferente, no entanto, no Exercício 17.30, em que as duas variáveis independentes x_1 e x_2 conjuntamente respondem por uma proporção muito maior da variação total em y do que x_1 ou x_1 isoladamente.

Quando abordamos o problema da correlação e causa, mostramos que uma forte correlação entre duas variáveis pode ser totalmente devido à sua dependência de uma terceira variável. Ilustramos isso com os exemplos das vendas de chicletes e a taxa de criminalidade, o número de nascimentos e número de cegonhas. Para dar outro exemplo, consideremos as duas variáveis, x_1, a venda semanal de taças de chocolate quente numa estação de veraneio, e x_2, o número semanal de visitantes da estação. Se, com base em dados adequados, obtemos $r = -0,30$ para essas variáveis, isso deveria constituir uma surpresa – afinal de contas, deveríamos esperar maior volume de vendas de chocolate quente quando há mais visitantes, e vice-versa, portanto uma correlação positiva.

Pensando um pouco mais, entretanto, podemos admitir que a correlação negativa de –0,30 seja causada pelo fato de as variáveis x_1 e x_2 estarem ambas relacionadas com uma terceira variável, x_3, a temperatura média semanal na estação de veraneio. Com uma temperatura alta, haverá mais visitantes, os quais, entretanto, preferirão bebidas frias ao chocolate quente; se a temperatura é baixa, haverá menos visitantes, mas que preferirão o chocolate quente às bebidas frias. Suponhamos, então, que outros dados forneçam $r = -0,70$ para x_1 e x_3, e $r = 0,80$ para x_2 e x_3. Esses valores parecem razoáveis, pois vendas baixas de chocolate quente devem acompanhar altas temperaturas e vice-versa, enquanto que o número de visitantes é elevado quando a temperatura é alta, e reduzido quando a temperatura é baixa.

No exemplo precedente, deveríamos, na realidade, ter pesquisado a relação entre x_1 e x_2 (vendas de chocolate quente e número de visitantes da estação) quando todos os outros fatores, especialmente a temperatura, são mantidos constantes. Como quase nunca é possível exercer um tal controle, constatou-se que uma estatística denominada **coeficiente de correlação parcial** desempenha satisfatoriamente a função de eliminar os efeitos de outras variáveis. Representando os coeficientes de correlação comuns de x_1 e x_2, x_1 e x_3, e x_2 e x_3, por r_{12}, r_{13} e r_{23}, o coeficiente de correlação parcial para x_1 e x_2, com x_3 considerado fixo, é dado por

COEFICIENTE DE CORRELAÇÃO PARCIAL

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

EXEMPLO 17.8 Calcule $r_{12.3}$ para o exemplo precedente, que tratava das vendas de chocolate quente e o número de visitantes de uma estação de veraneio.

Solução Substituindo $r_{12} = -0,30$, $r_{13} = -0,70$ e $r_{23} = 0,80$ na fórmula para $r_{12.3}$, obtemos

$$r_{12.3} = \frac{(-0,30) - (-0,70)(0,80)}{\sqrt{1 - (-0,70)^2}\sqrt{1 - (0,80)^2}} \approx 0,607$$

Como era de se esperar, esse resultado mostra que há uma relação positiva entre as vendas de chocolate quente e o número de visitantes da estação quando se elimina o efeito das diferenças de temperatura.

Esse exemplo foi dado principalmente para ilustrar o que queremos dizer com um coeficiente de correlação parcial, mas também serve para lembrar que os coeficientes de correlação podem muito bem dar a impressão errada se não forem interpretados com cuidado.

EXERCÍCIOS

*17.29 Num problema de regressão múltipla, a soma de quadrados de regressão é

$$\sum(\hat{y} - \bar{y})^2 = 45.225$$

e a soma de quadrados total é

$$\sum(y - \bar{y})^2 = 136.210$$

Encontre o valor do coeficiente de correlação múltipla.

*17.30 Com referência ao Exercício 16.34 à página 420, use o mesmo aplicativo utilizado anteriormente para obter o coeficiente de correlação múltipla. Também determine os coeficientes de correlação par y e x_1 (idade) sozinhos e para y e x_2 (anos de pós-graduação) sozinhos, comparando-os com o coeficiente de correlação múltipla.

17.31 Com referência ao Exercício 16.34 à página 420, use o mesmo aplicativo do que naquele exercício para obter o coeficiente de correlação múltipla.

17.32 Uma equipe de pesquisadores conduziu um experimento para ver se a altura de certas roseiras pode ser prevista com base na quantidade de fertilizante e na quantidade de irrigação que são aplicados ao solo. Para prever a altura com base em ambas variáveis, eles obtiveram um coeficiente de correlação múltipla de 0,58; para prever a altura com base no fertilizante sozinho, eles obtiveram $r = 0{,}66$. Comente esses resultados.

17.33 Com referência ao Exercício 16.35 à página 421, use um computador ou uma calculadora gráfica para determinar os coeficientes de correlação necessários para calcular o coeficiente de correlação parcial para a dureza do aço e a temperatura em que o aço é temperado quando o conteúdo de cobre é mantido fixo.

17.34 Um experimento forneceu os seguintes resultados: $r_{12} = 0{,}80$, $r_{13} = -0{,}70$ e $r_{23} = 0{,}90$. Explique por que essas quantidades não podem estar todas corretas.

17.5 LISTA DE TERMOS-CHAVE (com indicação das páginas de suas definições)

Análise de correlação normal, 442
Coeficiente de correlação, 433
Coeficiente de correlação do momento produto, 435
*Coeficiente de correlação múltipla, 446
*Coeficiente de correlação parcial, 447
Coeficiente de correlação populacional, 442
Coeficiente de determinação, 433
Correlação negativa, 433
Correlação positiva, 433
Covariância amostral, 436
Momento produto, 435
Soma de quadrados de regressão, 432
Soma de quadrados residual, 433
Soma de quadrados total, 432
Transformação Z de Fisher, 442

17.6 REFERÊNCIAS

Informação mais detalhada sobre as correlações múltipla e parcial podem ser encontradas em

EZEKIEL, M., and FOX, K. A., *Methods of Correlation and Regression Analysis*, 3rd ed. New York: John Wiley & Sons, Inc., 1959.

HARRIS, R. J., *A Primer of Multivariate Statistics*. New York: Academic Press, Inc., 1975.

e um tratamento teórico avançado é dado no Volume 2 de

KENDALL, M. G., and STUART, A., *The Advanced Theory of Statistics*, 3rd ed. New York: Hafner Press, 1973.

O Volume 1 desse livro fornece a fundamentação teórica dos testes de significância para r.

18 TESTES NÃO-PARAMÉTRICOS

A maioria dos testes citados nos Capítulos 12 a 17 exige suposições específicas sobre a população, ou populações, de onde provêm as amostras. Em muitos casos, devemos admitir que as populações tenham aproximadamente a forma de distribuições normais, que suas variâncias sejam conhecidas ou que se saiba que são iguais, ou que as amostras são independentes. Como há muitas situações em que é duvidoso se todas as suposições necessárias podem ser satisfeitas, os estatísticos elaboraram procedimentos alternativos baseados em suposições menos restritivas, que passaram a ser conhecidos como **testes não-paramétricos**.

Afora o fato de os testes não-paramétricos poderem ser usados sob condições mais gerais, do que os testes padrão que substituem, os testes não-paramétricos são, em geral, mais fáceis de explicar e de entender; além disso, em muitos testes não-paramétricos a carga computacional é tão leve que eles são considerados como técnicas "rápidas e fáceis" ou "de atalho". Por todas essas razões, os testes não-paramétricos têm se tornado muito populares, existindo vasta literatura voltada à sua teoria e aplicação.

Nas Seções 18.1 e 18.2, apresentamos o teste de sinais como uma alternativa não-paramétrica aos testes relativos a médias e aos testes relativos a diferenças entre médias, baseados em pares de dados. Nas Seções 18.3 e 18.4, damos outro teste não-paramétrico, que atende aos mesmos objetivos, mas que desperdiça menos informação. Nas Seções 18.5 a 18.7, apresentamos uma alternativa não-paramétrica aos testes referentes à diferença entre as médias de amostras independentes, e uma alternativa não-paramétrica um tanto semelhante para a análise da variância de um

critério. Nas Seções 18.8 a 18.10, vemos como testar a aleatoriedade de uma amostra após efetivamente obtidos os dados; e na Seção 18.11, apresentamos um teste não-paramétrico da significância de uma relação entre pares de dados. Finalmente, na Seção 18.12, mencionamos alguns pontos fracos dos testes não-paramétricos e, na Seção 18.13, damos uma tabela que lista os diversos testes não-paramétricos e os testes "padrão" que eles substituem.

18.1 O TESTE DE SINAIS

Exceto pelos testes para grandes amostras, todos os testes referentes a médias que estudamos no Capítulo 12 foram baseados na suposição de que as populações amostradas tinham aproximadamente a forma de distribuições normais. Quando essa suposição é insustentável na prática, esses testes-padrão podem ser substituídos por qualquer uma de várias alternativas não-paramétricas, e essas são o assunto das Seções 18.1 a 18.7. O mais simples dentre as alternativas é o **teste de sinais**, que vamos estudar nesta seção e na Seção 18.2.

Aplicamos o **teste de sinais para uma amostra** quando extraímos amostras de uma população contínua, de modo que a probabilidade de obtermos um valor amostral menor do que a mediana e a probabilidade de obter um valor amostral maior do que a mediana são ambas $\frac{1}{2}$. Naturalmente, quando a população é simétrica, a média μ e a mediana $\tilde{\mu}$ coincidem e podemos enunciar as hipóteses em termos de qualquer um desses dois parâmetros.

Para testar a hipótese nula $\tilde{\mu} = \tilde{\mu}_0$ contra uma alternativa apropriada com base numa amostra aleatória de tamanho n, substituímos cada valor amostral maior do que $\tilde{\mu}_0$ por um sinal de mais, e cada valor amostral menor do que $\tilde{\mu}_0$ por um sinal de menos. Testamos, então, a hipótese nula de que o número de sinais de mais são valores de uma variável aleatória de distribuição binomial com $p = \frac{1}{2}$. Se algum valor amostral for efetivamente igual a $\tilde{\mu}_0$, o que pode ocorrer facilmente quando tratamos com dados arredondados, simplesmente o descartamos.

Para fazermos um teste de sinais de uma amostra quando a amostra é razoavelmente pequena, recorremos diretamente a uma tabela de probabilidades binomiais, como a Tabela V do final do livro, ou à tabela do *National Bureau of Standards* à qual nos referimos à página 213. Alternativamente, podemos usar um computador ou uma calculadora para obter as probabilidades binomiais requeridas. Contudo, quando a amostra é grande, utilizamos a aproximação normal da distribuição binomial, conforme ilustrado na Seção 18.2.

EXEMPLO 18.1 Para conferir a alegação de um professor de que o valor publicado de 0,050 para o coeficiente de fricção de metais bem engraxados deve ser muito pequeno, numa turma de ciências fazem 18 determinações do coeficiente, obtendo 0,054; 0,052; 0,044; 0,056; 0,050; 0,051; 0,055; 0,053; 0,047; 0,053; 0,052; 0,050; 0,051; 0,051; 0,054; 0,046; 0,053 e 0,043. Normalmente, o teste t de uma amostra seria a escolha lógica para testar a alegação, mas a assimetria dos dados sugere o uso de uma alternativa não-paramétrica. Portanto, o professor sugere que a turma use o teste de sinais para uma amostra para testar a hipótese nula $\tilde{\mu} = 0{,}050$ contra a alternativa $\tilde{\mu} > 0{,}050$, ao nível 0,05 de significância.

Solução

1. $H_0 : \tilde{\mu} = 0{,}050$
$H_A : \tilde{\mu} > 0{,}050$

2. $\alpha = 0{,}05$

3'. A estatística de teste é o número de sinais de mais, ou seja, o número de valores acima de 0,050.

4'. Substituindo por um sinal de mais cada valor maior do que 0,050, por um sinal de menos cada valor menor do que 0,050, e descartando os dois valores iguais a 0,050, obtemos

$$++-++++-++++++-+-$$

Assim, $x = 12$, e a Tabela V mostra que, para $n = 16$ e $p = 0{,}50$, a probabilidade de $x \geq 12$, que é o valor p, é $0{,}028 + 0{,}009 + 0{,}002 = 0{,}039$.

5'. Como 0,039 é menor do que 0,05, a hipótese nula deve ser rejeitada. Os dados corroboram a alegação de que o valor publicado do coeficiente de fricção é muito pequeno. ■

Observe que utilizamos o método alternativo para testar hipóteses, conforme mencionado à página 309. Como na Seção 14.1, o método alternativo simplifica as coisas quando os testes são baseados diretamente em tabelas binomiais. Embora possa parecer não necessário, poderíamos ter usado um computador para o Exemplo 18.1. Se tivéssemos feito isso, teríamos obtido um impresso como o impresso mostrado na Figura 18.1. A diferença entre os dois valores de p, 0,039 e 0,0384, é devida, evidentemente, ao arredondamento.

O teste de sinais também pode ser aplicado quando trabalhamos com pares de dados, como na Seção 12.7. Em tais problemas, cada par de valores amostrais é substituído por um sinal de mais se o primeiro valor for maior do que o segundo valor, por um sinal de menos se o primeiro valor for menor do que o segundo valor, e descartamos pares de valores idênticos. Para pares de dados, o teste de sinais é usado para testar a hipótese nula de que a mediana da população dessas diferenças é zero. Quando é utilizado dessa maneira, o teste é denominado **teste de sinais com pares de dados**.

EXEMPLO 18.2 No Exemplo 12.9 foram fornecidos os dados relativos às perdas semanais médias de horas de trabalho devidas a acidentes em dez indústrias, antes e depois da adoção de um programa de segurança abrangente:

45 e 36	73 e 60	46 e 44	124 e 119	33 e 35
57 e 51	83 e 77	34 e 29	26 e 24	17 e 11

Utilizando o teste t de pares de amostras, mostramos a eficácia do programa de segurança ao nível 0,05 de significância. Use o teste de sinais com pares de dados para refazer esse exercício.

Solução

1. $H_0 : \tilde{\mu}_D = 0$, onde $\tilde{\mu}_D$ é a mediana da população de diferenças amostradas.
$H_A : \tilde{\mu}_D > 0$

2. $\alpha = 0{,}05$

3'. A estatística de teste é o número de sinais de mais, a saber, o número de fábricas nas quais decresceu o número de perdas semanais médias de horas de trabalho.

4'. Substituindo por um sinal de mais cada par de valores se o primeiro valor é maior do que o segundo, e por um sinal de menos se o primeiro valor for menor do que o segundo, obtemos

$$++++-+++++$$

Teste de Sinais para a Mediana: Dados

```
Sign test of median = 0.05000 versus > 0.05000

            N   BELOW  EQUAL  ABOVE       P    MEDIAN
Data       18       4      2     12  0.0384   0.05150
```

Figura 18.1 Impresso de MINITAB para o Exemplo 18.1.

Assim, $x = 9$, e a Tabela V mostra que, para $n = 10$ e $p = 0{,}50$, a probabilidade de $x \geq 9$, que é o valor p, é $0{,}010 + 0{,}001 = 0{,}011$.

5'. Como 0,011 é menor do que 0,05, a hipótese nula deve ser rejeitada. Como no Exemplo 12.9, concluímos que os programa de segurança é eficaz.

18.2 O TESTE DE SINAIS (Grandes Amostras)

Quando np e $n(1-p)$ são ambos maiores do que 5, permitindo-nos a utilização da aproximação normal da distribuição binomial, podemos basear o teste de sinais no teste de grandes amostras da Seção 14.2, a saber, na estatística

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

com $p_0 = 0{,}50$, que tem, aproximadamente, a distribuição normal padrão. Quando n é pequeno, pode ser indicado utilizar a correção de continuidade sugerida à página 334. Isso é o caso especialmente quando a hipótese nula só puder ser *mal e mal* rejeitada sem a correção de continuidade. Como já indicamos anteriormente, a correção de continuidade não precisa ser considerada quando, *sem ela*, a hipótese nula não puder ser rejeitada.

EXEMPLO 18.3 No Exercício 2.44 apresentamos os dados seguintes sobre os números de associados solteiros que participaram de 48 excursões patrocinadas pela associação de ex-alunos de uma universidade:

28	51	31	38	27	35	33	40	37	28	33	27
33	31	41	46	40	36	53	23	33	27	40	30
33	22	37	38	36	48	22	36	45	34	26	28
40	42	43	41	35	50	31	48	38	33	39	35

Use o teste de sinais de uma amostra para testar a hipótese nula $\tilde{\mu} = 32$ contra a hipótese alternativa que $\tilde{\mu} \neq 32$ ao nível 0,01 de significância.

Solução

1. $H_0 : \tilde{\mu} = 32$
$H_A : \tilde{\mu} \neq 32$

2. $\alpha = 0{,}01$

3. Rejeitar a hipótese nula se $z \leq -2{,}575$ ou $z \geq 2{,}575$, onde

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

com $p_0 = 0{,}50$; caso contrário, aceitar a hipótese nula ou reservar julgamento.

4. Contando o número de valores superiores a 32 (sinal de mais), o número de valores inferiores a 32 (sinal de menos), e o número de valores que são iguais a 32, obtemos 34, 14 e 0 (e, portanto, nenhum precisa ser descartado). Assim, $x = 14$, $n = 48$ e

$$z = \frac{34 - 48(0{,}50)}{\sqrt{48(0{,}50)(0{,}50)}} \approx 2{,}89$$

5. Como 2,89 é maior do que 2,575, a hipótese nula deve ser rejeitada. (Se tivéssemos utilizado correção de continuidade, teríamos obtido $z = 2{,}74$ e a conclusão teria sido a mesma.)

EXEMPLO 18.4 Dois supervisores classificaram o desempenho de uma amostra aleatória de empregados de uma grande companhia, numa escala de 0 a 100, com o resultado seguinte:

Supervisor 1	*Supervisor 2*
88	73
69	67
97	81
60	73
82	78
90	82
65	62
77	80
86	81
79	79
65	77
95	82
88	84
91	93
68	66
77	76
74	74
85	78

Use o teste de sinais para pares de amostras (baseado na aproximação normal da distribuição binomial) para testar ao nível 0,05 de significância se as diferenças entre os dois conjuntos de classificações podem ser atribuídas ao acaso,

(a) sem usar correção de continuidade;

(b) usando a correção de continuidade.

Solução (a) **1.** $H_0 : \tilde{\mu}_D = 0$, onde $\tilde{\mu}_D$ é a mediana da população de diferenças (entre as classificações dos supervisores) amostradas.
$H_A : \tilde{\mu}_D \neq 0$

2. $\alpha = 0{,}05$

3. Rejeitar a hipótese nula se $z \leq -1{,}96$ ou $z \geq 1{,}96$, onde

$$z = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

com $p_0 = 0{,}50$; caso contrário, aceitar a hipótese nula ou reservar julgamento.

4. Contando o número de diferenças positivas (sinal de mais), o número de diferenças negativas (sinal de menos), e o número de pares iguais (e que, portanto, devem ser descartados), verificamos que são, respectivamente, 12, 4 e 2. Assim, $x = 12$ e $n = 16$ e, como $np = 16(0{,}50) = 8$ e $n(1 - p) = 16(0{,}50) = 8$ são ambos maiores do que 5, temos uma justificativa para usar a aproximação normal da distribuição binomial. Substituindo na fórmula para z, obtemos

$$\frac{12 - 16(0{,}50)}{\sqrt{16(0{,}50)(0{,}50)}} = 2{,}00$$

5. Como 2,00 mal e mal excede 1,96, adiamos a tomada de decisão até recalcular z com correção de continuidade.

(b) **4.** Com a correção de continuidade, obtemos

$$\frac{11{,}5 - 16(0{,}50)}{\sqrt{16(0{,}50)(0{,}50)}} = 1{,}75$$

5. Como $z = 1{,}75$ cai entre $-1{,}96$ e $1{,}96$, verificamos que a hipótese nula não pode ser rejeitada. Concluímos que as diferenças entre as classificações dos supervisores podem ser atribuídas ao acaso. (Se tivéssemos baseado nossa decisão na Tabela V, o valor p teria sido maior do que 0,05, e a decisão final teria sido a mesma.)

EXERCÍCIOS

18.1 Numa amostra aleatória de 14 ocasiões, um empregado de uma firma na cidade teve de esperar 4,5; 8,6; 7,3; 7,0; 2,5; 6,1; 8,9; 6,5; 6,3; 1,6; 5,8; 6,3; 5,9; e 9,0 minutos pelo ônibus que o leva ao trabalho. Use o teste de sinais baseado na Tabela V e o nível 0,05 de significância, para testar a hipótese nula $\tilde{\mu} = 6{,}0$ (que sua espera mediana é 6,0 minutos) contra a hipótese alternativa $\tilde{\mu} \neq 6{,}0$.

18.2 Use um computador para refazer o Exercício 18.1.

18.3 Os dados abaixo constituem uma amostra aleatória de pesos (em gramas) do gengibre cristalizado de 20 caixinhas: 110,6; 113,5; 111,2; 109,8; 110,5; 111,1; 110,4; 109,7; 112,6; 110,8; 110,5; 110,0; 110,2; 111,4; 110,9; 110,5; 110,0; 109,4; 110,8; e 109,7. Use o teste de sinais baseado na Tabela V e o nível 0,01 de significância para testar a hipótese nula $\tilde{\mu} = 110{,}0$ (que o peso mediano de tais caixinhas de gengibre é 110,0 gramas) contra a hipótese alternativa $\tilde{\mu} > 110{,}0$.

18.4 Use um computador para refazer o Exercício 18.3.

18.5 Depois de jogar quatro partidas de golfe num clube carioca, uma amostra aleatória de 15 golfistas profissionais totalizou os escores de 279, 281, 278, 279, 276, 280, 280, 277, 282, 278, 281, 288 (nossa!), 276, 279 e 280. Use o teste de sinais ao nível 0,05 de significância para testar a hipótese nula $\tilde{\mu} = 278$ (que o escore mediano dos golfistas profissionais naquele campo é 278, dois abaixo do par) contra a hipótese alternativa $\tilde{\mu} > 278$. Baseie o teste na

(a) Tabela V;
(b) na aproximação normal da distribuição binomial.

18.6 Use um computador para refazer a parte (a) do exercício precedente.

18.7 As milhagens por galão que foram obtidas com 40 tanques cheios de certa marca de gasolina são as seguintes:

24,1	25,0	24,8	24,3	24,2	25,3	24,2	23,6
24,5	24,4	24,5	23,2	24,0	23,8	23,8	25,3
24,5	24,6	24,0	25,2	25,2	24,4	24,7	24,1
24,6	24,9	24,1	25,8	24,2	24,2	24,8	24,1
25,6	24,5	25,1	24,6	24,3	25,2	24,7	23,3

Use o teste de sinais baseado na aproximação normal da distribuição binomial para testar a hipótese nula $\tilde{\mu} = 24{,}2$ (que a mediana da população de milhagens amostradas é de 24,2 milhas por galão) contra a hipótese alternativa $\tilde{\mu} > 24{,}2$. Use o nível 0,01 de significância.

18.8 Os números de passageiros em dezesseis dias nos vôos de ida e de volta entre Los Angeles e Chicago foram os seguintes: 199 e 232, 231 e 265, 236 e 250, 238 e 251, 218 e 226,

258 e 269, 253 e 247, 248 e 252, 220 e 245, 237 e 245, 239 e 235, 248 e 260, 239 e 245, 240 e 240, 233 e 239, 247 e 236. Use o teste de sinais baseado na Tabela V e o nível 0,03 de significância para testar a hipótese nula $\tilde{\mu}_D = 0$, onde $\tilde{\mu}_D$ é a mediana das diferenças da população amostrada, contra a hipótese alternativa $\tilde{\mu}_D < 0$.

18.9 Use um computador para refazer o Exercício 18.8.

18.10 Os números de faltas em 20 dias dos funcionários de duas repartições governamentais foram os seguintes: 29 e 24, 45 e 32, 38 e 38, 39 e 34, 46 e 42, 35 e 41, 42 e 36, 39 e 37, 40 e 45, 38 e 35, 31 e 37, 44 e 35, 42 e 40, 40 e 32, 42 e 45, 51 e 38, 36 e 33, 45 e 39, 33 e 28, 32 e 38. Use o teste de sinais ao nível 0,05 de significância para testar a hipótese nula $\tilde{\mu}_D = 0$, onde $\tilde{\mu}_D$ é a mediana da população de diferenças entre as faltas diárias das duas repartições governamentais. Use a hipótese alternativa $\tilde{\mu}_D > 0$.

18.11 Use a aproximação normal da distribuição binomial para refazer o Exercício 18.10.

18.12 Os números de artefatos descobertos em 30 dias por dois arqueólogos numa antiga população indígena abandonada foram os seguintes: 2 e 0, 4 e 1, 2 e 0, 0 e 1, 2 e 0, 3 e 1, 1 e 2, 4 e 0, 2 e 3, 3 e 2, 1 e 0, 2 e 6, 5 e 2, 3 e 2, 1 e 0, 2 e 1, 1 e 1, 4 e 2, 1 e 1, 1 e 0, 0 e 2, 3 e 1, 2 e 1, 2 e 0, 0 e 0, 1 e 3, 4 e 1, 2 e 1, 1 e 1, e 3 e 0. Use o teste de sinais ao nível 0,05 de significância para testar a hipótese nula de que os dois arqueólogos são igualmente bons para encontrar artefatos contra a hipótese alternativa de que eles não são igualmente bons.

18.3 O TESTE DE SINAL COM POSTO

O teste de sinais é fácil de aplicar e possui um apelo intuitivo, mas desperdiça informação porque utiliza somente os sinais das diferenças entre as observações e $\tilde{\mu}_0$, no caso de uma amostra, ou os sinais das diferenças entre os pares de observações, no caso de pares de amostras. É por essa razão que geralmente é preferido um teste não-paramétrico alternativo, o **teste de sinais com posto** (também conhecido como **teste de sinais com posto de Wilcoxon**).

Nesse teste, ordenamos as diferenças independentemente de seus sinais, atribuindo o posto 1 à menor diferença numérica (isto é, à menor diferença em valor absoluto), o ponto 2 à segunda menor diferença numérica,..., e o posto n à maior diferença numérica. Novamente descartamos as diferenças zero e, se duas ou mais diferenças são numericamente iguais, atribuímos a cada uma delas a média dos postos que ocupam conjuntamente. Baseamos, então, o teste em T^+, que é a soma dos postos das diferenças positivas, em T^-, que é a soma dos postos das diferenças negativas, ou em T, a menor das duas somas.

O teste de sinais com posto serve como uma alternativa tanto ao teste de sinais com uma amostra quanto ao teste de sinais com pares de amostras; como tal, ele é aplicável quando a probabilidade de obter um valor inferior à mediana é igual à probabilidade de obter um valor superior à mediana. Ilustraremos esse teste aqui com as medições da octanagem de uma certa marca de gasolina especial, baseadas nas quais testaremos a hipótese nula $\tilde{\mu} = 98{,}5$ contra a hipótese alternativa $\tilde{\mu} < 98{,}5$, ao nível 0,01 de significância.

As medições estão exibidas na coluna da esquerda na tabela a seguir, sendo que na coluna do meio aparecem as diferenças obtidas subtraindo 98,5 de cada medição:

* Como o teste de sinal com posto é uma alternativa ao teste de sinal, esta seção e a Seção 18.4 podem ser omitidas sem perda de continuidade.

Medições	*Diferenças*	*Postos*
97,5	−1,0	4
95,2	−3,3	12
97,3	−1,2	6
96,0	−2,5	10
96,8	−1,7	7
100,3	1,8	8
97,4	−1,1	5
95,3	−3,2	11
93,2	−5,3	14
99,1	0,6	2
96,1	−2,4	9
97,6	−0,9	3
98,2	−0,3	1
98,5	0,0	
94,9	−3,6	13

Depois de descartar a diferença zero, verificamos que a menor diferença numérica é 0,3, a próxima menor diferença numérica é 0,6, a seguinte menor diferença numérica é 0,9,..., e a maior diferença numérica é 5,3. Esses postos estão mostrados na terceira coluna, e segue que

$$T^+ = 8 + 2 = 10$$
$$T^- = 4 + 12 + 6 + 10 + 7 + 5 + 11 + 14 + 9 + 3 + 1 + 13$$
$$= 95$$

e, portanto, $T = 10$. Como $T^+ + T^-$ é igual à soma dos inteiros de 1 a n, ou seja, $\frac{n(n+1)}{2}$, poderíamos ter obtido T^- mais facilmente subtraindo $T^+ = 10$ de $\frac{14 \cdot 15}{2} = 105$. [Não ocorreram empates de posto nesse exemplo mas, como observamos anteriormente, se ocorrerem empates, atribuímos a cada um dos valores empatados (diferenças) a média dos postos que elas ocupam conjuntamente.]

A estreita relação entre T^+, T^- e T é também refletida por suas distribuições amostrais, um exemplo disso, para $n = 5$, sendo ilustrado na Figura 18.2. Como há uma chance meio a meio para cada posto cair numa das diferenças positivas ou numa das diferenças negativas, existe uma totalidade de 2^n possibilidades, cada uma com a probabilidade $(\frac{1}{2})^n$. Para obter as probabilidades associadas com os diversos valores de T^+, T^- e T, contamos o número de maneiras pelas quais esses valores de T^+, T^- e T podem ser obtidos e multiplicamos por $(\frac{1}{2})^n$. Por exemplo, para $n = 5$ e $T^+ = 6$, existem as três possibilidades 1 e 5, 2 e 4, e 1 e 2 e 3, e a probabilidade é $3 \cdot (\frac{1}{2})^5 = \frac{3}{32}$, como mostra a Figura 18.2.

Para simplificar a construção de tabelas de valores críticos, basearemos todos os testes da hipótese nula $\tilde{\mu} = \tilde{\mu}_0$ na distribuição de T, e a rejeitaremos para os valores que caem na cauda esquerda. Devemos ter o cuidado, entretanto, para utilizar a estatística e o valor crítico corretos. Quando $\tilde{\mu} < \tilde{\mu}_0$, então T^+ tende a ser pequeno e, assim, quando a hipótese alternativa é $\tilde{\mu} < \tilde{\mu}_0$, baseamos o teste em T^+; quando $\tilde{\mu} > \tilde{\mu}_0$, então T^- tende a ser pequeno e, assim, quando a hipótese alternativa é $\tilde{\mu} > \tilde{\mu}_0$, baseamos o teste em T^-; e quando $\tilde{\mu} \neq \tilde{\mu}_0$, então ou T^+ ou T^- tende a ser pequeno, e assim, quando a hipótese alternativa é $\tilde{\mu} \neq \tilde{\mu}_0$, baseamos o teste em T. Essas relações estão sintetizadas na tabela seguinte:

Hipótese alternativa	*Rejeitar a hipótese nula se*	*Aceitar a hipótese nula ou reservar julgamento se*
$\tilde{\mu} < \tilde{\mu}_0$	$T^+ \leq T_{2\alpha}$	$T^+ > T_{2\alpha}$
$\tilde{\mu} > \tilde{\mu}_0$	$T^- \leq T_{2\alpha}$	$T^- > T_{2\alpha}$
$\tilde{\mu} \neq \tilde{\mu}_0$	$T \leq T_\alpha$	$T > T_\alpha$

Na Tabela VI, no fim do livro, encontram-se os valores necessários de T_α, que são os maiores valores de T para os quais a probabilidade de $T \leq T_\alpha$ não excede α; os espaços em branco na tabela indicam que a hipótese nula não pode ser rejeitada, independentemente do valor obtido para a estatística de teste. Note que os mesmos valores críticos servem para testes em diversos níveis de significância, dependendo de a hipótese alternativa ser unilateral ou bilateral.

EXEMPLO 18.5 Com referência às medições da octanagem à página 457, use o teste de sinais com posto, ao nível 0,01 de significância, para testar a hipótese nula $\tilde{\mu} = 98{,}5$ contra a hipótese alternativa $\tilde{\mu} < 98{,}5$.

Solução

1. $H_0 : \tilde{\mu} = 98{,}5$
   $H_A : \tilde{\mu} < 98{,}5$
2. $\alpha = 0{,}01$
3. Rejeitar a hipótese nula se $T^+ \leq 16$, onde 16 é o valor de $T_{0{,}02}$ para $n = 14$; caso contrário, aceitá-la ou reservar julgamento.
4. Conforme mostrado à página 457, $T^+ = 10$.
5. Como $T^+ = 10$ é menor do que 16, a hipótese nula deve ser rejeitada. Concluímos que a medição de octanagem mediana da marca de gasolina especial considerada é menor do que 98,5.

Se nesse exemplo tivéssemos optado por usar um computador, teríamos obtido um impresso de MINITAB como o da Figura 18.3. Uma vantagem de utilizar um computador é que não precisamos nos referir a uma tabela especial; a Figura 18.3 dá o valor de p como sendo 0,004. Isso também teria levado à rejeição da hipótese nula.

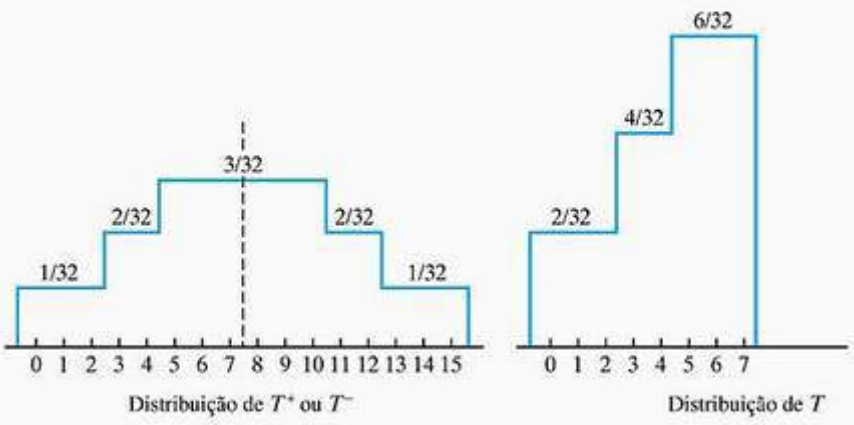

Figura 18.2 Distribuição de T^+ e T^- para $n = 5$.

Teste de Sinais com Posto de Wilcoxon: Octanagem

```
Test of median = 98.50 versus median < 98.50

            N for  Wilcoxon        Estimated
         N  Test Statistic     P    Median
Octanes 15    14      10.0 0.004     96.85
```

Figura 18.3
Impresso de computador para o Exemplo 18.5.

O teste de sinais com posto também pode ser usado como uma alternativa para o teste de sinais para pares de amostras. O procedimento é exatamente o mesmo, mas quando escrevemos a hipótese nula como $\tilde{\mu}_D = 0$, então $\tilde{\mu}_D$ é a mediana da população de diferenças amostrada.

EXEMPLO 18.6 Use o teste de sinais com posto para refazer o Exemplo 18.2. Os dados originais, relativos às perdas semanais médias de horas de trabalho devidas a acidentes em dez indústrias, antes e depois da adoção do programa de segurança, estão exibidos na coluna da esquerda da tabela seguinte. A coluna do meio contém suas diferenças e, descartando os sinais, os postos das diferenças numéricas estão exibidos na coluna da direita.

Perdas de homens-hora antes e depois	*Diferenças*	*Postos*
45 e 36	9	9
73 e 60	13	10
46 e 44	2	2
124 e 119	5	4,5
33 e 35	–2	2
57 e 51	6	7
83 e 77	6	7
34 e 29	5	4,5
26 e 24	2	2
17 e 11	6	7

Assim, $T^- = 2$ e $T^+ = 53$.

Solução

1. $H_0 : \tilde{\mu}_D = 0$, onde $\tilde{\mu}_D$ é a mediana da população de diferenças amostradas (entre as perdas de horas de trabalho antes e depois da adoção do programa se segurança).
   $H_A : \tilde{\mu}_D > 0$
2. $\alpha = 0{,}05$
3. Rejeitar a hipótese nula se $T^- \leq 11$, onde 11 é o valor de $T_{0,10}$ para $n = 10$; caso contrário, aceitar a hipótese nula ou reservar julgamento.
4. Conforme mostrado anteriormente, $T^- = 2$.
5. Como $T^- = 2$ é menor do que 11, a hipótese nula deve ser rejeitada. Concluímos que o programa de segurança é eficaz. (Se nesse exemplo tivéssemos usado um computador, teríamos obtido um valor de p como sendo 0,005, e a conclusão teria sido a mesma.)

*18.4 O TESTE DE SINAL COM POSTO (Grandes Amostras)

Quando n é 15 ou mais, é considerado razoável admitir que as distribuições de T^+ e T^- sejam satisfatoriamente aproximadas por curvas normais. Nesse caso, podemos basear todos os testes em T^+ ou em T^- e, como não interessa qual estatística tomamos, vamos trabalhar aqui com a estatística T^+.

Com base na suposição de que cada diferença tanto pode ser positiva como negativa, pode ser mostrado que a média e o desvio-padrão da distribuição amostral de T^+ são

MÉDIA E DESVIO-PADRÃO DA ESTATÍSTICA T^+

$$\mu_{T^+} = \frac{n(n+1)}{4}$$

e

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Assim, para grandes amostras, o que nesse caso é $n \geq 15$, podemos basear o teste de sinais com posto na estatística

TESTE DE SINAIS COM POSTO PARA GRANDES AMOSTRAS

$$z = \frac{T^+ - \mu_{T^+}}{\sigma_{T^+}}$$

que é um valor de uma variável aleatória aproximadamente de distribuição normal padrão. Quando a hipótese alternativa é $\tilde{\mu} \neq \tilde{\mu}_0$ (ou $\tilde{\mu}_D \neq 0$), rejeitamos a hipótese nula se $z \leq -z_{\alpha/2}$ ou $z \geq z_{\alpha/2}$; quando a hipótese alternativa é $\tilde{\mu} > \tilde{\mu}_0$ (ou $\tilde{\mu}_D > 0$), rejeitamos a hipótese nula se $z \geq z_\alpha$; e quando a hipótese alternativa é $\tilde{\mu} < \tilde{\mu}_0$ (ou $\tilde{\mu}_D < 0$), rejeitamos a hipótese nula se $z \leq z_\alpha$.

EXEMPLO 18.7 Os pesos de 16 pessoas, antes e depois de submeterem-se a uma dieta para emagrecimento durante duas semanas, são os seguintes, em libras: 169,0 e 159,9; 188,6 e 181,3; 222,1 e 209,0; 160,1 e 162,3; 187,5 e 183,5; 202,5 e 197,6; 167,8 e 171,4; 214,3 e 202,1; 143,8 e 145,1; 198,2 e 185,5; 166,9 e 158,6; 142,9 e 145,4; 160,5 e 159,5; 198,7 e 190,6; 149,7 e 149,0; e 181,6 e 183,1. Use o teste de sinais com posto para grandes amostras ao nível 0,05 de significância para testar se a dieta para emagrecimento é eficaz.

Solução

1. $H_0 : \tilde{\mu}_D = 0$, onde $\tilde{\mu}_D$ é a mediana da população de diferenças (entre os pesos antes e depois) amostradas.
   $H_A : \tilde{\mu}_D > 0$
2. $\alpha = 0{,}05$
3. Rejeitar a hipótese nula se $z \geq 1{,}645$, onde

$$z = \frac{T^+ - \mu_{T^+}}{\sigma_{T^+}}$$

   e, caso contrário, aceitar a hipótese nula ou reservar julgamento.
4. Os dados originais, as diferenças e os postos de seus valores absolutos estão exibidos na tabela seguinte:

Pesos antes e depois	*Diferenças*	*Postos*
169,0 e 159,9	9,1	13
188,6 e 181,3	7,3	10
222,1 e 209,0	13,1	16
160,1 e 162,3	−2,2	5
187,5 e 183,5	4,0	8
202,5 e 197,6	4,9	9
167,8 e 171,4	−3,6	7
214,3 e 202,1	12,2	14
143,8 e 145,1	−1,3	3
198,2 e 185,5	12,7	15
166,9 e 158,6	8,3	12
142,9 e 145,4	−2,5	6
160,5 e 159,5	1,0	2
198,7 e 190,6	8,1	11
149,7 e 149,0	0,7	1
181,6 e 183,1	−1,5	4

Decorre que

$$T^+ = 13 + 10 + 16 + 8 + 9 + 14 + 15 + 12 + 2 + 11 + 1 = 111$$

e, como

$$\mu_{T^+} = \frac{16 \cdot 17}{4} = 68 \quad \text{e} \quad \sigma_{T^+} = \sqrt{\frac{16 \cdot 17 \cdot 33}{24}} \approx 19{,}34$$

obtemos finalmente

$$z = \frac{111 - 68}{19{,}34} \approx 2{,}22$$

5. Como $z = 2{,}22$ é maior do que 1,645, a hipótese nula deve ser rejeitada; concluímos que a dieta para emagrecimento é eficaz.

EXERCÍCIOS

***18.13** Em que estatística baseamos nossa decisão, e para que valores da estatística rejeitamos a hipótese nula, se temos uma amostra aleatória de tamanho $n = 10$ e estamos utilizando o teste de sinais com posto ao nível 0,05 de significância para testar a hipótese nula $\tilde{\mu} = \tilde{\mu}_0$ contra a hipótese alternativa

(a) $\tilde{\mu} \neq \tilde{\mu}_0$;
(b) $\tilde{\mu} > \tilde{\mu}_0$;
(c) $\tilde{\mu} < \tilde{\mu}_0$?

***18.14** Refaça o Exercício 18.13 com o nível de significância trocado para 0,01.

***18.15** Em que estatística baseamos nossa decisão, e para que valores da estatística rejeitamos a hipótese nula, se temos uma amostra aleatória de $n = 12$ pares de valores e estamos utilizando o teste de sinais com posto ao nível 0,01 de significância para testar a hipótese nula $\tilde{\mu}_D = 0$ contra a hipótese alternativa

(a) $\tilde{\mu}_D \neq 0$;
(b) $\tilde{\mu}_D > 0$;
(c) $\tilde{\mu}_D < 0$?

***18.16** Refaça o Exercício 18.15 com o nível de significância trocado para 0,05.

***18.17** Numa amostra aleatória de 13 edições, um jornal relacionou 40, 52, 43, 27, 35, 36, 57, 39, 41, 34, 46, 32 e 37 apartamentos para alugar. Use o teste de sinais com posto, ao nível 0,05 de significância, para testar a hipótese nula $\tilde{\mu} = 45$ contra a hipótese alternativa

(a) $\tilde{\mu} < 45$;
(b) $\tilde{\mu} \neq 45$.

***18.18** Use o teste de sinais com posto para refazer o Exercício 18.1.

***18.19** Use o teste de sinais com posto para refazer o Exercício 18.3.

***18.20** Numa amostra aleatória obtida num parque de recreação público, foram necessários 38, 43, 36, 29, 44, 28, 40, 50, 39, 47 e 33 minutos para jogar uma partida de tênis. Use o teste de sinais com posto, ao nível 0,05 de significância, para testar se são, ou não, necessários 35 minutos, em média, para jogar uma partida de tênis naquele parque.

***18.21** Numa amostra aleatória de 15 dias de verão, duas cidades do estado norte-americano do Arizona reportaram as seguintes temperaturas máximas em graus Fahrenheit: 102 e 106, 103 e 110, 106 e 106, 104 e 107, 105 e 108, 102 e 109, 103 e 102, 104 e 107, 110 e 112, 109 e 110, 100 e 104, 110 e 109, 105 e 108, 111 e 114, e 105 e 106. Use o teste de sinais com posto, ao nível 0,05 de significância, para testar a hipótese nula $\tilde{\mu}_D = 0$ contra a hipótese alternativa $\tilde{\mu}_D < 0$.

***18.22** A seguir estão os números de Certificados de Depósito (CD), de três meses e de seis meses, que um banco vendeu numa amostra aleatória de 16 dias úteis: 37 e 32, 33 e 22, 29 e 26, 18 e 33, 41 e 25, 42 e 34, 33 e 43, 51 e 31, 36 e 24, 29 e 22, 23 e 30, 28 e 29, 44 e 30, 24 e 26, 27 e 18, e 30 e 35. Teste ao nível 0,05 de significância se o banco vende igualmente os dois tipos de CD contra a hipótese alternativa de que vende mais CD de três meses, usando

(a) o teste de sinais com posto baseado na Tabela VI;
(b) o teste de sinais com posto para grandes amostras.

***18.23** Use o teste de sinais com posto para grandes amostras para refazer o Exercício 18.21.

***18.24** Use o teste de sinais com posto para grandes amostras para refazer o Exercício 18.7.

***18.25** Use o teste de sinais com posto para grandes amostras para refazer o Exercício 18.10.

***18.26** A seguir é dada uma amostra aleatória das notas obtidas por maridos e suas esposas num teste de reconhecimento espacial:

Maridos	*Esposas*
108	103
104	116
103	106
112	104
99	99
105	94
102	110
112	128
119	106
106	103

Maridos	*Esposas*
125	120
96	98
107	117
115	130
110	101
101	100
103	96
105	99
124	120
113	116

Use o teste de sinais com posto para grandes amostras, ao nível 0,05 de significância, para testar se maridos e esposas se saem igualmente bem nesse teste.

*18.27 Use um computador para refazer o Exercício 18.20.

*18.28 Use um computador para refazer o Exercício 18.21.

*18.29 Use um computador para refazer a parte (a) do Exercício 18.22.

18.5 O TESTE U

Vamos agora apresentar uma alternativa não-paramétrica para teste t de duas amostras relativa à diferença entre duas médias populacionais. Essa alternativa é denominada o **teste *U*** ou, às vezes, **teste da soma de postos de Wilcoxon**, ou ainda, **teste de Mann-Whitney**, em homenagem aos estatísticos que contribuíram para seu desenvolvimento. Os diferentes nomes refletem a maneira na qual são organizados os cálculos; do ponto de vista lógico, esses testes são todos equivalentes.

Com esse teste, conseguimos verificar se duas amostras independentes provêm de populações idênticas. Em particular, podemos testar a hipótese nula $\mu_1 = \mu_2$ sem precisar supor que as populações amostradas tenham aproximadamente a forma de distribuições normais. Na realidade, o teste exige apenas que as populações sejam contínuas (para evitar empates), e mesmo essa exigência não é crítica, desde que o número de empates seja pequeno. Observe, entretanto, que de acordo com o dicionário de termos estatísticos de Kendall e Buckland, o teste U testa a igualdade dos parâmetros de localização de duas populações que são, no mais, idênticas. E é claro que existem muitos parâmetros de localização. Por exemplo, na Figura 18.5 à página 466, os parâmetros de localização em questão são as medianas populacionais. Aqui elas são denotadas por ETA1 e ETA2, onde anteriormente tinham sido denotadas por $\tilde{\mu}_1$ e $\tilde{\mu}_2$.

A fim de ilustrar como aplicar o teste U, suponha que queiramos comparar o tamanho do grão de areia obtido em duas localidades diferentes da superfície da Lua, com base nos diâmetros seguintes (em milímetros):

Localidade 1:	0,37	0,70	0,75	0,30	0,45	0,16	0,62	0,73	0,33	
Localidade 2:	0,86	0,55	0,80	0,42	0,97	0,84	0,24	0,51	0,92	0,69

As médias dessas duas amostras são 0,49 e 0,68, e sua diferença parece grande, mas resta ver se é significante.

Para aplicar o teste U, primeiramente dispomos os dados conjuntamente, como se constituíssem uma única amostra, em ordem crescente de magnitude. Com nossos dados obtemos

0,16	0,24	0,30	0,33	0,37	0,42	0,45	0,51	0,55	0,62
1	2	1	1	1	2	1	2	2	1
0,69	0,70	0,73	0,75	0,80	0,84	0,86	0,92	0,97	
2	1	1	1	2	2	2	2	2	

onde, para cada valor, indicamos se provém da localidade 1 ou da localidade 2. Atribuindo aos dados, nessa ordem, os postos 1, 2, 3, . . ., e 19, vemos que os valores da primeira amostra (localidade 1) ocupam os postos 1, 3, 4, 5, 7, 10, 12, 13 e 14, enquanto que os da segunda amostra (localidade 2) ocupam os postos 2, 6, 8, 9, 11, 15, 16, 17, 18 e 19. Não há empates aqui entre valores nas diferentes amostras, mas, se houvesse, atribuiríamos a cada uma das observações empatadas a média dos postos que elas ocupam conjuntamente. Por exemplo, se o terceiro e o quarto valores são iguais, atribuímos a cada um o posto $\frac{3+4}{2} = 3{,}5$ e se o nono, o décimo e o décimo primeiro valores são iguais, atribuímos a cada um o posto $\frac{9+10+11}{3} = 10$. Quando há empates entre valores pertencentes à mesma amostra, não importa como sejam seus postos. Por exemplo, se o

terceiro e o quarto valores são iguais, mas pertencem à mesma amostra, não importa qual seja considerado como posto 3 e qual como posto 4.

Porém, se há uma diferença considerável entre as médias das duas populações, a maioria dos postos mais baixos tende a acompanhar os valores de uma das amostras, enquanto a maioria dos postos mais altos tende a acompanhar os valores da outra amostra. O teste da hipótese nula, de que as duas amostra provenham de populações idênticas, pode, pois basear-se em W_1, a soma dos postos dos valores da primeira amostra, ou em W_2, a soma dos postos dos valores da segunda amostra. Na prática, não importa qual seja a designada como a amostra 1 e qual seja designada como a amostra 2, nem se baseamos o teste em W_1 ou em W_2. (Quando os tamanhos das amostras são diferentes, costumamos rotular por amostra 1 a menor das duas amostras, mas isso não será necessário para o nosso trabalho.)

Se os tamanhos das amostras são n_1 e n_2, a soma de W_1 e W_2 é simplesmente a soma dos primeiros $n_1 + n_2$ inteiros positivos, que sabemos ser

$$\frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

Essa fórmula permite encontrar W_2 se conhecermos W_1, e vice-versa. Para nossa ilustração obtemos

$$W_1 = 1 + 3 + 4 + 5 + 7 + 10 + 12 + 13 + 14 = 69$$

e como a soma dos 19 primeiros inteiros positivos é $\frac{19 \cdot 20}{2} = 190$, decorre que

$$W_2 = 190 - 69 = 121$$

(Esse valor é a soma dos postos 2, 6, 8, 9, 11, 15, 16, 17, 18 e 19).

Quando primeiro propusemos a utilização de **somas de postos** como uma alternativa não-paramétrica do teste t de duas amostras, a decisão foi baseada em W_1 e W_2. Hoje em dia, é mais comum basear a decisão em uma das duas estatísticas

ESTATÍSTICAS U_1 E U_2

$$U_1 = W_1 - \frac{n_1(n_1 + 1)}{2}$$

ou

$$U_2 = W_2 - \frac{n_2(n_2 + 1)}{2}$$

ou na estatística U, que é sempre igual à menor das duas. Os testes resultantes são equivalentes ao testes baseados em W_1 ou W_2, mas têm a vantagem de prestarem-se mais facilmente à construção de tabelas de valores críticos. Não só os testes U_1 e U_2 tomam valores no mesmo intervalo de 0 a $n_1 n_2$ — de fato, sua soma é sempre igual a $n_1 n_2$ — mas têm distribuições idênticas que são simétricas em relação a $\frac{n_1 n_2}{2}$. A Figura 18.4 ilustra a relação entre as distribuições amostrais de U_1, U_2 e U para o caso especial em que $n_1 = 3$ e $n_2 = 3$.

Como já foi dito anteriormente, supomos que estamos tratando com amostras aleatórias independentes de populações idênticas, mas estamos mais interessados no caso $\mu_1 = \mu_2$. Como na Seção 18.3, basearemos todos os testes na distribuição amostral de uma única e mesma estatística. Entretanto, aqui essa é a estatística U, e rejeitaremos a hipótese nula para valores caindo em sua cauda esquerda. Contudo, novamente devemos ter cuidado para utilizar a estatística e o valor crítico corretos. Se $\mu_1 < \mu_2$, então U_1 tende a ser pequeno e, assim, quando a hipótese alternativa é $\mu_1 < \mu_2$, baseamos o teste em U_1; se $\mu_1 > \mu_2$, então U_2 tende a ser pequeno e assim, quando a hipótese alternativa é $\mu_1 > \mu_2$, baseamos o teste em U_2; e se $\mu_1 \neq \mu_2$, então ou U_1 ou U_2 tende a ser pequeno, e assim, quando a hipótese alternativa é $\mu_1 \neq \mu_2$, baseamos o teste em U. Tudo isso está sintetizado na tabela seguinte:

Hipótese alternativa	*Rejeitar a hipótese nula se*	*Aceitar a hipótese nula ou reservar julgamento se*
$\mu_1 < \mu_2$	$U_1 \leq U_{2\alpha}$	$U_1 > U_{2\alpha}$
$\mu_1 > \mu_2$	$U_2 \leq U_{2\alpha}$	$U_2 > U_{2\alpha}$
$\mu_1 \neq \mu_2$	$U \leq U_\alpha$	$U > U_\alpha$

Os valores necessários de U_α, que são os maiores valores de U para os quais a probabilidade de $U \leq U_\alpha$ não é maior do que α, podem ser encontrados na Tabela VII, no fim do livro; os espaços em branco na tabela indicam que a hipótese nula não pode ser rejeitada, independentemente do valor obtido para a estatística de teste. Note que os mesmos valores críticos servem para testes em diferentes níveis de significância, dependendo de a hipótese alternativa ser unilateral ou bilateral.

EXEMPLO 18.8 Com referência aos dados dos tamanhos de grãos à página 463, use o teste U ao nível 0,05 de significância para testar se as duas amostras provêm, ou não, de populações com médias iguais.

Solução

1. $H_0 : \mu_1 = \mu_2$
   $H_A : \mu_1 \neq \mu_2$
2. $\alpha = 0{,}05$
3. Rejeitar a hipótese nula se $U \leq 20$, onde 20 é o valor de $U_{0{,}05}$ para $n_1 = 9$ e $n_2 = 10$; caso contrário, reservar julgamento.
4. Já tendo mostrado à página 464 que $W_1 = 69$ e $W_2 = 121$, obtemos

$$U_1 = 69 - \frac{9 \cdot 10}{2} = 24$$

$$U_2 = 121 - \frac{10 \cdot 11}{2} = 66$$

   e, portanto, $U = 24$. Note que $U_1 + U_2 = 24 + 66 = 90$, que é igual a $n_1 n_2 = 9 \cdot 10$.
5. Como $U = 24$ é maior do que 20, a hipótese nula não pode ser rejeitada; em outras palavras, não podemos concluir que haja diferença na média do tamanho do grão de areia das duas localidades na Lua.

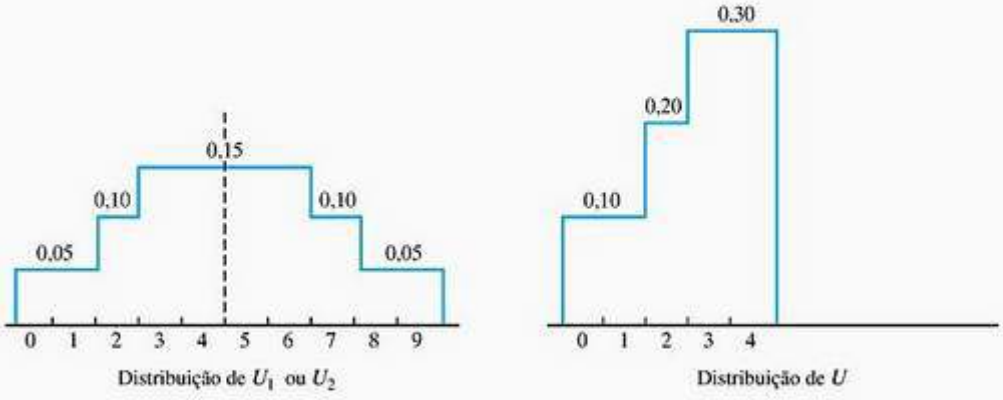

Figura 18.4 Distribuição de U_1, U_2 e U para $n_1 = 3$ e $n_2 = 3$.

```
Teste de Mann-Whitney e C1: Loc. 1, Loc. 2

Loc. 1     N = 9     Median =      0.4500
Loc. 2     N = 10    Median =      0.7450

W = 69.0
Test of ETA1 = ETA2  vs ETA1 not = ETA2 is significant at 0.0942

Cannot reject at alpha = 0.05
```

Figura 18.5 Impresso de computador para o Exemplo 18.8.

Se nesse exemplo tivéssemos optado por usar um computador, o MINITAB teria produzido a Figura 18.5. No impresso, ETA (a letra grega η) denota a população mediana, que antes tinha sido denotada por $\tilde{\mu}$. Também $W = 69$ é a estatística que antes foi denominada W_1, e o valor de p é novamente 0,0942. Como 0,0942 excede 0,05, concluímos (como antes) que a hipótese nula não pode ser rejeitada.

EXEMPLO 18.9 Os tempos de queima (arredondados até o décimo de minuto mais proximo) de amostras aleatórias de dois tipos de sinais de emergências:

Marca 1:	17,2	18,1	19,3	21,1	14,4	13,7	18,8	15,2	20,3	17,5
Marca 2:	13,6	19,1	11,8	14,6	14,3	22,5	12,3	13,5	10,9	14,8

Use o teste U ao nível 0,05 de significância para testar se é razoável dizer que, em média, os sinais da marca 1 são melhores (duram mais) do que os sinais da marca 2.

Solução

1. $H_0 : \mu_1 = \mu_2$
   $H_A : \mu_1 > \mu_2$
2. $\alpha = 0{,}05$
3. Rejeitar a hipótese nula se $U_2 \leq 27$, onde 27 é o valor de $U_{0,10}$ para $n_1 = 10$ e $n_2 = 10$; caso contrário, aceitá-la ou reservar julgamento.
4. Dispondo os dados conjuntamente de acordo com o tamanho, vemos que os valores da segunda amostra ocupam os postos 5, 16, 2, 9, 7, 20, 3, 4, 1 e 10, de modo que
$$W_2 = 5 + 16 + 2 + 9 + 7 + 20 + 3 + 4 + 1 + 10 = 77$$
e
$$U_2 = 77 - \frac{10 \cdot 11}{2} = 22$$
5. Como $U_2 = 22$ é menor do que 27, a hipótese nula deve ser rejeitada; concluímos que os sinais da marca 1 são, realmente, melhores do que os da marca 2.

18.6 O TESTE U (Grandes Amostras)

O teste U para grandes amostras pode ser baseado tanto em U_1 quanto em U_2, conforme definido à página 464, mas como os testes resultantes são equivalentes, não importando qual amostra denotamos por amostra 1 e qual denotamos por amostra 2, vamos utilizar aqui a estatística U_1.

Com base na suposição de que as duas amostras provenham de populações contínuas idênticas, pode ser mostrado que a média e o desvio-padrão da distribuição amostral de U_1 são*

MÉDIA E DESVIO-PADRÃO DA ESTATÍSTICA U_1

$$\mu_{U_1} = \frac{n_1 n_2}{2}$$

e

$$\sigma_{U_1} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Observe que essas fórmulas permanecem as mesmas quando permutamos os índices 1 e 2, mas isso não deve constituir surpresa – como salientamos à página 464, as distribuições de U_1 e U_2 são as mesmas.

Além disso, se n_1 e n_2 são ambos maiores do que 8, a distribuição amostral de U_1 pode ser satisfatoriamente aproximada por uma distribuição normal. Assim, baseamos o teste da hipótese nula $\mu_1 = \mu_2$ na estatística

ESTATÍSTICA PARA O TESTE U DE GRANDES AMOSTRAS

$$z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

que tem aproximadamente a distribuição normal padrão. Quando a hipótese alternativa é $\mu_1 \neq \mu_2$, rejeitamos a hipótese nula se $z \leq -z_{\alpha/2}$ ou $z \geq z_{\alpha/2}$; quando a hipótese alternativa é $\mu_1 > \mu_2$, rejeitamos a hipótese nula se $z \geq z_\alpha$; e quando a hipótese alternativa é $\mu_1 < \mu_2$, rejeitamos a hipótese nula se $z \leq -z_{\alpha/2}$.

EXEMPLO 18.10 Os aumentos de peso de duas amostras aleatórias de perus novos alimentados com duas rações diferentes, mas, afora isso, mantidos em condições idênticas, são os seguintes, em libras:

Ração 1:	16,3	10,1	10,7	13,5	14,9	11,8	14,3	10,2
	12,0	14,7	23,6	15,1	14,5	18,4	13,2	14,0
Ração 2:	21,3	23,8	15,4	19,6	12,0	13,9	18,8	19,2
	15,3	20,1	14,8	18,9	20,7	21,1	15,8	16,2

Use o teste U para grandes amostras ao nível 0,01 de significância, para testar a hipótese nula de que as duas populações amostradas são idênticas, contra a hipótese alternativa de que, em média, a segunda ração produz um aumento maior de peso.

Solução

1. $H_0 : \mu_1 = \mu_2$ (populações são idênticas)

   $H_A : \mu_1 < \mu_2$

2. $\alpha = 0{,}01$

3. Rejeitar a hipótese nula se $z \leq -2{,}33$, onde

$$z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

* Quando há empates em postos, a fórmula do desvio-padrão dá apenas uma aproximação mas, a menos que o número de empates seja muito grande, raramente há a necessidade de fazer alguma correção.

e, caso contrário, aceitar a hipótese nula ou reservar julgamento.

4. Dispondo os dados conjuntamente de acordo com o tamanho, vemos que os valores da primeira amostra ocupam os postos 21; 1; 3; 8; 15; 4; 11; 2; 5,5; 13; 31; 16; 12; 22; 7; e 10. (O quinto e sexto valores são iguais a 12,0, de modo que atribuímos a cada um deles o posto 5,5.) Assim,

$$\begin{aligned} W_1 &= 1+2+3+4+5{,}5+7+8+10+11+12+13 \\ &\quad +15+16+21+22+31 \\ &= 181{,}5 \end{aligned}$$

e

$$U_1 = 181{,}5 - \frac{16 \cdot 17}{2} = 45{,}5$$

Como $\mu_{U_1} = \frac{16 \cdot 16}{2} = 128$ e $\sigma_{U_1} = \sqrt{\frac{16 \cdot 16 \cdot 33}{12}} \approx 26{,}53$, segue que

$$z = \frac{45{,}5 - 128}{26{,}53} \approx -3{,}11$$

5. Como $z = -3{,}11$ é menor do que $-2{,}33$, a hipótese nula deve ser rejeitada; concluímos que, em média, a segunda ração produz maior aumento de peso.

18.7 O TESTE *H*

O **teste *H***, ou **teste de Kruskal-Wallis**, é um teste de soma de postos que serve para testar a suposição de que k amostras aleatórias independentes provêm de populações idênticas e, em particular, a hipótese nula de que $\mu_1 = \mu_2 = \cdots = \mu_k$, contra a hipótese alternativa de que essas médias não são todas iguais. Ao contrário do teste padrão que ele substitui, a análise da variância de um critério, da Seção 15.3, o teste H não exige a suposição de que as populações amostradas tenham, pelo menos aproximadamente, distribuições normais.

Tal como no teste U, os dados são dispostos conjuntamente de baixo para cima, como se constituíssem uma única amostra. Então, se R_i é a soma dos postos atribuídos aos n_i valores da i-ésima amostra e $n = n_1 + n_2 + \cdots + n_k$, o teste H é baseado na estatística

ESTATÍSTICA PARA O TESTE *H*

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Se a hipótese nula é verdadeira e se cada amostra tem, pelo menos, cinco observações, em geral é considerado razoável aproximar a distribuição amostral de H por uma distribuição qui-quadrado com $k-1$ graus de liberdade. Conseqüentemente, rejeitamos a hipótese nula $\mu_1 = \mu_2 = \cdots = \mu_k$ e aceitamos a hipótese alternativa de que essas médias não são todas iguais, quando o valor obtido para H é maior do que ou igual a χ^2_α para $k-1$ graus de liberdade.

EXEMPLO 18.11 Distribuem-se os estudantes aleatoriamente por grupos que estudam espanhol por três métodos diferentes: (1) ensino em sala de aula e laboratório de linguagem, (2) apenas ensino em sala de aula e (3) apenas auto-estudo em laboratório de linguagem. São apresentadas, a seguir, as notas do exame final de amostras de estudantes dos três grupos:

Método 1:	94	88	91	74	86	97	
Método 2:	85	82	79	84	61	72	80
Método 3:	89	67	72	76	69		

Use o teste H ao nível 0,05 de significância para testar a hipótese nula de que as populações originais são idênticas, contra a hipótese alternativa de que suas médias não são todas iguais.

Solução

1. $H_0 : \mu_1 = \mu_2 = \mu_3$ (As populações são idênticas.)
   $H_A :$ μ_1, μ_2 e μ_3 não são todas iguais.
2. $\alpha = 0{,}05$
3. Rejeitar a hipótese nula se $H \geq 5{,}991$, que é o valor de $\chi^2_{0,05}$ para $3 - 1 = 2$ graus de liberdade; caso contrário, aceitá-la ou reservar julgamento.
4. Dispondo os dados conjuntamente de acordo com o tamanho, obtemos 61, 67, 69, 72, 72, 74, 76, 79, 80, 82, 84, 85, 86, 88, 89, 91, 94 e 97. Atribuindo aos dados, nessa ordem, os postos 1, 2, 3, . . ., e 18, obtemos

$$R_1 = 6 + 13 + 14 + 16 + 17 + 18 = 84$$
$$R_2 = 1 + 4{,}5 + 8 + 9 + 10 + 11 + 12 = 55{,}5$$
$$R_3 = 2 + 3 + 4{,}5 + 7 + 15 = 31{,}5$$

   e segue que

$$H = \frac{12}{18 \cdot 19}\left(\frac{84^2}{6} + \frac{55{,}5^2}{7} + \frac{31{,}5^2}{5}\right) - 3 \cdot 19$$
$$\approx 6{,}67$$

5. Como $H = 6{,}67$ é maior do que 5,991, a hipótese nula deve ser rejeitada; concluímos que os três métodos de ensino não são todos igualmente eficazes.

Se tivéssemos utilizado um computador para esse exemplo, teríamos encontrado que o valor p correspondente a $H = 6{,}67$ é 0,036, e que o valor p ajustado ao empate também é 0,036. Como 0,036 é menor do que 0,05, teríamos concluído, como antes, que a hipótese nula deve ser rejeitada.

EXERCÍCIOS

18.30 Em que estatística baseamos a decisão e para quais valores da estatística rejeitamos a hipótese nula $\mu_1 = \mu_2$ se temos amostras aleatórias de tamanhos $n_1 = 9$ e $n_2 = 9$ e estamos usando o teste U baseado na Tabela VII e o nível 0,05 de significância para testar a hipótese nula contra a hipótese alternativa
(a) $\mu_1 > \mu_2$;
(b) $\mu_1 \neq \mu_2$;
(c) $\mu_1 < \mu_2$?

18.31 Refaça o Exercício 18.30 com o nível de significância trocado para 0,01.

18.32 Em que estatística baseamos a decisão e para quais valores da estatística rejeitamos a hipótese nula $\mu_1 = \mu_2$ se temos amostras aleatórias de tamanhos $n_1 = 10$ e $n_2 = 14$ e estamos usando o teste U baseado na Tabela VII e o nível 0,01 de significância para testar a hipótese nula contra a hipótese alternativa
(a) $\mu_1 > \mu_2$;
(b) $\mu_1 \neq \mu_2$;
(c) $\mu_1 < \mu_2$?

18.33 Refaça o Exercício 18.32 com o nível de significância trocado para 0,05.

18.34 Em que estatística baseamos a decisão e para quais valores da estatística rejeitamos a hipótese nula $\mu_1 = \mu_2$ contra a hipótese alternativa $\mu_1 \neq \mu_2$ se estamos usando o teste U baseado na Tabela VII e o nível 0,05 de significância, e

(a) $n_1 = 4$ e $n_2 = 6$;
(b) $n_1 = 9$ e $n_2 = 8$;
(c) $n_1 = 5$ e $n_2 = 12$;
(d) $n_1 = 7$ e $n_2 = 3$?

18.35 Refaça o Exercício 18.34 com a hipótese alternativa trocada para $\mu_1 > \mu_2$.

18.36 Explique por que não há valor na Tabela VII para $U_{0,05}$ correspondente a $n_1 = 3$ e $n_2 = 3$. (*Sugestão*: Recorra à Figura 18.4.)

18.37 As notas obtidas por amostras aleatórias de estudantes de dois grupos de minorias num teste de eventos contemporâneos foram as seguintes:

Minoria 1:	73	82	39	68	91	75
	89	67	50	86	57	65
Minoria 2:	51	42	36	53	88	59
	49	66	25	64	18	76

Use o teste U baseado na Tabela VII para testar ao nível 0,05 de significância se pode ser esperado que os estudantes das duas minorias se saiam igualmente bem nesse teste.

18.38 Use um computador para refazer o Exercício 18.37.

18.39 Os números de minutos que amostras aleatórias de 15 homens e 12 mulheres levaram para completar um teste escrito para renovação de sua carteira de motorista são os seguintes:

Homem:	9,9	7,4	8,9	9,1	7,7	9,7	11,8	7,5
	9,2	10,0	10,2	9,5	10,8	8,0	11,0	
Mulher:	8,6	10,9	9,8	10,7	9,4	10,3		
	7,3	11,5	7,6	9,3	8,8	9,6		

Use o teste U baseado na Tabela VII para testar ao nível 0,05 de significância se vale $\mu_1 = \mu_2$ ou não, onde μ_1 e μ_2 são os tempos médios que homens e mulheres levam para completar o teste.

18.40 Use o teste U para grandes amostras para refazer o Exercício 18.39.

18.41 Os números de Rockwell da dureza de seis moldes de alumínio selecionados aleatoriamente do lote de produção A e oito moldes selecionados do lote B são os seguintes:

Lote de Produção A:	75	56	63	70	58	74		
Lote de Produção A:	63	85	77	80	86	76	72	82

Use o teste U baseado na Tabela VII para testar ao nível 0,05 de significância se a dureza dos moldes do lote B é, em média, superior à dos moldes do lote A.

18.42 Use um computador para refazer o Exercício 18.41.

18.43 Use o teste U para grandes amostras para refazer o Exemplo 18.8.

18.44 Use o teste U para grandes amostras para refazer o Exemplo 18.9.

18.45 Use um computador para refazer o Exemplo 18.10.

18.8 TESTES DE ALEATORIEDADE: REPETIÇÕES

Todos os métodos de inferência estudados neste livro baseiam-se na suposição de que nossas amostras sejam aleatórias; todavia, há muitas aplicações em que é difícil determinar se tal suposição é justificável. Isso é verdade, particularmente, quando temos pouco ou nenhum controle sobre a seleção dos dados, como é o caso, por exemplo, quando nos baseamos nos dados que estão disponíveis para fazer previsões meteorológicas a longo prazo, quando utilizamos os dados que estão disponíveis para estimar a taxa de mortalidade de uma doença, ou quando utilizamos registros de vendas de meses passados para fazer previsões sobre as vendas futuras de uma loja de departamentos. Nenhuma dessas informações constitui uma amostra aleatória no sentido estrito.

Existem vários métodos para julgar a aleatoriedade de uma amostra com base na ordem em que se obtêm as informações; eles nos permitem decidir, depois de coletados os dados, se os padrões que parecem suspeitamente não-aleatórios podem ser atribuídos ao acaso. A técnica que vamos desenvolver aqui e nas duas próximas seções, o **teste *u***, é baseada na **teoria de repetições**.

Uma **repetição** é uma sucessão de letras (ou qualquer outro símbolo) idênticas seguida e precedida por letras diferentes (ou por nenhuma letra). A título de ilustração, consideremos o seguinte arranjo de árvores sãs, *S*, e doentes, *D*, plantadas há muitos anos ao longo de uma estrada secundária:

$$\underline{H\,H\,H\,H}\quad \underline{D\,D\,D}\quad \underline{H\,H\,H\,H\,H\,H\,H}\quad \underline{D\,D}\quad \underline{H\,H}\quad \underline{D\,D\,D\,D}$$

Sublinhando as letras que constituem as repetições, vemos que há repetições de quatro *S*, depois uma repetição de três *D*, depois uma repetição de sete *S*, depois uma repetição de dois *D*, depois uma repetição de dois *S* e finalmente uma repetição de quatro *D*.

O **número total de repetições** que aparecem num arranjo desse tipo costuma ser uma boa indicação de uma possível falta de aleatoriedade. Se há muito poucas repetições, podemos suspeitar de um agrupamento ou conglomerado definido, ou talvez de uma tendência; se há demasiado repetições, suspeitamos de algum padrão alternativo repetido ou cíclico. No exemplo precedente parece haver um conglomerado definido – as árvores doentes parecem vir aos grupos – mas resta saber se isso é realmente significante ou se pode ser atribuído ao acaso.

Se há n_1 letras de um tipo, n_2 letras de um outro tipo, e u repetições, baseamos esse tipo de decisão no seguinte critério:

Rejeitar a hipótese nula de aleatoriedade se

$$u \leqslant u'_{\alpha/2} \quad \text{ou} \quad u \geqslant u_{\alpha/2}$$

onde $u'_{\alpha/2}$ e $u_{\alpha/2}$ são dados na Tabela VIII para valores de n_1 e n_2 até 15, e $\alpha = 0{,}05$ e $\alpha = 0{,}01$.

Na construção da Tabela VIII, $u'_{\alpha/2}$ é o maior valor de u para o qual a probabilidade de $u \leq u'_{\alpha/2}$ não excede $\alpha/2$, e $u_{\alpha/2}$ é o menor valor de u para o qual a probabilidade de $u \geq u_{\alpha/2}$ não excede $\alpha/2$; os espaços em branco na tabela indicam que a hipótese nula não pode ser rejeitada para valores naquela cauda da distribuição amostral, independentemente do valor obtido para u.

EXEMPLO 18.12 Com referência ao arranjo das árvores sãs e doentes citado anteriormente, aplique o teste u ao nível 0,05 de significância para testar a hipótese nula de aleatoriedade contra a hipótese alternativa de que o arranjo não é aleatório.

Solução 1. H_0: O arranjo é aleatório.
H_A: O arranjo não é aleatório.

2. $\alpha = 0{,}05$.

3. Rejeitar a hipótese nula se $u \leq 6$ ou $u \geq 17$, onde 6 e 17 são os valores de $u'_{0{,}025}$ e $u_{0{,}025}$ para $n_1 = 13$ e $n_2 = 9$; caso contrário, aceitá-la ou reservar julgamento.

4. $u = 6$, por inspeção dos dados.

5. Como $u = 6$ é igual ao valor de $u'_{0{,}025}$, a hipótese nula deve ser rejeitada; concluímos que o arranjo de árvores sãs e doentes não é aleatório. Há menos repetições do que seria esperado e parece que as árvores doentes dispõem-se em conglomerados. ■

18.9 TESTES DE ALEATORIEDADE: REPETIÇÕES (Grandes Amostras)

Sob a hipótese nula de que n_1 letras de um tipo e n_2 letras de um outro tipo estejam dispostas aleatoriamente, pode ser mostrado que a média e o desvio-padrão de u, o número total de repetições, são

MÉDIA E DESVIO-PADRÃO DE u

$$\mu_u = \frac{2n_1n_2}{n_1 + n_2} + 1$$

e

$$\sigma_u = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

Além disso, se nem n_1 nem n_2 é menor do que 10, a distribuição amostral de u pode ser aproximada satisfatoriamente por uma distribuição normal. Assim, baseamos o teste da hipótese nula de aleatoriedade na estatística

ESTATÍSTICA PARA O TESTE u DE GRANDES AMOSTRAS

$$z = \frac{u - \mu_u}{\sigma_u}$$

que tem aproximadamente a distribuição normal padrão. Se a hipótese alternativa é que a disposição das letras não é aleatória, rejeitamos a hipótese nula para $z \leq -z_{\alpha/2}$ ou $z \geq z_{\alpha/2}$; se a hipótese alternativa é que há um aglomerado ou uma tendência, rejeitamos a hipótese nula para $z \leq -z_{\alpha/2}$; e se a hipótese alternativa é que há um padrão alternante, ou cíclico, rejeitamos a hipótese nula para $z \geq z_{\alpha/2}$.

EXEMPLO 18.13 O arranjo de homens, H, e mulheres, M, numa fila única para compra de entradas para um concerto de rock é o seguinte:

$$H\ M\ H\ M\ H\ H\ H\ M\ H\ M\ H\ H\ H\ M\ M\ H\ H\ H\ H\ M\ M\ H\ M\ H$$
$$H\ H\ M\ H\ H\ H\ M\ M\ M\ H\ M\ H\ H\ H\ M\ H\ M\ H\ H\ H\ H\ M\ M\ H$$

Teste a aleatoriedade ao nível 0,05 de significância.

Solução 1. H_0: O arranjo é aleatório.
H_A: O arranjo não é aleatório.

2. $\alpha = 0{,}05$

3. Rejeitar a hipótese nula, se $z \leq -1{,}96$ ou $z \geq 1{,}96$, onde

$$z = \frac{u - \mu_u}{\sigma_u}$$

e, caso contrário, aceitar a hipótese nula ou reservar julgamento.

4. Como $n_1 = 30$, $n_2 = 18$ e $u = 27$, obtemos

$$\mu_u = \frac{2 \cdot 30 \cdot 18}{30 + 18} + 1 = 23{,}5$$

$$\sigma_u = \sqrt{\frac{2 \cdot 30 \cdot 18(2 \cdot 30 \cdot 18 - 30 - 18)}{(30 + 18)^2(30 + 18 - 1)}} = 3{,}21$$

e, então

$$z = \frac{27 - 23{,}5}{3{,}21} \approx 1{,}09$$

5. Como $z = 1{,}09$ cai entre $-1{,}96$ e $1{,}96$, a hipótese nula não pode ser rejeitada; em outras palavras, não há evidência real que possa sugerir de que o arranjo não seja aleatório.

18.10 TESTES DE ALEATORIEDADE: REPETIÇÕES ACIMA E ABAIXO DA MEDIANA

O teste u não está limitado ao teste da aleatoriedade de seqüências de atributos, como os S e D, ou H e M, de nossos exemplos. Qualquer amostra que consista em medidas ou observações numéricas pode ser tratada de modo análogo, utilizando-se as letras a e b para denotar valores que estão acima e abaixo da mediana amostral. Omitem-se os números iguais à mediana. A seqüência resultante de a e b (representando os dados em sua ordem original) pode, então, ser testada quanto à aleatoriedade com base no número total de repetições de a e b, a saber, o número total de **repetições acima e abaixo da mediana**. Dependendo do tamanho de n_1 e n_2, usamos a Tabela VIII ou o teste de grandes amostras da Seção 18.9.

EXEMPLO 18.14 Em 24 viagens (repetições) sucessivas entre duas cidades, um ônibus transportou

24	19	32	28	21	23	26	17	20	28	30	24
13	35	26	21	19	29	27	18	26	14	21	23

passageiros. Usando o número total de repetições acima e abaixo da mediana, teste ao nível 0,01 de significância se é razoável tratar esses dados como constituindo uma amostra aleatória.

Solução Como a mediana dos dados é 23,5 obtemos o seguinte arranjo de valores acima e abaixo dela:

$$a\ b\ a\ a\ b\ b\ a\ b\ b\ a\ a\ a\ b\ a\ a\ b\ b\ a\ a\ b\ a\ b\ b\ b$$

1. H_0: O arranjo é aleatório.
   H_A: O arranjo não é aleatório.

2. $\alpha = 0{,}01$

3. Rejeitar a hipótese nula se $u \leq 6$ ou $u \geq 20$, onde 6 e 20 são os valores de $u'_{0{,}005}$ e $u_{0{,}005}$ para $n_1 = 12$ e $n_2 = 12$; caso contrário, aceitar a hipótese nula ou reservar julgamento.

4. $u = 14$ por inspeção do arranjo precedente dos a e b.

5. Como $u = 14$ cai entre 6 e 20, a hipótese nula não pode ser rejeitada; em outras palavras, não há evidência real para indicar que os dados não constituam uma amostra aleatória.

EXERCÍCIOS

18.55 A ordem em que um corretor recebeu 25 ordens para comprar, C, ou para vender, V, de uma certa ação é:

$$S\ S\ S\ B\ B\ B\ B\ B\ B\ B\ B\ S\ B\ S\ S\ S\ S\ S\ S\ B\ B\ B\ B\ S\ S$$

Use a Tabela VIII para testar a aleatoriedade ao nível 0,05 de significância.

18.56 Use o teste de grandes amostras para refazer o Exercício 18.55.

18.57 Um motorista compra gasolina ou num posto da Petrobrás, P, ou num posto da Ipiranga, I, e o arranjo a seguir mostra a ordem dos postos dos quais comprou gasolina 29 vezes do longo de um determinado período de tempo:

$$A\ C\ A\ C\ C\ A\ C\ A\ C\ A\ C\ C\ A\ A\ A\ C\ A\ C\ C\ C\ A\ C\ A\ A\ A\ C\ C\ A\ C$$

Use a Tabela VIII para testar a aleatoriedade ao nível 0,01 de significância.

18.58 Use o teste de grandes amostras para refazer o Exercício 18.57.

18.59 Teste, ao nível 0,05 de significância, se pode ser considerado aleatório o arranjo seguinte de motores defeituosos, D, e não-defeituosos, N, provenientes de uma linha de montagem:

$$N\ N\ N\ N\ N\ N\ N\ D\ N\ N\ N\ N\ N\ D\ D\ D\ N\ N\ N\ D\ D\ D\ N\ N$$

18.60 O arranjo a seguir indica se 60 pessoas, entrevistadas consecutivamente por um pesquisador, são favoráveis, F, ou contrários, C, a um aumento de 1% no imposto sobre a gasolina para financiar a recuperação de estradas:

$$C\ F\ F\ F\ C\ C\ C\ C\ F\ C\ C\ F\ F\ C\ C\ C\ C\ F\ F\ F\ C\ F\ C\ C\ C\ F\ F\ F\ C\ C$$
$$C\ C\ F\ F\ F\ F\ C\ C\ F\ C\ C\ F\ F\ F\ F\ F\ C\ F\ C\ C\ C\ C\ F\ F\ C\ F\ F\ F\ F\ C$$

Teste a aleatoriedade ao nível 0,01 de significância.

18.61 Para testar se um sinal de rádio contém uma mensagem ou se constitui um ruído aleatório, certo intervalo de tempo foi subdividido num número de subintervalos muito pequenos e, para cada um desses, foi determinado se a intensidade do sinal excede, E, ou não excede, N, um determinado nível de ruído de fundo. Ao nível 0,05 de significância, teste se o arranjo seguinte, assim obtido, pode ser considerado aleatório, indicando que o sinal não contém qualquer mensagem e que pode ser considerado como ruído aleatório:

$$E\ N\ N\ N\ N\ E\ N\ E\ N\ N\ N\ E\ E\ N\ N\ N\ E\ E\ N\ E\ N\ E\ N\ N\ N\ E\ E\ N\ N\ N$$
$$N\ N\ E\ E\ N\ E\ N\ N\ E\ N\ N\ N\ E\ E\ E\ N\ N\ N\ E\ N\ E\ N\ E\ N\ N\ N\ N\ N\ E\ N$$

18.62 Jogue uma moeda 50 vezes e teste, ao nível 0,05 de confiança, se a seqüência resultante de caras (K) e coroas (C) pode ser considerada aleatória.

18.63 Registre se 60 carros consecutivos chegando num estacionamento do aeroporto são locais, L, da cidade, ou de outra cidade, O. Teste a aleatoriedade ao nível 0,05 de significância.

18.64 No início da Seção 11.1 foram fornecidos os seguintes números relativos aos minutos que 36 pessoas levaram para montar um brinquedo "fácil de montar": 17, 13, 18, 19, 17, 21, 29, 22, 16, 28, 21, 15, 26, 23, 24, 20, 8, 17, 17, 21, 32, 18, 25, 22, 16, 10, 20, 22, 19, 14, 30, 22, 12, 24, 28 e 11. Teste a aleatoriedade ao nível 0,05 de significância.

18.65 Os pesos de 40 ovelhas de uma certa raça são 66,2; 59,2; 70,8; 58,0; 64,3; 50,7; 62,5; 58,4; 48,7; 52,4; 51,0; 35,7; 62,6; 52,3; 41,2; 61,1; 52,9; 58,8; 64,1; 48,9; 74,3; 50,3; 55,7; 55,5; 51,8; 55,8; 48,9; 51,8; 63,1; 44,6; 47,0; 49,0; 62,5; 45,0; 78,6; 54,2; 72,2; 52,4; 60,5; e 46,8 quilogramas. Sabendo que a mediana desses pesos é 54,85 quilogramas, teste a aleatoriedade ao nível 0,05 de significância.

18.66 As notas de 42 estudantes, na ordem em que eles terminaram um exame:

75	95	77	93	89	83	69	77	92	88	62	64	91	72
76	83	50	65	84	67	63	54	58	76	70	62	65	41
63	55	32	58	61	68	54	28	35	49	82	60	66	57

Teste a aleatoriedade ao nível 0,05 de significância.

18.67 Numa grande cidade, o número total de lojas de varejo iniciando e encerrando atividades durante os anos de 1970 a 1999 foram os seguintes:

107	125	142	147	122	116	153	144	106	138
126	125	129	134	137	143	150	148	152	145
112	162	139	132	122	143	148	155	146	158

Utilizando o fato de que a mediana é 140,5, teste ao nível 0,05 de significância, se há uma tendência significante.

18.68 As vendas trimestrais (em milhões de dólares) durante seis anos de um fabricante de maquinaria pesada são as seguintes:

83,8	102,5	121,0	90,5	106,6	104,8	114,7	93,6
98,9	96,9	122,6	85,6	103,2	96,9	118,0	92,1
100,5	92,9	125,6	79,2	110,8	95,1	125,6	86,7

Utilizando o fato de que a mediana é 99,7, teste ao nível 0,05 de significância, se há um autêntico padrão cíclico.

18.11 CORRELAÇÃO POR POSTO

Como o teste de significância de r dado na Seção 17.3 é baseado em suposições bastante restritivas, às vezes recorremos a uma alternativa não-paramétrica que pode ser aplicada sob condições bem mais gerais. Esse teste da hipótese nula sem correlação é baseado no **coeficiente de correlação por posto**, muitas vezes denominado **coeficiente de correlação por posto de Sperman**, e denotado por r_s.

Para calcular o coeficiente de correlação por posto para um conjunto de pares de dados, primeiro ordenamos os x entre si em ordem crescente ou decrescente; em seguida ordenamos os y da mesma maneira, encontramos a soma dos quadrados das diferenças, d, entre os postos dos x e dos y, e substituímos na fórmula

COEFICIENTE DE CORRELAÇÃO POR POSTO

$$r_S = 1 - \frac{6\left(\sum d^2\right)}{n(n^2 - 1)}$$

onde n é o número de pares de x e y. Quando há empates nos postos, procedemos como anteriormente, atribuindo a cada uma das observações empatadas a média dos postos que elas ocupam conjuntamente.

EXEMPLO 18.15 Os números de horas durante as quais 10 alunos estudaram para um exame, e as notas que obtiveram, são os seguintes:

Número de horas de estudo x	*Nota no exame* y
9	56
5	44
11	79
13	72
10	70
5	54
18	94
15	85
2	33
8	65

Calcule r_s.

Solução Ordenando os x entre si em ordem crescente e também os y, obtemos os postos exibidos nas duas primeiras colunas da tabela seguinte:

Posto de x	*Posto de y*	d	d^2
5	4	1,0	1,00
2,5	2	0,5	0,25
7	8	−1,0	1,00
8	7	1,0	1,00
6	6	0,0	0,00
2,5	3	−0,5	0,25
10	10	0,0	0,00
9	9	0,0	0,00
1	1	0,0	0,00
4	5	−1,0	1,00
			4,50

Observe que o segundo e o terceiro menor valor dentre os x são ambos iguais a 5, de modo que a cada um associamos o posto $\frac{2+3}{2} = 2{,}5$. Então, determinando os d (as diferenças entre os postos) e seus quadrados, e substituindo $n = 10$ e $\sum d^2 = 4{,}50$ na fórmula de r_s, obtemos

$$r_S = 1 - \frac{6(4{,}50)}{10(10^2 - 1)} \approx 0{,}97$$

Como se pode ver nesse exemplo, é fácil calcular r_s manualmente, e essa é a razão por que às vezes ele é usado em lugar de r, quando não dispomos de uma calculadora. Quando não há empates, r_s é efetivamente igual ao coeficiente de correlação r calculado para os dois conjuntos de postos; quando existem empates, pode haver uma pequena diferença (que, em geral, é desprezível). É claro que, trabalhando com postos em lugar dos dados originais, perdemos alguma informação, mas isso geralmente é compensado pela facilidade do cálculo do coeficiente de correlação por posto. É interessante observar que se tivéssemos calculado r para os x e y originais do exemplo precedente, teríamos obtido 0,96, em vez de 0,97; pelo menos nesse caso, a diferença entre r e r_s é muito pequena.

A principal vantagem em utilizar r_s é que podemos testar a hipótese nula de não haver qualquer correlação sem ter de fazer quaisquer suposições sobre as populações amostradas. Sob a hipótese