



Tornando tudo mais fácil!

# Estatística II

PARA  
**LEIGOS**  
FOR  
DUMMIES

## Aprenda a:

- Melhorar suas habilidades em análise de dados
- Selecionar e testar modelos
- Fazer previsões
- Aplicar a Estatística em situações reais

**Deborah Rumsey**

*Autora de Estatística Para Leigos*



A compra deste conteúdo não prevê atendimento e fornecimento de suporte técnico operacional, instalação ou configuração do sistema de leitor de ebooks. Em alguns casos, e dependendo da plataforma, o suporte poderá ser obtido com o fabricante do equipamento e/ou loja de comércio de ebooks.



# Saiba Qual Análise de Dados Usar

Esta tabela vai ajudá-lo a comparar, contrastar e decidir qual análise de dados usar e quando. Use-a para uma consulta rápida ou para uma revisão antes das provas.

<i>Análise</i>	<i>Propósito</i>	<i>Quando Usar</i>	<i>Capítulo</i>
Regressão Linear Simples	Usa $x$ para estimar $y$ através de uma reta	A variável de resposta $y$ é quantitativa; variação constante através de $x$ , que também é quantitativa	4
Regressão Múltipla	Usa várias variáveis $x$ ( $x_i, i = 1 \dots, k$ ) para estimar $y$ através de um plano	$y$ é quantitativa; distribuição normal para cada $x_i$ com variação constante	5
Regressão Não-Linear	Usa $x$ para estimar $y$ através de uma curva	$y$ é quantitativa; distribuição normal; variação constante através de $x$	7
Regressão Logística	Usa $x$ para estimar $p$ = probabilidade da ocorrência de $y$	$y$ é uma variável de sim/não	8
ANOVA com um fator	Compara mais do que duas médias populacionais usando um fator	$y$ é quantitativa; o fator é $x$	10
Teste de Tukey	Comparações Múltiplas	Intervalos de confiança para todos os pares de médias; mantém as taxas de erro baixas	10
LSD de Fisher	Comparações Múltiplas	Intervalos de confiança para todos os pares de médias; taxas de erro globais mais altas do que as de Tukey	10
Método Scheffe	Comparações Múltiplas	Examina as combinações lineares das médias, não apenas dos pares	10
Ajuste de Bonferroni	Comparações Múltiplas	Todos os pares de testes $t$ ajustados para número de teste	10
Teste de Dunnett	Comparações Múltiplas	Experimentos; compara tratamento versus controle	10
Teste de Student Newman-Keuls (SNK)	Comparações Múltiplas	Abordagem gradual, compara pares ordenados do menor para o maior	10
Teste de Duncan (MRT)	Comparações Múltiplas	Ajusta SNK para mais força	10
ANOVA com dois fatores	Compara mais do que duas médias populacionais usando dois fatores mais interação	$y$ é quantitativa; os fatores são $(x_1, x_2)$	11
Testes do Qui-quadrado	Testa a independência de duas variáveis ou a qualidade de ajuste para uma variável qualitativa	Todas as variáveis são qualitativas	14, 15
Teste do sinal/ Teste dos postos sinalizados	Testa uma mediana populacional	$y$ é quantitativa ou ordinal (baseada nos postos)	17
Teste da soma de postos	Compara duas medianas populacionais	$y$ é quantitativa ou ordinal (baseada nos postos)	18
Teste de Kruskal-Wallis	Compara mais de duas medianas populacionais usando um fator	$y$ é quantitativa ou ordinal (baseada nos postos); o fator é $x$	19

***Para Leigos: A série de livros para iniciantes que mais vende no mundo.***



## Entendendo os resultados fornecidos pelo computador

Esta página mostra o dissecamento dos resultados fornecidos pelo programa estatístico para a regressão múltipla e para a ANOVA. Os professores adoram dar esses dados nas provas e pedir que você os interprete. Às vezes, eles deixam espaços em branco e pedem que você os preencha usando a informação dada — esteja preparado! (Observação: Para mais informação sobre como incorporei os resultados fornecidos pelo programa aos tópicos deste livro, veja a Introdução e o Capítulo 1).

### Análise de regressão Y versus $X_1, X_2$

The regression equation is

$Y = 2.34 + 0.00741X_1 + 0.0261X_2$  [row 1]

Predictor	Coef	SE Coef	T	P
Constant	2.3405	0.6821	3.43	0.002
$X_1$	0.007406	0.003435	2.16	0.040 [row 2]
$X_2$	0.02610	0.01176	2.22	0.035 [row 3]

S = 2.44958 [4] R-Sq = 39.4% [5] R-Sq(adj) = 34.9% [6]

[linha 1] = Este é o modelo para estimar  $y$  usando  $x_1$  e  $x_2$  (equação do plano).

[linha 2] = o coeficiente de  $x_1$  é 0,007; a estatística- $t$  para testar sua significância (dado que  $x_2$  está no modelo) é 2,16, valor que é significativo (valor- $p$  = 0,04, o que é menor do que 0,05).

[linha 3] = o coeficiente de  $x_2$  é 0,026; a estatística- $t$  para testar sua significância (dado que  $x_1$  está no modelo) é 2,22, valor que é significativo (valor- $p$  = 0,035, o que é menor do que 0,05).

[4] = Variabilidade de  $y$  sobre os valores previstos (um valor pequeno é desejável).

[5] =  $R^2$  = Porcentagem de variabilidade em  $y$  explicada por  $x_1$  e  $x_2$  (uma alta porcentagem é um bom sinal).

[6] =  $R^2$  (de [5]) ajustado para o número de variáveis no modelo. Este é o chamado " $R^2$  Ajustado". (Um valor alto é bom.)

### ANOVA com um fator: Y versus Group

Source	DF	SS	MS	F	P
Group	2	20.58	10.29	1.13	0.329 [row 1]
Error	63	572.45	9.09		[row 2]
Total	65	593.03			[row 3]

S = 3.014 R-Sq = 3.47% R-Sq(adj) = 0.41% [row 4]

[linha 1] = tratamento (trt) = grupo;  $k = 3$  grupos, pois  $gl = k - 1 = 2$ ; SQT = 20,58; MQT = SQT /  $gl = 20,58/2 = 10,29$ .  $F = MQT / MQE = 1,13$  não significativo (valor- $p$  = 0,329 > 0,05). (Veja a linha 2 para MQR.) Assim, não há diferença entre os os grupos com relação à variável  $y$ .

[linha 2] =  $gl = n - k = 63$ , assim,  $n = 66$  (pois  $k = 3$ , segundo a linha 1). MQE = SQD /  $gl = 572,45 / 63 = 9,09$ . MQE é o denominador do teste- $F$  na linha 1.

[linha 3] =  $gl$  Total =  $n - 1$ , assim,  $n = 66$ . Lembre-se SQTOT = SQT + SQE.

[linha 4] = Veja [4], [5], e [6] da saída para Regressão. Você pode ver que distinguir os grupos não influencia  $y$ , pois  $R^2$  é muito pequeno e  $R^2$  ajustado (para o número de grupos) é ainda menor.

**Para Leigos: A série de livros para iniciantes que mais vende no mundo.**





# ***Estatística II*** PARA **LEIGOS®**

**por Deborah Rumsey, PhD**

  
ALTA BOOKS  
EDITORA  
Rio de Janeiro, 2014

**Estatística II Para Leigos** Copyright © 2014 da Starlin Alta Editora e Consultoria Eireli.

ISBN: 978-85-7608-636-9

*Translated from original Statistics II For Dummies© 2009 by John Wiley & Sons, Inc. ISBN 978-0-470-46646-9. is translation is published and sold by permission John Wiley & Sons, Inc, the owner of all rights to publish and sell the same. PORTUGUESE language edition published by Starlin Alta Editora e Consultoria Eireli, Copyright © 2014 by Starlin Alta Editora e Consultoria Eireli.*

Todos os direitos reservados e protegidos por Lei. Nenhuma parte deste livro, sem autorização prévia por escrito da editora, poderá ser reproduzida ou transmitida.

**Erratas:** No site da editora relatamos, com a devida correção, qualquer erro encontrado em nossos livros (procure pelo título do livro).

**Marcas Registradas:** Todos os termos mencionados e reconhecidos como Marca Registrada e/ou Comercial são de responsabilidade de seus proprietários. A Editora informa não estar associada a nenhum produto e/ou fornecedor apresentado no livro.

Impresso no Brasil — 1ª Edição, 2014

Vedada, nos termos da lei, a reprodução total ou parcial deste livro.

---

## **Produção Editorial**

Editora Alta Books

## **Gerência Editorial**

Anderson Vieira

## **Editoria Para Leigos**

Evellyn Pacheco

## **Supervisão Gráfica**

Angel Cabeza

## **Supervisão de Qualidade Editorial**

Sergio Luiz de Souza

## **Supervisão de Texto**

Jaciara Lima

## **Conselho de Qualidade Editorial**

Anderson Vieira

Angel Cabeza

Jaciara Lima

Rodrigo Araujo

Sergio Luiz de Souza

## **Design Editorial**

Auleriano Messias

Aurélio Corrêa

## **Marketing e Promoção**

marketing@altabooks.com.br

## **Equipe Editorial**

Claudia Braga

Cristiane Santos  
Daniel Siqueira  
Elaine Mendonça  
Hannah Carriello  
Letícia Vitoria  
Livia Brazil  
Marcelo Vieira  
Milena Lepsch  
Milena Souza  
Natália Gonçalves  
Thiê Alves

### **Tradução**

Larissa Franzin

### **Copidesque**

Joris Bianca da Silva

### **Revisão Técnica**

Christina Ghiaroni Tiziano

*Bacharel em Estatística – UFRJ* Manuel Martins Filho

*Doutor pelo Programa de Engenharia e Computação — COPPE/UFRJ,*

*Linha de Pesquisa:*

*Inteligência Artificial*

### **Revisão Gramatical**

Jorge Guimarães

Tais Nunes Garcia

### **Diagramação**

Lúcia Quaresma

### **Produção de ePub**

Tatiana Medeiros

---

#### Dados Internacionais de Catalogação na Publicação (CIP)

R938e Rumsey, Deborah.

Estatística II para leigos / por Deborah Rumsey. – Rio de Janeiro, RJ : Alta Books, 2014.

408 p. : il. ; 24 cm. – (Para leigos)

Inclui índice e apêndice.

Tradução de: *Selling for Dummies* (3 ed.).

ISBN 978-85-7608-636-9

1. Estatística - 2. Análise de regressão. 3. Análise de variância.  
4. Teste qui-quadrado. 5. Estatística não-paramétrica. I Título. II  
Série.

CDU 519.22

CDD 519.2

**Índice para catálogo sistemático:**

(Bibliotecária responsável: Sabrina Leal Araujo – CRB 10/1507)



Rua Viúva Cláudio, 291 – Bairro Industrial do Jacaré  
CEP: 20970-031 – Rio de Janeiro – Tels.: (21) 3278-8069/8419  
[www.altabooks.com.br](http://www.altabooks.com.br) – e-mail: [altabooks@altabooks.com.br](mailto:altabooks@altabooks.com.br)  
[www.facebook.com/altabooks](http://www.facebook.com/altabooks) – [www.twitter.com/alta\\_books](http://www.twitter.com/alta_books)

# ***Dedicatória***

Para meu marido, Eric: meu sol nasce e se põe por você. Para meu filho, Clint: amo você daqui até a lua, ida e volta.

## ***Sobre a Autora***

**Deborah Rumsey** é PhD em Estatística pela Ohio State University (1993), onde é especialista no ensino de Estatística e auxiliar docente do Departamento de Estatística. Dra. Rumsey recebeu o privilégio de ser nomeada membro da Associação Americana de Estatística. Também ganhou o Prêmio Presidencial de Ensino da Kansas State University. Ela ainda é a autora de *Estatística Para Leigos*, *Statistics Workbook For Dummies* e *Probability For Dummies*, além de ter publicado inúmeros artigos e apresentado várias palestras sobre o ensino da Estatística. Suas paixões incluem estar com a família, observar pássaros, ficando mais tempo em seu trator Kubota, e torcer para o Ohio State Buckeyes em mais um campeonato nacional.

## *Agradecimentos da Autora*

Obrigada, novamente, Lindsay Lefevere e Kathy Cox, por me darem a oportunidade de escrever este livro; Natalie Harris e Guthrie Chrissy, pelo apoio inabalável e pela perfeita moldagem de minhas palavras e ideias; Kim Gilbert, da Universidade da Geórgia, por sua minuciosa revisão técnica; Elizabeth Rea e Sarah Westfall, pelo maravilhoso copidesque. Um agradecimento especial a Elizabeth Stasny, por sua orientação e apoio desde o primeiro dia, e a Joan Garfield, pela constante inspiração e encorajamento.

# Sumário Resumido

---

## ***Introdução***

## ***Parte I: Encarando os Fundamentos da Análise de Dados e da Construção de Modelos***

Capítulo 1: Além das Operações Numéricas: A Arte e a Ciência da Análise de Dados

Capítulo 2: Encontre a Análise Certa para o Problema

Capítulo 3: Revendo Intervalos de Confiança e Testes de Hipótese

## ***Parte II: Usando Diferentes Tipos de Regressão para Fazer Previsões***

Capítulo 4: Em Linha com a Regressão Linear Simples

Capítulo 5: Regressão Múltipla com Duas Variáveis X

Capítulo 6: Como Vou Sentir Sua Falta se Você Não Sair? Escolha do Modelo de Regressão

Capítulo 7: Subindo na Curva de Aprendizagem com a Regressão Não Linear

Capítulo 8: Sim, Não, Talvez: Fazendo Previsões Usando a Regressão Logística

## ***Parte III: Analisando a Variância com ANOVA***

Capítulo 9: Precisando Testar Várias Médias? Venha para a ANOVA!

Capítulo 10: Organizando as Médias Através das Comparações Múltiplas

Capítulo 11: Percorrendo os Caminhos da ANOVA com Dois Fatores

Capítulo 12: Regressão e ANOVA: Uma Relação Inesperada!

## ***Parte IV: Construindo Fortes Ligações com os Testes Qui-quadrado***

Capítulo 13: Fazendo Associações com Tabelas de Dupla Entrada

Capítulo 14: Independente o Suficiente para o Teste do Qui-quadrado

Capítulo 15: Usando os Testes do Qui-quadrado para Qualidade de Ajuste (dos Dados, e Não de Seu Jeans)

## ***Parte V: Estatística Não Paramétrica: Rebeldes sem Distribuição***

Capítulo 16: Ficando Não Paramétrico

Capítulo 17: Todos os Sinais Apontam para o Teste dos Sinais e o Teste de Postos Sinalizados

Capítulo 18: Subindo de Posto com o Teste das Somas dos Postos

Capítulo 19: Faça o Kruskal-Wallis e Ordene as Somas com Wilcoxon

Capítulo 20: Apontando Correlações com o Posto de Spearman

## ***Parte VI: A Parte dos Dez***

Capítulo 21: Os Dez Erros Mais Comuns nas Conclusões Estatísticas

Capítulo 22: Dez Formas de Chegar na Frente por Saber Estatística



*Apêndice: Tabelas de Referência*

# Sumário

---

## *Introdução*

Sobre Este Livro

Convenções Usadas Neste Livro

Só de Passagem

Penso que

Como Este Livro Está Organizado

Parte I: Encarando os Fundamentos da Análise de Dados e da Construção de Modelos

Parte II: Usando Diferentes Tipos de Regressão para Fazer Previsões

Parte III: Analisando a Variância com ANOVA

Parte IV: Construindo Fortes Ligações com os Testes Qui-quadrado

Parte V: Estatística Não Paramétrica: Rebeldes sem Distribuição

Parte VI: A Parte dos Dez

Ícones Usados Neste Livro

De Lá para Cá, Daqui para Lá

## *Parte I: Encarando os Fundamentos da Análise de Dados e da Construção de Modelos*

### **Capítulo 1: Além das Operações Numéricas: A Arte e a Ciência da Análise de Dados**

Análise de Dados: Olhe Antes de Mastigar

Nada(nem mesmo uma reta) dura para sempre

Bisbilhotar os dados não é coisa que se faça!

Proibido pescar(dados)

Veja o Quadro como um Todo: Um Panorama sobre Estatística II

Parâmetro da população

Estatística amostral

Intervalo de confiança

Teste de hipótese

Análise de variância (ANOVA)

Comparações múltiplas

Efeitos de interação

Correlação

Regressão linear

Testes Qui-quadrados

Estatística não paramétrica

### **Capítulo 2: Encontre a Análise Certa para o Problema**

Variáveis Categóricas versus Variáveis Quantitativas

Estatísticas para Variáveis Categóricas

Estimando uma proporção

Comparando proporções

Procurando relações entre variáveis categóricas

Construindo modelos para fazer previsões

Estatísticas para Variáveis Quantitativas

Fazendo estimativas

Fazendo comparações

Explorando relações

Prevendo y através de x

Evitando o Viés

Medindo a Precisão Através da Margem de Erro

Conhecendo Seus Limites

## **Capítulo 3: Revendo Intervalos de Confiança e Testes de Hipótese**

Estimando Parâmetros Usando os Intervalos de Confiança

Entendendo o básico: A forma geral de um intervalo de confiança

Encontrando o intervalo de confiança para uma média populacional

O que altera a margem de erro?

Interpretando um intervalo de confiança

O que É que os Testes de Hipótese Têm?

O que  $H_0$  e  $H_a$  realmente representam?

Reunindo evidências em uma estatística de teste

Determinando a força da evidência através do valor- $p$

Alarmes falsos e oportunidades perdidas: Erros Tipo I e Tipo II

O poder de um teste de hipótese

## ***Parte II: Usando Diferentes Tipos de Regressão para Fazer Previsões***

### **Capítulo 4: Em Linha com a Regressão Linear Simples**

Investigando Relações com Diagramas de Dispersão e Correlações

Usando diagramas de dispersão para investigar relações

Comparando informações através do coeficiente de correlação

Construindo um Modelo de Regressão Linear Simples

Encontrando a reta certa para modelar seus dados

O intercepto y da reta de regressão

O coeficiente angular da reta de regressão

Estimando pontos através da regressão linear

Sem Deixar Nenhuma Conclusão para Trás: Testes e Intervalos de Confiança para a Regressão

Analisando o coeficiente angular

Inspecionando o intercepto y

Construindo intervalos de confiança para a resposta média

- Previendo o futuro com os intervalos de previsão
- Checando a Adequação do Modelo (dos Dados, Não das Roupas!)
- Definindo as condições
- Encontrando e investigando os resíduos
- Usando  $r^2$  para medir o ajuste do modelo
- Analizando outliers
- Conhecendo as Limitações de Sua Análise de Regressão
- Evitando cair no modo causa e efeito
- Extrapolação: N-A-O-Til, NUNCA!
- Às vezes é preciso ter mais do que uma variável

## **Capítulo 5: Regressão Múltipla com Duas Variáveis X**

- Conhecendo o Modelo de Regressão Múltipla
  - Descobrendo os usos da regressão múltipla
  - A fórmula geral do modelo de regressão múltipla
  - Seguindo os passos rumo a uma análise
- Observando x's e y's
- Coletando Dados
- Identificando Possíveis Relações
  - Construindo diagramas de dispersão
  - Correlações: Examinando os vínculos
- Checando a Multicolinearidade
- Encontrando o Modelo sob Medida para Duas Variáveis X
  - Obtendo os coeficientes de regressão múltipla
  - Interpretando os coeficientes
  - Testando os coeficientes
- Previendo y Através das Variáveis x
- Verificando o Ajuste do Modelo de Regressão Múltipla
  - Observando as condições
  - Traçando um plano para checar as condições
  - Verificando as três condições

## **Capítulo 6: Como Vou Sentir Sua Falta se Você Não Sair? Escolha do Modelo de Regressão**

- Dando o Pontapé Inicial na Estimativa para a Distância de um Punt
  - Fazendo o brainstorm das variáveis e coletando os dados
  - Examinando diagramas de dispersão e correlações
- Igual a Comprar Sapatos: O Modelo É Lindo, Mas Serve?
  - Avaliando o ajuste do modelo de regressão múltipla
  - Processo de seleção de modelo

## **Capítulo 7: Subindo na Curva de Aprendizagem com a Regressão Não Linear**

Antecipando a Regressão Não Linear  
Começando com Diagramas de Dispersão  
Nas Curvas da Estrada com os Polinômios  
Relembrando o que é um polinômio  
Em busca do melhor modelo polinomial  
Usando um polinômio de segundo grau para passar na prova  
Avaliando o ajuste de um modelo polinomial  
Fazendo previsões  
Subiu? Desceu? Então É Exponencial!  
Recordando os modelos exponenciais  
Em busca do melhor modelo exponencial  
Espalhando segredos de forma exponencial

## **Capítulo 8: Sim, Não, Talvez: Fazendo Previsões Usando a Regressão Logística**

Entendendo o Modelo de Regressão Logística  
Qual é a diferença entre a regressão logística e as outras regressões?  
Utilizando uma curva em S para estimar as probabilidades  
Interpretando os coeficientes do modelo de regressão logística  
O modelo de regressão linear em ação  
Fazendo uma Análise de Regressão Logística  
Fazendo a análise no Minitab  
Encontrando os coeficientes e construindo o modelo  
Estimando p  
Verificando o ajuste do modelo  
Ajustando o modelo

## ***Parte III: Analisando a Variância com ANOVA***

## **Capítulo 9: Precisando Testar Várias Médias? Venha para a ANOVA!**

Comparando Duas Médias com um Teste-t  
Avaliando Mais Médias com ANOVA  
Cuspe de sementes: Uma situação perfeita para a ANOVA  
Seguindo os passos da ANOVA  
Verificando as Condições  
Verificando a independência  
Procurando o que é normal  
Notando a dispersão  
Estabelecendo as Hipóteses  
Realizando o Teste- $F$   
ANOVA no Minitab  
Desmembrando a variância em somas de quadrados  
Localizando as médias das somas de quadrados

Chegando à estatística-*F*  
Tirando conclusões a partir da ANOVA  
O que fazer agora?  
Verificando o Ajuste do Modelo ANOVA

## **Capítulo 10: Organizando as Médias Através das Comparações Múltiplas**

Acompanhando a ANOVA  
Comparando o uso de minutos no celular: Um exemplo  
Preparando o terreno para os procedimentos de comparação múltipla  
Identificando as Médias Diferentes com Fisher e Tukey  
Pescando diferenças com o LSD de Fisher  
Usando o novo e aperfeiçoado LSD de Fisher  
O teste de Tukey  
Examinando a Saída para Determinar a Análise  
Tantos Outros Procedimentos, Tão Pouco Tempo!  
Cortando a conversa fiada com o ajuste de Bonferroni  
Comparando combinações usando o método de Scheffe  
O teste de Dunnett  
O teste de Student Newman-Keuls  
O teste de Duncan  
Ficando não paramétrico com o teste de Kruskal-Wallis

## **Capítulo 11: Percorrendo os Caminhos da ANOVA com Dois Fatores**

Configurando o Modelo ANOVA com Dois Fatores  
Determinando os tratamentos  
Em busca das somas de quadrados  
Entendendo os Efeitos da Interação  
Mas, afinal, o que é interação?  
Interagindo com os gráficos de interação  
Testando os Termos na ANOVA com Dois Fatores  
Executando uma Tabela ANOVA  
Interpretando os resultados: Números e gráficos  
O Branco Fica Mais Branco na Água Quente? Mais um Caso para a ANOVA com Dois Fatores

## **Capítulo 12: Regressão e ANOVA: Uma Relação Inesperada!**

Vendo a Regressão Através dos Olhos da Variação  
Localizando a variabilidade e encontrando uma “x-plicação”  
Chegando aos resultados com a regressão  
Avaliando o ajuste do modelo de regressão  
Regressão e ANOVA: O Encontro dos Modelos  
Comparando as somas de quadrados  
Dividindo os graus de liberdade

Levando a regressão até a tabela ANOVA  
Relacionando as estatísticas  $F$  e  $t$ : a última fronteira

## ***Parte IV: Construindo Fortes Ligações com os Testes Qui-quadrado***

### **Capítulo 13: Fazendo Associações com Tabelas de Dupla Entrada**

- Decompondo uma Tabela de Dupla Entrada
  - Organizando dados em uma tabela de dupla entrada
  - Preenchendo as células
  - Totais marginais
- Desmembrando as Probabilidades
  - Probabilidades marginais
  - Probabilidades conjuntas
  - Probabilidades condicionais
- Tentando Ser Independente
  - Verificando a independência entre duas categorias
  - Verificando a independência entre duas variáveis
- Desmistificando o Paradoxo de Simpson
  - Experimentando o Paradoxo de Simpson
  - Descobrimo o porquê do Paradoxo de Simpson
  - De olho no Paradoxo de Simpson

### **Capítulo 14: Independente o Suficiente para o Teste do Qui-quadrado**

- O Teste do Qui-quadrado para a Independência
  - Coletando e organizando os dados
  - Determinando as hipóteses
  - Calculando as frequências esperadas
  - Verificando as condições para o teste
  - Calculando a estatística Qui-quadrado
  - Encontrando seus resultados na tabela do Qui-quadrado
  - Tirando conclusões
  - Colocando o Qui-quadrado à prova
- Comparando Dois Testes para Comparar Duas Proporções
  - Refamiliarizando-se com o teste-Z para duas proporções populacionais
  - Igualando os testes do Qui-quadrado e testes-Z para uma tabela dois por dois

### **Capítulo 15: Usando os Testes do Qui-quadrado para Qualidade de Ajuste (dos Dados, e Não de Seu Jeans)**

- Encontrando a Estatística de Qualidade de Ajuste
  - O observado versus o esperado
  - Calculando a estatística de qualidade de ajuste
- Interpretando a Estatística da Qualidade de Ajuste Através do Qui-quadrado
  - Verificando as condições antes de começar

## ***Parte V: Estatística Não Paramétrica: Rebeldes sem Distribuição***

### **Capítulo 16: Ficando Não Paramétrico**

Em Favor da Estatística Não Paramétrica

Não precisa se preocupar se as condições não forem atendidas

Uma chance para a mediana mostrar seu potencial

Então, qual é a pegadinha?

Dominando o Básico das Estatísticas Não Paramétricas

Sinal

Postos

Postos com sinais

Soma de postos

### **Capítulo 17: Todos os Sinais Apontam para o Teste dos Sinais e o Teste de Postos Sinalizados**

Interpretando os Sinais: O Teste dos Sinais

Testando a mediana

Estimando a mediana

Testando os pares combinados

Um Passo Adiante com o Teste de Postos Sinalizados

Uma limitação do teste dos sinais

Seguindo os passos para realizar um teste de postos sinalizados

Emagrecendo com os postos sinalizados

### **Capítulo 18: Subindo de Posto com o Teste das Somas dos Postos**

Realizando o Teste da Soma dos Postos

Verificando as condições

Seguindo os passos para a realização de um teste

Aumentando o tamanho da amostra

Realizando um Teste da Soma dos Postos: Qual Corretor de Imóveis Vende Casas Mais Rápido?

Verificando as condições para este teste

Testando a hipótese

### **Capítulo 19: Faça o Kruskal-Wallis e Ordene as Somas com Wilcoxon**

Fazendo o Teste de Kruskal-Wallis para Comparar Mais de Duas Populações

Verificando as condições

Estabelecendo o teste

Realizando o teste passo a passo

Localizando as Diferenças: O Teste da Soma dos Postos de Wilcoxon

Comparações pareadas



Realizando testes de comparação para ver quem é diferente  
Examinando as medianas para ver como elas se diferem

## **Capítulo 20: Apontando Correlações com o Posto de Spearman**

Pearson e Suas Preciosas Condições

Correlação de Posto de Spearman

Calculando a correlação de posto de Spearman

Spearman em ação: Relacionando aptidão ao desempenho

## ***Parte VI: A Parte dos Dez***

## **Capítulo 21: Os Dez Erros Mais Comuns nas Conclusões Estatísticas**

Dizer o que as Estatísticas Provam

Tecnicamente Não É Estatisticamente Significativo, Mas

Concluir que  $x$  Causa  $y$

Supor que os Dados São Normais

Relatar Apenas os Resultados “Importantes”

Supor que uma Amostra Grande É Sempre Melhor

Não É Tecnicamente Aleatória, Mas

Supor que 1.000 Respostas São 1.000 Respostas

Naturalmente, os Resultados se Aplicam à População em Geral

Omitir

## **Capítulo 22: Dez Formas de Chegar na Frente por Saber Estatística**

Faça as Perguntas Certas

Seja Cético

Colete e Analise os Dados Corretamente

Pedindo Ajuda

Refazendo os Passos de Outras Pessoas

Juntando as Peças

Verificando Suas Respostas

Explicando a Saída

Fazendo Recomendações Convincentes

Estabelecendo-se como o Cara da Estatística

## **Capítulo 23: Dez Empregos Legais que Usam Estatística**

Pesquisadores de Opinião Pública

Ornitólogo (Observador de Pássaros)

Comentarista ou Jornalista Esportivo

Jornalista

Combatentes do Crime

Profissional da Área Médica

Executivo de Marketing

Advogado  
Corretor de Ações

## ***Apêndice: Tabelas de Referência***

Tabela- $t$

Tabela Binomial

Tabela do Qui-quadrado

Tabela da Soma de Postos

Tabela- $F$

# Introdução

---

**E**ntão, você já sabe alguma coisa sobre Estatística. Médias, medianas e desvios padrão são todos termos que lhe soam familiares. Também conhece um pouco de pesquisa e experimentos e entende as ideias básicas de correlação e regressão simples. Estudou probabilidade, margem de erro e alguma coisa sobre testes de hipótese e intervalos de confiança. Está pronto para acrescentar ferramentas mais sofisticadas à sua caixa de ferramenta estatísticas? *Estatística II Para Leigos* inicia justamente onde o *Estatística Para Leigos* parou e faz com que você continue passo a passo sua caminhada pela trilha de ideias e técnicas estatísticas.

O foco do *Estatística II Para Leigos* está em encontrar mais formas de analisar dados. Aqui, você encontrará instruções passo a passo para usar técnicas como a de regressão múltipla, regressão não linear, análise de variância com um ou dois fatores (ANOVA), testes Qui-quadrado e estatísticas não paramétricas. Com essas novas técnicas, é possível estimar, investigar, correlacionar e congrega ainda mais variáveis baseadas nas informações que você tem à mão.

# Sobre Este Livro

Este livro foi concebido para aqueles que concluíram os conceitos básicos de estatística, indo desde os intervalos de confiança até os testes de hipótese (encontrados no *Estatística Para Leigos*), e que estão prontos para desbravar o caminho e seguir em direção à parte final da Estatística I ou encarar a Estatística II. Contudo, sempre que necessário, voltarei brevemente a alguns pontos de Estatística I, apenas para lembrá-lo da matéria e garantir que você está tinindo. A cada nova técnica, você terá um panorama geral de quando e por que ela é usada, como saber quando ela é necessária, além de instruções passo a passo sobre como fazê-la, e as dicas e truques de uma analista de dados veterana (eu mesma, ao seu dispor!). Uma vez que é muito importante ser capaz de saber qual método usar e quando, enfatizo o que distingue cada técnica e o que dizem os resultados. Você também verá muitas aplicações das técnicas em situações reais.

Também incluo a interpretação do resultado para fim de análise de dados. Mostro como usar os programas para chegar aos resultados, mas foco mais em como interpretar os resultados encontrados no resultado (saída), uma vez que é mais provável que você tenha que interpretar esse tipo de informação em vez de fazer a programação. E já que as equações e cálculos podem ficar muito complexos se feitos à mão, você usará muito o computador para chegar aos resultados. Incluo instruções para o uso do Minitab na realização de muitos dos cálculos apresentados neste livro. A maioria dos professores que ministram aulas sobre esses tópicos também mantém a mesma opinião. (Que alívio!)

Este livro se diferencia dos outros livros de Estatística II por muitos motivos, pois traz:

- ✓ **Explicações completas dos conceitos de Estatística II.** Muitos livros enfiam todos os tópicos de Estatísticas II no finalzinho da matéria de Estatística I; o resultado disso é que esses pontos tendem a ficar condensados e aparecerem como se fossem opcionais. Mas não se preocupe, dediquei tempo para explicar clara e completamente todas as informações de que você precisa para sobreviver e prosperar.
- ✓ **Dissecação do resultado.** Ao longo de todo o livro, apresento muitos exemplos que utilizam programa de estatística para analisar dados. Em cada caso, apresento o resultado e explico como eu o obtive e o que ele significa.
- ✓ **Um grande número de exemplos.** Incluo muitos exemplos para abordar os diferentes tipos de problemas que você irá encarar.
- ✓ **Muitas dicas, estratégias e alertas.** Compartilho com você alguns segredos fundamentados em minha experiência como professora, orientadora e na correção de provas.
- ✓ **Linguagem simples.** Tento manter a linguagem informal a fim de ajudá-lo a entender, memorizar e praticar as definições, as técnicas e os processos estatísticos.

- ✓ **Passo a passo conciso e objetivo.** Na maioria dos capítulos, é possível encontrar instruções passo a passo que, de forma intuitiva, explicam como trabalhar os problemas de Estatística II — e o ajudam a lembrar como resolvê-los quando tiver que fazê-lo sozinho.

# *Convenções Usadas Neste Livro*

Ao longo de todo este livro, usei várias convenções, as quais gostaria que você conhecesse:

- ✓ Indico a multiplicação, usando um sinal de vezes indicado por um asterisco (\*).
- ✓ Indico as hipóteses nula e alternativa como  $H_0$  (para a hipótese nula) e  $H_a$  (para a hipótese alternativa).
- ✓ O pacote de software estatístico que utilizo e mostro ao longo do livro é o Minitab 14, mas me refiro a ele simplesmente por Minitab.
- ✓ Sempre que introduzo um novo termo, o escrevo em itálico.
- ✓ Palavras-chave e passo a passos numerados aparecem em **negrito**.
- ✓ Sites e endereços de e-mail aparecem em courier new.

# *Só de Passagem*

Às vezes, apresento alguns detalhes mais técnicos de fórmulas e procedimentos para os leitores que possam precisar deles — ou que apenas queiram conhecer os mínimos detalhes. Essas minúcias estão marcadas com um ícone. Também incluo barras laterais como informações à parte do texto essencial, em geral na forma de um exemplo real ou de informação extra que você possa achar interessante. Fique à vontade para pular esses ícones e essas barras laterais, pois não perderá nenhuma das informações principais de que precisa (mas, se você for lê-las, poderá deixar seu professor de estatística impressionado com seu conhecimento além da média!).

## *Penso que...*

Já que este livro aborda a Estatística II, presumo que você já tenha feito um curso de introdução à Estatística (ou, pelo menos, tenha lido *Estatística Para Leigos*), que tenha apresentado a você o Teorema do Limite Central e, talvez, algo sobre intervalos de confiança e testes de hipótese (embora eu faça uma breve revisão desses conceitos no Capítulo 3). Não é preciso ter experiência com regressão linear simples. Apenas a álgebra aprendida no ensino superior é necessária para os detalhes matemáticos. Experiência com softwares estatísticos é algo a mais, mas não é necessária.

Como estudante, você pode abordar estes tópicos destas formas: como prosseguimento do curso de Estatística I (talvez de forma apressada, mas, de qualquer modo, está vendo); ou como um curso de duas etapas, sendo os tópicos deste livro o foco da segunda fase. Se for o caso, este livro lhe oferece a informação necessária para que você se dê bem, seja qual for o método que esteja seguindo.

Você pode estar apenas interessado em Estatística II para compreender situações do dia a dia ou, talvez, queira aprimorar sua compreensão de estudos e resultados estatísticos mostrados na mídia. Caso seja este o seu caso, você encontrará vários exemplos e aplicações dessas técnicas estatísticas na vida real, assim como cuidados para interpretá-los.



# ***Como Este Livro Está Organizado***

Este livro está organizado em cinco partes principais que exploram os tópicos mais importantes em Estatística II, além de uma parte bônus que oferece uma série de dez referências rápidas para você usar. Cada parte contém capítulos que dividem o principal objetivo da parte em fragmentos compreensíveis. A configuração não linear deste livro lhe permite pular capítulos e, ainda assim, acessar e compreender facilmente qualquer tópico dado.

## ***Parte I: Encarando os Fundamentos da Análise de Dados e da Construção de Modelos***

Esta parte aborda as grandes ideias das estatísticas descritiva e inferencial, além da regressão linear simples no contexto da construção de modelos e do processo decisório. Alguns tópicos da Estatística I recebem uma rápida revisão. Também apresento o jargão típico da Estatística II.

## ***Parte II: Usando Diferentes Tipos de Regressão para Fazer Previsões***

Nesta parte, você pode revisar e expandir as ideias da regressão linear simples para o processo de utilização de mais de uma variável preditora. Essa parte apresenta técnicas para manipular dados que seguem uma curva (modelos não lineares) e modelos para dados do tipo sim ou não usados para fazer previsões sobre o acontecimento ou não de um evento (regressão logística). Nesta parte, você encontra tudo aquilo de que precisa para saber sobre condições, diagnósticos, construção de modelos, técnicas de análise de dados e interpretação de resultados.

## ***Parte III: Analisando a Variância com ANOVA***

Você pode querer comparar as médias de mais de duas populações, e isso requer a utilização da análise da variância (ANOVA). Essa parte discute as condições básicas necessárias, o teste-F, a ANOVA de um e dois fatores e as comparações múltiplas. O objetivo final dessas análises é mostrar se as médias das populações em questão são diferentes e, caso sejam, quais são mais altas ou mais baixas do que o restante.

## ***Parte IV: Construindo Fortes Ligações com os Testes Qui-quadrado***

Esta parte abrange a distribuição Qui-quadrado e como você pode usá-la para modelar e testar dados categóricos (qualitativos). Você vai descobrir como testar a independência de

duas variáveis usando o teste Qui-quadrado. (Você não vai mais precisar fazer especulações apenas por meio da observação dos dados em tabelas 2X2!) Você também verá como usar um Qui-quadrado para testar a capacidade de adequação do modelo aos dados.

## ***Parte V: Estatística Não Paramétrica: Rebeldes sem Distribuição***

Esta parte o ajuda com as técnicas usadas em situações em que você não pode (ou não quer) partir do princípio de que seus dados vêm de uma população com determinada distribuição, como, por exemplo, quando sua população não é normal (condição exigida pela maioria dos outros métodos em Estatística II).

## ***Parte VI: A Parte dos Dez***

A leitura desta parte pode lhe dar uma vantagem dentro de uma importante área que vai além de fórmulas e técnicas da Estatística II: terminar o problema da forma correta (sabendo que tipos de conclusões você pode ou não tirar). Você também vai conhecer a Estatística II no mundo real, ou seja, de que forma ela pode ajudá-lo a se sobressair na multidão.

No final do livro, você também encontrará um apêndice que contém todas as tabelas necessárias para o entendimento e para a realização dos cálculos presentes nesta obra.

# Ícones Usados Neste Livro

Neste livro, utilizo ícones a fim de chamar sua atenção para determinados textos que aparecem frequentemente. Pense nos ícones como placas com as quais você se depara durante uma viagem. Algumas placas mostram atalhos, outras oferecem mais informações que você pode precisar; algumas o alertam sobre possíveis perigos, enquanto outras dão algum lembrete.



Este ícone significa que vou explicar como realizar uma determinada análise de dados usando o Minitab. Também explico as informações obtidas no resultado para que você consiga interpretá-lo.



Uso esse ícone para reforçar certas ideias que são cruciais para o sucesso em Estatística II, tais como coisas que acredito ser importante revisar durante a preparação para uma prova.



Quando você vir esse ícone, poderá pular a informação caso não queira saber os pormenores. Tais informações estão presentes para as pessoas que tenham um interesse especial ou a obrigação de saber mais sobre os aspectos mais técnicos de certas questões estatísticas.



Este ícone aponta dicas úteis, ideias ou atalhos que podem ser usados para economizar tempo; também inclui formas alternativas de compreender determinado conceito.



Utilizo os ícones de cuidado para ajudá-lo a ficar longe de erros e armadilhas comuns com os quais você pode se deparar ao lidar com ideias e técnicas relacionadas à Estatística II.

# *De Lá para Cá, Daqui para Lá*

Este livro foi escrito de forma não linear. Portanto, é possível começar por qualquer capítulo e, ainda assim, entender o que está acontecendo. Entretanto, quero fazer algumas recomendações caso você queira instruções de por onde começar.

Se você estiver bem familiarizado com as ideias de testes de hipótese e de regressão linear simples, comece pelo Capítulo 5 (regressão múltipla). Utilize o Capítulo 1 se precisar de uma referência para o jargão que os profissionais usam em Estatística II.

Se já tiver abordado todos os tópicos relacionados aos vários tipos de regressão (simples, múltipla, não linear e logística) ou um subconjunto dos tópicos que seu professor considerou importantes, vá ao Capítulo 9, o fundamento da análise de variância (ANOVA).

O Capítulo 14 é a parte por onde começar caso você queira enfrentar as variáveis categóricas (qualitativas) antes de acertar as quantitativas. Lá você pode trabalhar com o teste Qui-quadrado.

A estatística não paramétrica é apresentada no Capítulo 16. Essa área, hoje em dia, é um tópico significativo nos cursos de Estatística, no entanto, parece não receber o espaço devido nos livros didáticos. Comece aqui caso queira detalhes completos sobre os procedimentos não paramétricos mais comuns.

# **Parte I:**

## **Encarando os Fundamentos da Análise de Dados e da Construção de modelos**



**“Fiz uma avaliação do nosso último gráfico de pizza. Aparentemente, é de quatro queijos.”**



## *Nesta parte...*

**P**ara que você comece a deixar os conceitos básicos de estatística (abordados em seu livro de Estatística I, assim como no *Estatística Para Leigos*) para conhecer os novos e instigantes métodos apresentados neste livro, primeiro introduzo o básico da análise de dados, as terminologias mais importantes, os principais objetivos e conceitos da construção de modelos e as dicas para a escolha da estatística adequada ao trabalho. Além disso, vou refrescar sua memória com relação a itens de grande referência em Estatística I, e você também começará a fazer e observar alguns resultados básicos produzidos no Minitab.

# Capítulo 1

## Além das Operações Numéricas: A Arte e a Ciência da Análise de Dados

---

### *Neste Capítulo*

- ▶ Entendendo seu papel como analista de dados
  - ▶ Evitando gafes estatísticas
  - ▶ Bisbilhotando o jargão da Estatística II
- 

**J**á que está lendo este livro, você provavelmente já está familiarizado com o básico de Estatística e está pronto para fazer mais um avanço. O próximo passo envolve o uso do que você já conhece, mais a aprendizagem de algumas ferramentas e técnicas e, finalmente, a mistura de tudo irá ajudá-lo a resolver questões mais realistas através do uso de dados reais. Em termos estatísticos, você está pronto para entrar no mundo dos *analistas de dados*.

Neste capítulo, você vai rever os termos envolvidos em Estatística que fazem parte da análise de dados no nível da Estatística II. Você terá uma ideia do impacto que seus resultados podem causar ao ver o que essas técnicas de análise são capazes de fazer. Você também terá uma boa visão sobre o mal uso da análise de dados e seus efeitos.



# *Análise de Dados: Olhe Antes de Mastigar*

Apenas os estatísticos costumavam analisar dados, já que os únicos programas de computador disponíveis eram complicados demais e requeriam um grande conhecimento sobre Estatística para organizar e conduzir as análises. Os cálculos eram entediante e, às vezes, imprevisíveis, além de requererem um bom conhecimento sobre teorias e métodos para a obtenção de respostas corretas e confiáveis.

Hoje em dia, qualquer um que queira analisar dados pode fazê-lo sem grandes esforços. Muitos pacotes de programas estatísticos são feitos justamente com esse propósito — Microsoft Excel, Minitab, SAS e SPSS são apenas alguns exemplos. Também existem programas online gratuitos, como o Stat Crunch, para ajudá-lo a fazer exatamente o que seu nome sugere — esmiuçar os números e chegar a uma solução.

Cada programa tem seus prós e contras (e seus próprios usuários e discordantes). O meu preferido, ao qual vou me referir ao longo do livro, é o Minitab, pois é muito fácil de usar. Os resultados são precisos e o software vem carregado com todas as técnicas de análise de dados usadas em Estatística II. Embora a licença do Minitab não seja barata, a versão para estudantes pode ser alugada por um preço bem baixo.



A ideia mais importante durante a aplicação das técnicas estatísticas para a análise de dados é saber o que se passa por trás do cálculo, para que, assim, você (e não o computador) fique no controle da análise. É por isso que o conhecimento em Estatística II é crucial.



Muitas pessoas não se dão conta de que o software não lhes diz quando usar ou não usar uma determinada técnica estatística. Você é que tem de determinar isso. Como resultado, as pessoas acham que estão fazendo suas análises da forma correta, mas podem acabar cometendo todos os tipos de erros. Nas seções a seguir, dou exemplos de algumas situações em que inocentes análises de dados podem dar errado, e por que é importante identificar e evitar esses erros antes de começar os cálculos.

Conclusão: os pacotes de software atuais parecem mágica se você não entende os princípios de Estatística II envolvidos.

## **Voltando aos velhos tempos**

Antigamente, a fim de determinar se métodos diferentes geravam diferentes resultados, era preciso escrever um programa com um código que você só aprendia depois de ter uma aula. Era preciso digitar seus dados da forma requerida pelo programa, enviá-los para o computador e esperar os resultados. Esse método consumia muito tempo e dava muita dor de cabeça.

A boa nova é que os programas de estatística passaram por uma evolução inacreditável nos últimos 15 anos, a ponto de, hoje, você conseguir inserir seus dados de forma rápida e fácil em quase

todos os formatos. Além disso, as opções para a análise de dados são bem organizadas e listadas em barras de menu. “Pull-down menus” refere-se à característica das barras de menu de serem “estendidas” para baixo. Os resultados são instantâneos e eficazes, e você pode recortá-los e colá-los em um editor de texto em um piscar de olhos.

## ***Nada(nem mesmo uma reta) dura para sempre***

Bill Prediction é um estudante de estatística que estuda o efeito do tempo de estudo sobre o resultados obtidos em provas. Bill coleta dados sobre estudantes de estatística e utiliza seu confiabilíssimo software para prever as notas obtidas nas provas utilizando o tempo de estudo. O computador dele apresenta a equação  $y = 10x + 30$ , onde  $y$  representa a nota que você obtém se estudar um certo número de horas ( $x$ ). Observe que esse modelo é a equação de uma reta com intercepto  $y$  (coordenada  $y$  quando  $x = 0$ ) de 30 e coeficiente angular de 10.

Sendo assim, Bill prevê, usando este modelo, que se você não estudar nada, sua nota na prova será 30 (usando  $x = 0$  na equação para chegar ao valor de  $y$ , ponto que representa a intercepção  $y$  da reta). E ele prevê, usando este modelo, que se você estudar durante 5 horas, sua nota na prova será  $y = (10 * 5) + 30 = 80$ . Assim, o ponto (5,80) também está nesta reta.

Mas, então, Bill se empolga um pouco e quer saber o que aconteceria se você estudasse durante 40 horas (uma vez que, sempre que está estudando, ele tem a impressão de ter estudado todo esse tempo). O computador, então, lhe diz que se ele estudasse durante 40 horas, a previsão de sua nota seria  $(10 * 40) + 30 = 430$ . Essa, sim, seria uma nota alta! O problema é que a nota máxima em uma prova é 100. Dessa forma, Bill se questiona onde seu computador errou.

Mas acaba colocando a culpa no lugar errado. Na verdade, ele precisa se lembrar de que os valores de  $x$  precisam ter um limite para que esta equação faça sentido. Por exemplo, uma vez que  $x$  representa o tempo de estudo,  $x$  nunca pode ser menor do que zero. Se você substituir  $x$  por um número negativo, digamos  $x = -10$ , terá  $y = (10 * -10) + 30 = -70$ , resultado que não faz sentido. No entanto, nem a equação nem o computador que a descobriu sabem disso. O computador apenas faz o gráfico da reta que você dá, presumido que ela é infinita tanto na direção positiva quanto na negativa.

Depois de conseguir uma equação ou um modelo estatístico, é preciso especificar os valores que se aplicam à equação. As equações não sabem quando esses valores funcionam ou não; é tarefa do analista de dados determinar isso. A ideia é a mesma para a aplicação dos resultados de qualquer análise de dados que você realizar.

## ***Bisbilhotar os dados não é coisa que se faça!***

Os estatísticos inventaram um ditado que você já deve ter ouvido: “Os números não





mentem. Mas os mentirosos fabricam números.” Conheça todas as análises realizadas em um conjunto de dados, e não apenas as que foram relatadas como sendo estatisticamente significativas.

Suponha que Bill Prediction (da seção anterior) decida tentar prever as notas obtidas em uma prova de biologia, baseando-se no tempo de estudo, mas, desta vez, seu modelo não se adequa. Não querendo dar o braço a torcer, Bill insiste na ideia de que deva haver outros fatores que prevejam as notas da prova de biologia além do tempo de estudo e, então, sai em busca deles.

Bill vasculha todas as possibilidades. Seu conjunto de 20 possíveis variáveis inclui o tempo de estudo, a média de notas, a experiência anterior em estatística, as notas de matemática no Ensino Médio e se você masca chiclete durante a prova. Depois de muitas análises de correlação, as variáveis que Bill descobre estar relacionadas à nota da prova foram o tempo de estudo, as notas de matemática no Ensino Médio, a média de notas e se você masca chiclete durante a prova. No final, este modelo em particular serve muito bem (por critérios que discuto no Capítulo 5 sobre os modelos de regressão linear múltipla).

Mas aqui está o problema: ao observar todas as possíveis correlações entre suas 20 variáveis e a nota da prova, Bill está, na verdade, fazendo 20 análises estatísticas separadas. Sob as condições normais descritas no Capítulo 3, cada análise estatística tem 5% de chances de estar errada apenas por acaso. Aposto que você consegue acertar qual das correlações de Bill provavelmente está errada nesse caso. E ainda bem que ele não vai contar com um chiclete para melhorar sua nota em Biologia.



A observação exaustiva de dados em busca de algo é conhecida como *data snooping*. O data snooping concede ao pesquisador seus cinco minutos de fama, mas, depois, faz com que ele perca toda a sua credibilidade, pois ninguém mais quer repetir seus resultados.

## ***Proibido pescar(dados)***

Algumas pessoas simplesmente não aceitam um não como resposta, e, quando se trata da análise de dados, isso pode trazer problemas.

Sue Gonnafindit é uma pesquisadora determinada. Ela acredita que seu cavalo consegue contar batendo a pata no chão. (Por exemplo, ela diz dois, e o cavalo bate com a pata no chão duas vezes.) Sue coleta dados sobre seu cavalo durante quatro semanas, registrando a porcentagem de vezes que o cavalo conseguiu contar corretamente. Em seguida, conduz a análise estatística adequada a seus dados e se surpreende por não ter descoberto nenhuma diferença significativa entre os resultados de seu cavalo e os resultados que você teria obtido apenas por palpite.

Determinada a provar que seus resultados são reais, Sue procura outros tipos de análises existentes e coloca seus dados em toda e qualquer coisa que encontra (não importando o fato de que essas análises não sejam adequadas à sua situação). Usando o famoso método

da água mole em pedra dura tanto bate até que fura, em um determinado ponto, ela finalmente tropeça em um resultado significativo. Entretanto, o resultado é artificial, pois ela usou muitas análises inadequadas e ignorou os resultados da análise adequada, apenas porque estes não diziam o que ela queria ouvir.

Outro fato interessante: quando Sue foi a um programa de TV para mostrar ao mundo seu magnífico cavalo, alguém na plateia percebeu que sempre que o cavalo chegava ao número correto de batidas, Sue o interrompia dizendo: “Bom trabalho!”, e ele parava. Na verdade, ele não sabia contar; tudo o que ele sabia era parar de bater a pata quando ela dizia: “Bom trabalho!”

O ato de refazer análises de forma diferente a fim de tentar chegar aos resultados que você deseja é chamado de *data fishing* e, no mundo da estatística, isso é considerado uma infração gravíssima. (Entretanto, infelizmente, as pessoas cometem esse erro com frequência, a fim de verificar suas crenças mais fortes.) Ao usar a análise errada para conseguir os resultados desejados, você leva o público a pensar que sua hipótese é realmente correta quando, na verdade, ela pode não ser.



# *Veja o Quadro como um Todo: Um Panorama sobre Estatística II*

A Estatística II é uma extensão da Estatística I (estatística introdutória). Sendo assim, o jargão continua o mesmo, e as técnicas se baseiam no que você já conhece. Nesta seção, você encontra uma introdução à terminologia usada em Estatística II, além de uma ampla visão geral das técnicas utilizadas por estatísticos para analisar os dados e descobrir o que está por trás de toda história. (Caso você ainda tenha dúvidas sobre alguns termos da Estatística I, consulte seu livro de Estatística I ou veja meu outro livro, *Estatística Para Leigos*, da Alta Books, para uma revisão completa.)

## *Parâmetro da população*



O *parâmetro* é um número que resume a *população*, grupo de interesse de sua pesquisa. Alguns exemplos de parâmetros incluem a média de uma população, sua mediana ou uma proporção que se enquadra em determinada categoria.

Suponha que você queira determinar a duração média de uma chamada de telefone celular entre adolescentes (com idades entre 13 e 18 anos). Você não está interessado em fazer comparações; o que você quer é uma boa estimativa do tempo médio de duração. Para isso, você deve, então, estimar um parâmetro populacional (como a média aritmética). A população é composta por todos os usuários de telefone celular com idades entre 13 e 18 anos. O parâmetro é a duração média de uma chamada feita por essa população.

## *Estatística amostral*

Normalmente, não é possível determinar com exatidão os parâmetros populacionais, podemos apenas estimá-los. Mas nem tudo está perdido; através da coleta de uma *amostra* (um subconjunto de indivíduos) da população e de seu estudo, é possível obter uma boa estimativa do parâmetro populacional. A *estatística amostral* é um único número que resume esse subconjunto.

Por exemplo, no caso do telefone celular descrito na seção anterior, você selecionaria uma amostra de adolescentes e mediria a duração de suas chamadas durante um período de tempo (ou obteria esses dados através dos registros em seus telefones, se pudesse ter acesso legal a eles). Em seguida, calcularia a média da duração das chamadas. Por exemplo, a duração média de 100 chamadas poderia ser igual a 12,2 minutos — essa média é uma estatística. Essa estatística em particular é chamada de *média amostral*, pois trata-se de um valor médio retirado de seus dados amostrais.

Existem muitas estatísticas para o estudo de diferentes características de uma amostra, tais como a proporção, a mediana e o desvio padrão.

## ***Intervalo de confiança***

O *intervalo de confiança* é um conjunto de possíveis valores para um parâmetro populacional com base em uma amostra e nas estatísticas que resultam dessa amostra. A principal razão para que você tenha um conjunto de possíveis valores, em vez de apenas um único número, é que os resultados das amostras variam.

Por exemplo, suponha que você queira estimar a porcentagem de pessoas que comem chocolate. De acordo com o Simmons Research Bureau, 78% dos adultos entrevistados comem chocolate, e desses, 18% admitem comer doces com regularidade. O que está faltando nesses resultados? Esses números resultaram de uma única amostra de pessoas, e esses valores amostrais, com certeza, variam de amostra para amostra. Por isso, é preciso ter uma medida para o quanto esses resultados mudariam caso você repetisse o estudo.

Essa variação de amostra para amostra, esperada para sua estatística, é medida pela *margem de erro*, que reflete um certo número de desvios padrões somados e subtraídos à estatística, para que você obtenha uma determinada confiança em seus resultados (veja o Capítulo 3 para mais informações sobre margem de erro). Se os resultados dos que comem chocolate se baseassem em uma amostra formada por mil pessoas, a margem de erro seria de aproximadamente 3%. Isso significa que a porcentagem real de pessoas que comem chocolate esperada para a população total seria de  $78\% \pm 3\%$  (ou seja, entre 75 e 81%).

## ***Teste de hipótese***

O *teste de hipótese* é um procedimento estatístico usado para avaliar uma afirmação existente sobre uma população usando seus dados. A afirmação é representada por  $H_0$  (hipótese nula). Caso seus dados comprovem a hipótese, você não pode rejeitar a  $H_0$ . Entretanto, caso seus dados não comprovem a hipótese, você deve rejeitar a  $H_0$  e elaborar uma hipótese alternativa,  $H_a$ . A razão pela qual muitos conduzem um teste de afirmação não é a de meramente mostrar que seus dados comprovam uma hipótese existente, mas a de mostrar que a hipótese existente é falsa, em favor da hipótese alternativa.

O Pew Research Center estudou a porcentagem de pessoas que assistem à ESPN para saber as notícias do mundo esportivo. Suas estatísticas, baseadas em uma pesquisa com cerca de mil pessoas, apontaram que, no ano 2000, 23% das pessoas diziam preferir a ESPN, e, em 2004, esse número caiu para apenas 20%. A pergunta é: essa redução de 3% de telespectadores de 2000 para 2004 representa uma tendência significativa com a qual a ESPN deveria se preocupar?

Para testar formalmente essas diferenças, você pode elaborar um teste de hipótese. Você toma sua hipótese nula como o resultado que você tem de acreditar, sem estudo,  $H_0 =$  não existe diferença entre os dados de 2000 e 2004 para a audiência da ESPN. Sua hipótese alternativa ( $H_a$ ) é a de que existe uma diferença. Para conduzir um teste de hipótese, observe a diferença entre a estatística obtida a partir de seus dados e a afirmação que já foi feita sobre a população (em  $H_0$ ); Então, meça o quanto elas se distanciam em unidades de

desvios padrão.

Com relação ao exemplo, usando as técnicas do Capítulo 3, o teste de hipótese mostra que 23% e 20% não se distanciam o suficiente, em termos de desvios padrão, para disputar a hipótese ( $H_0$ ). Sendo assim, você não pode afirmar que a porcentagem de telespectadores da ESPN na população total tenha sofrido uma alteração de 2000 para 2004.



Assim como em qualquer análise estatística, suas conclusões podem estar erradas apenas por acaso, uma vez que seus resultados se baseiam em dados amostrais e os resultados amostrais variam. No Capítulo 3, discuto os tipos de erros que podem ser cometidos em conclusões tiradas a partir de um teste de hipótese.

## ***Análise de variância (ANOVA)***

ANOVA é o acrônimo para *análise de variância* (do inglês *analysis of variance*). A ANOVA é utilizada quando você quer comparar as médias de mais de duas populações. Por exemplo, digamos que você queira comparar o tempo de vida de quatro marcas de pneus em número de milhas. Você coleta uma amostra aleatória de 50 pneus de cada grupo, somando um total de 200 pneus, elabora um experimento para comparar o tempo de vida de cada um e faz registros. Ao final, você obtém quatro médias e quatro desvios padrão, um para cada conjunto de dados.

Em seguida, para testar as diferenças no tempo de vida médio das quatro marcas, você, basicamente, compara a variabilidade entre os quatro conjuntos de dados à variabilidade dentro de um conjunto inteiro, usando uma razão. Essa razão é denominada *estatística-F*. Caso essa razão seja grande, a variabilidade entre as marcas é maior do que a variabilidade dentro de cada marca, deixando claro que nem todas as médias são iguais para as diferentes marcas de pneus. No entanto, se a estatística- $F$  for pequena, isso indica que não existe uma diferença suficiente entre as médias do tratamento comparadas à variabilidade geral dentro de cada tratamento. Neste caso, não se pode dizer que as médias para cada grupo são diferentes. (Nos Capítulos 9 e 10, você vai encontrar mais detalhes sobre ANOVA, além de todos os jargões, todas as fórmulas e o resultado obtido por computador.)

## ***Comparações múltiplas***

Suponha que você realize uma ANOVA e descubra uma diferença na média de vida útil das quatro marcas de pneus (veja a seção anterior). Suas próximas perguntas, provavelmente, serão: “Quais marcas se diferem?” e “Quão diferentes elas são?” Para responder a essas perguntas, utilize os procedimentos de comparação múltipla.

O *procedimento de comparação múltipla* é uma técnica estatística que compara as médias entre si e descobre as que se diferem. Com essa informação, você consegue classificá-las da maior para a menor, levando em conta que, em alguns casos, dois ou mais grupos



poderão ter médias tão próximas, a ponto de terem de ser classificados na mesma posição.

Existem muitos tipos diferentes de procedimentos de comparação múltipla que comparam as médias individuais e estabelecem uma ordem caso sua estatística- $F$  realmente tenha encontrado a existência de uma diferença. Alguns dos procedimentos de comparação múltipla incluem o teste de Tukey, LSD e os testes- $t$  pareados. Alguns procedimentos serão melhores do que outros, dependendo das condições e de seu objetivo como analista de dados. No Capítulo 11, discuto os procedimentos de comparação múltipla com mais detalhes.

Não dê o segundo passo a fim de comparar as médias dos grupos caso a ANOVA não encontre qualquer resultado significativo durante a primeira fase. O programa nunca vai impedi-lo de fazer uma análise dessas, mesmo quando a realização de tal procedimento for errada.

## *Efeitos de interação*

O *efeito de interação* na estatística opera da mesma forma que no mundo da medicina. Às vezes, quando você toma dois medicamentos distintos no mesmo dia, o efeito combinado dos dois é muito diferente do obtido quando você os toma separadamente.

Os efeitos de interação podem aparecer em modelos estatísticos que utilizam duas ou mais variáveis para explicar ou comparar resultados. Nesse caso, você não pode, automaticamente, estudar o efeito de cada variável de forma separada; primeiro, é preciso verificar a presença do efeito de interação.

Por exemplo, suponha que pesquisadores da área médica estejam estudando um novo medicamento para depressão e desejem saber como essa droga afeta a mudança da pressão arterial tanto com grandes quanto com pequenas doses. Também será comparado o contraste entre os efeitos em crianças e adultos. Também pode ser que o nível da dosagem afete de modo diferente a pressão arterial dos adultos e das crianças. Esse tipo de modelo se chama *ANOVA com dois fatores*, com um provável efeito de interação entre os dois fatores (idade e nível da dosagem). O Capítulo 11 abrange esse assunto com mais profundidade.

## *Correlação*

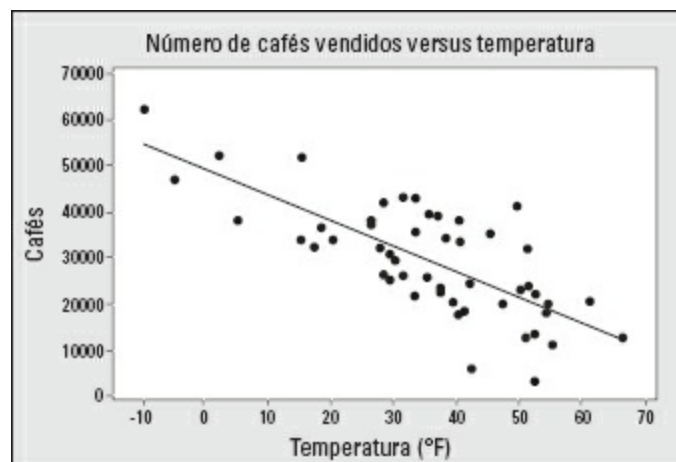
O termo *correlação* é frequentemente utilizado de forma equivocada. Em estatística, a correlação mede a força e a direção de uma relação linear entre duas *variáveis quantitativas* (variáveis que apenas representam contagens e medidas).

Não utilize a correlação para descrever relações, a não ser quando as variáveis são quantitativas. Por exemplo, é errado dizer que existe uma correlação entre a cor dos olhos e a dos cabelos. (No Capítulo 14, você vai explorar as associações entre as variáveis categoriais.)



A correlação é um número entre  $-1,0$  e  $+1,0$ . Uma correlação de  $+1,0$  indica uma relação positiva perfeita; à medida que uma variável aumenta, a outra aumenta em perfeita sincronia. Uma correlação de  $-1,0$  indica uma relação negativa perfeita; à medida que uma variável aumenta, a outra diminui em perfeita sincronia. Uma correlação igual a zero indica a ausência de uma relação linear entre todas as variáveis. A maioria das correlações no mundo real fica entre  $-1,0$  e  $+1,0$ ; quanto mais próximas estiverem de  $-1,0$  ou de  $+1,0$ , mais forte será a relação; quanto mais próximas estiverem de 0, mais fraca será a relação.

A Figura 1-1 mostra o gráfico do número de cafés vendidos em partidas de futebol americano realizadas em Buffalo, Nova York, bem como a temperatura do ar (em graus Fahrenheit) em cada partida. Este conjunto de dados parece seguir uma reta descendente, indicando a existência de uma correlação negativa. A correlação se revela igual a  $-0,741$ ; o número de cafés vendidos possui uma relação negativa relativamente forte com a temperatura no dia da partida. Isso faz sentido, uma vez que, nos dias em que a temperatura está baixa, as pessoas sentem frio e querem mais café. Discuto a correlação mais adiante, no Capítulo 4, já que ela se aplica à construção de modelos.



---

**Figura 1-1:** Cafés vendidos em dias de partidas de futebol de acordo com diferentes temperaturas.

---

## ***Regressão linear***

Depois de encontrar uma correlação e determinar a existência de uma relação linear relativamente forte entre duas variáveis, você pode tentar fazer previsões para uma variável com base no valor da outra. Por exemplo, caso você saiba da existência de uma relação linear negativa relativamente forte entre o número de cafés vendidos e a temperatura no dia da partida (veja a seção anterior), poderá usar essa informação para prever a quantidade de café necessária para o jogo com base na temperatura. O método para encontrar a reta que melhor represente a relação entre duas variáveis é chamado de *regressão linear*.

Há muitos tipos diferentes de análise de regressão, dependendo de sua situação. Quando apenas uma variável é utilizada para prever um resultado, o método de regressão é

chamado de *regressão linear simples* (veja o Capítulo 4). A regressão linear simples é a mais conhecida entre todas as análises de regressão e é essencial à continuação de um curso sequencial de estatística.

Entretanto, há outros tipos de regressão para outras situações.

- ✓ Quando mais do que uma variável é utilizada para prever um resultado, o método de regressão é chamado de *regressão linear múltipla* (veja o Capítulo 5).
- ✓ Quando a variável utilizada para a previsão de um resultado tiver apenas dois resultados, sim ou não, o método usado é a *regressão logística* (veja o Capítulo 8).
- ✓ Para relações que não seguem uma reta, existe uma técnica chamada (não se surpreenda) *regressão não linear* (veja o Capítulo 7).

## Testes Qui-quadrados

Todas as técnicas de regressão e correlação partem do princípio de que a variável que está sendo estudada mais detalhadamente (a variável de resposta) é quantitativa — ou seja, a variável mede ou conta alguma coisa. Você também pode se deparar com situações em que os dados que estão sendo estudados não sejam quantitativos, mas categóricos — ou seja, os dados representam categorias, e não medidas ou contagens. Para estudar as relações entre dados categóricos, você vai utilizar um teste do Qui-quadrado para independência. Caso se descubra que as variáveis não se relacionam entre si, estas são declaradas independentes. Mas, se existir uma relação entre elas, estas são declaradas dependentes.

Suponha que você queira explorar a relação entre o gênero e o hábito de tomar café da manhã. Uma vez que cada uma dessas variáveis é categórica ou qualitativa, você vai utilizar um teste do Qui-quadrado para independência. Em uma entrevista com 70 homens e 70 mulheres, você descobre que 25 homens tomam café da manhã e 45 não; no caso das mulheres, 35 fazem essa primeira refeição do dia e 35 não. A Tabela 1-1 organiza esses dados e os prepara para um cenário da utilização do teste Qui-quadrado.

**Tabela 1-1** Configuração para questionário Café da manhã e gênero

	<i>Tomam café da manhã</i>	<i>Não Tomam café da manhã</i>	<i>Total</i>
<i>Homens</i>	25	45	70
<i>Mulheres</i>	35	35	70

O teste Qui-quadrado primeiro calcula o valor possível em cada célula da tabela caso as variáveis sejam independentes (esses valores recebem o brilhante nome de *frequências esperadas*). O teste Qui-quadrado, então, compara esses valores possíveis aos valores observados nos dados (chamados de *frequências observadas*), usando a estatística Qui-quadrado.

Na comparação café da manhã e gênero, o número de homens que toma café da manhã é



menor do que o número de mulheres ( $25 \div 70 = 35.7\%$  comparado a  $35 \div 70 = 50\%$ ). Embora você saiba que os resultados variam de amostra para amostra, essa diferença é suficiente para declarar a existência de uma relação entre o gênero e o hábito de tomar café da manhã, segundo o teste Qui-quadrado de independência. O Capítulo 14 revela todos os detalhes da realização de um teste Qui-quadrado.

Você também pode usar o teste Qui-quadrado para verificar se sua teoria sobre a porcentagem de cada grupo que se enquadra em determinada categoria é verdadeira ou falsa. Por exemplo, você consegue adivinhar a porcentagem de M&M's que se enquadra em cada categoria de cor? Encontre mais sobre essas variações do Qui-quadrado, bem como a resposta para a questão dos M&M's, no Capítulo 15.

## ***Estatística não paramétrica***

A *estatística não paramétrica* é uma área da Estatística que fornece técnicas de análise usadas quando as condições para o uso de métodos mais tradicionais e mais comuns não são atendidas. No entanto, às vezes, as pessoas se esquecem ou não se preocupam em checar tais condições. Logo, se as condições realmente não forem atendidas, toda a análise é jogada fora com as conclusões.

Suponha que você esteja tentando testar uma hipótese sobre a média de uma população. A abordagem mais comum nessa situação é o uso do teste-*t*. No entanto, para usar um teste-*t*, os dados precisam ser coletados de uma população que tenha distribuição normal (ou seja, ela deve ter uma curva simétrica em forma de sino). Então, depois de coletar os dados e fazer o gráfico, você descobre que, em vez de uma distribuição normal, tem uma assimétrica. Você se vê em um beco sem saída — não pode usar os procedimentos comuns de teste de hipótese que conhece e adora (pelo menos, não deveria).

É aí que os procedimentos não paramétricos entram em cena. Esses procedimentos não requerem tantas condições quanto os procedimentos paramétricos. No caso de dados assimétricos, é sensato realizar um teste de hipótese para a mediana, em vez de usar a média, e muitos são os procedimentos não paramétricos que existem para isso.

Caso as condições para o procedimento de análise dos dados que você deseja realizar não sejam atendidas, não entre em pânico, é muito provável que um procedimento não paramétrico esteja esperando por você. Grande parte dos programas pode realizá-los com a mesma facilidade que realizariam os procedimentos mais comuns (paramétricos).

Os programas de estatística não verificam as condições antes de conduzir uma análise de dados. Você é quem deve checar todas e quaisquer condições necessárias, e, caso elas estejam seriamente violadas, altere o curso de sua jornada. Muitas vezes, um procedimento não paramétrico é apenas a passagem. Para mais informações sobre os diferentes procedimentos não paramétricos, veja os Capítulos 16 a 19.



## Capítulo 2

# Encontre a Análise Certa para o Problema

---

### *Neste Capítulo*

- ▶ Decifre a diferença entre variáveis categóricas e quantitativas
  - ▶ Selecione as técnicas adequadas para a tarefa em mãos
  - ▶ Avalie os níveis de viés e de precisão
  - ▶ Interprete os resultados de forma apropriada
- 

**U**m dos elementos mais críticos da estatística e da análise de dados é a habilidade de escolher a técnica certa para cada tarefa a ser realizada. Carpinteiros e mecânicos sabem a importância de se ter a ferramenta certa na hora certa e os problemas que podem ocorrer caso a ferramenta errada seja utilizada. Também sabem que usar a ferramenta certa aumenta suas chances de conseguir o resultado desejado logo na primeira tentativa, segundo a abordagem do “seja inteligente e economize esforço”.

Neste capítulo, você verá algumas das principais técnicas de análise estatística do ponto de vista de carpinteiros e mecânicos — conhecendo para o que serve cada ferramenta estatística, como e quando usá-la. Também vamos focar nos erros que alguns devoradores de números cometem ao aplicarem a análise errada ou realizarem análises demais.



Saber identificar esses problemas pode ajudá-lo a evitá-los, mas também o ajuda a navegar no oceano de estatísticas que pode estar à sua espera no trabalho ou em situações do dia a dia.

Caso muitas das ideias contidas neste capítulo lhe pareçam grego e você ache que precisa de mais informações, não se aflija. Antes de continuar a ler, consulte um livro de Estatística I ou leia *Estatística Para Leigos* (Alta Books), o primeiro livro da série.

# Variáveis Categóricas versus Variáveis Quantitativas

Depois de coletar todos os dados necessários de uma amostra, você deve organizá-los, resumi-los e analisá-los. Porém, antes de se lançar a todos aqueles cálculos numéricos é preciso, primeiro, identificar o tipo de dado com o qual você está lidando e que direciona o ao tipo correto de gráfico, estatística e análise que pode ser usado.



Antes de começar, aqui vai um jargão importante: os estatísticos denominam qualquer quantidade ou característica medida em um indivíduo de *variável*; os dados coletados sobre uma variável podem mudar de pessoa para pessoa (daí esse nome tão criativo).

Os dois principais tipos de variáveis são:

- ✓ **Categóricas:** A *variável categórica*, também conhecida como *variável qualitativa*, classifica o indivíduo segundo categorias. Por exemplo, a afiliação política nos Estados Unidos poderia ser classificada em quatro categorias: Democratas, Republicanos, Independentes e Outros; o gênero dos seres vivos, quando considerado uma variável, possui apenas duas possíveis categorias: macho e fêmea. As variáveis categóricas podem utilizar valores numéricos apenas como representantes (sem significado numérico).
- ✓ **Quantitativas:** A *variável quantitativa* mede ou conta uma característica quantificável, tal como altura, peso, número de filhos, média de notas escolares ou o número de horas que você dormiu na noite passada. O valor da variável quantitativa representa uma quantidade (contagem) ou uma medida e possui um significado numérico. Ou seja, você pode somar, subtrair, multiplicar ou dividir os valores de uma variável quantitativa, e os resultados obtidos terão valor numérico real.

Uma vez que os dois modelos de variáveis representam tipos de dados tão diferentes, faz sentido que cada um tenha seu próprio conjunto de estatísticas. As variáveis categóricas, tais como o gênero, são de alguma forma limitadas com relação ao número de estatísticas que podem ser aplicadas.

Por exemplo, suponha que você tenha uma amostra de 500 alunos classificados por gênero — 180 são do sexo masculino e 320, do feminino. Como resumir essa informação? Você já tem o número total em cada categoria (essa estatística é denominada *frequência*). Já é um bom começo, mas é muito difícil interpretar frequências pois você vai se pegar tentando compará-las a um total em sua mente, a fim de chegar a uma comparação adequada. Por exemplo, neste caso, você deve estar pensando: “Cento e oitenta de quanto?” Vejamos... de 500. Hmmm! Qual é a porcentagem disso? O próximo passo é encontrar um meio de relacionar esses números, uns aos outros, de forma fácil. Podemos fazer isso usando a *frequência relativa*, que é a porcentagem dos dados que se enquadram em uma categoria específica de uma variável categórica. Você pode encontrar a frequência relativa de uma categoria ao dividi-la pelo total da amostra e multiplicar o resultado por 100. Nesse caso,

you have  $\frac{180}{500} = 0,36 * 100 = 36\%$  of male students and  $\frac{320}{500} = 0,64 * 100 = 64\%$  of female students.

It is also possible to express the relative frequency as a proportion in each group, leaving the result in decimal form, and not multiplying by 100. This statistic is called *sample proportion*. In this example, the sample proportion of male students is 0,36 and of female students is 0,64.

Basically, categorical variables are summarized in two statistics: the number in each category (frequency) and the percentage (relative frequency) in each category.



# *Estatísticas para Variáveis Categóricas*

Os tipos de estatística realizados em dados categóricos podem parecer limitados; no entanto, a grande variedade de análises que podem ser realizadas com o uso das frequências e frequências relativas geralmente responde a uma vasta gama de possíveis perguntas que você pode querer explorar.

Nesta seção, você vai ver que a proporção dentro de cada grupo é a principal estatística para o resumo dos dados categóricos. Além disso, veremos como usar as proporções para estimar, comparar e procurar relações entre os grupos em que se dividem os dados categóricos.

## *Estimando uma proporção*

As frequências relativas podem ser usadas para fazer estimativas a respeito de uma única proporção populacional. (Consulte na seção anterior, “Variáveis Categóricas versus Variáveis Quantitativas”, a explicação sobre frequências relativas.)

Suponha que você queira saber qual é a proporção de mulheres democratas nos Estados Unidos. Segundo uma pesquisa com 29.839 eleitoras americanas, conduzida pela Pew Research Foundation em 2003, a porcentagem de mulheres democratas era 36. Agora, já que os pesquisadores da Pew basearam seus resultados em apenas uma amostra da população e não na população inteira, esses resultados irão variar se outra amostra for usada. Essa variação nos resultados obtidos por meio de amostras é sabiamente chamado de...Adivinha? Variabilidade amostral.

A variabilidade amostral é medida por meio da *margem de erro* (a quantia somada e subtraída da estatística amostral), que, para essa amostra, é de apenas 0,5%. (Para descobrir como calcular a margem de erro, vá ao Capítulo 3.) Isso significa que a porcentagem estimada de mulheres democratas na população de eleitores americanos está entre 35,5% e 36,5%.



A margem de erro, combinada à proporção amostral, forma o que os estatísticos chamam de intervalo de confiança para a proporção populacional. Lembre-se de que, em Estatística I, o *intervalo de confiança* é um conjunto de possíveis valores para um parâmetro populacional, formado a partir da soma e da subtração da margem de erro à estatística amostral. (Para mais informações sobre intervalos de confiança, veja o Capítulo 3.)

## *Comparando proporções*

Os pesquisadores, a mídia e até mesmo pessoas como eu e você adoram comparar grupos (admita você ou não). Por exemplo, qual a proporção de democratas que apoiam a exploração de petróleo no Alaska, comparada à de republicanos? Qual a porcentagem de mulheres que assistem futebol comparada à de homens? Qual a proporção de leitores do

*Estatística II Para Leigos* que passam em suas provas, comparada aos não leitores?

Para responder a essas perguntas, é preciso comparar as proporções amostrais usando um teste de hipótese para duas proporções. (Veja o Capítulo 3 ou consulte seu livro de Estatística I.)

Suponha que você tenha coletado dados em uma amostra aleatória de mil eleitores americanos e queira comparar a proporção de mulheres à proporção de homens, e descobrir se são iguais. Imagine que, em sua amostra, você perceba que a proporção de mulheres é 0,53 e a de homens é 0,47. Assim, para essa amostra de mil pessoas, a proporção de mulheres é maior do que a de homens.

Mas eis a grande questão: essas proporções são diferentes o bastante para dizer que há mais mulheres do que homens em toda a população de eleitores americanos? Afinal de contas, os resultados amostrais variam de amostra para amostra. A resposta para essa pergunta requer a comparação das proporções amostrais por meio de um teste de hipótese para duas proporções. No Capítulo 3, vou demonstrar e falar mais sobre essa técnica.

## ***Procurando relações entre variáveis categóricas***

Suponha que você queira saber se existe uma relação entre duas variáveis categóricas; por exemplo, a afiliação política está relacionada ao gênero de uma pessoa? A resposta a essa pergunta requer a organização dos dados amostrais em uma tabela 2X2 (tabela que usa linhas e colunas para representar duas variáveis, também chamada de tabela de contingência de dupla entrada ou cruzada) e sua análise através do teste Qui-quadrado (ver Capítulo 14).

Seguindo esse processo, você poderá determinar se duas variáveis categóricas são independentes (não se relacionam) ou se existe uma relação entre elas. Caso encontre uma relação, você pode usar a porcentagem para descrevê-la.

A Tabela 2-1 mostra um exemplo de dados organizados em uma tabela 2X2. Os dados foram coletados pela Pew Research Foundation.

**Tabela 2-1 Relação entre Gênero e Afiliação Política de 56.735 Eleitores Americanos**


<i><b>Gênero</b></i>	<i><b>Republicano</b></i>	<i><b>Democrata</b></i>	<i><b>Outros</b></i>
Homens	32%	27%	41%
Mulheres	29%	36%	35%

Observe que a porcentagem de homens republicanos na amostra é 32 e a porcentagem de mulheres republicanas é 29. Tais porcentagens estão muito próximas em termos relativos. Entretanto, a porcentagem de mulheres democratas parece muito mais alta do que a de homens democratas (36% versus 27%); além disso, a porcentagem de homens na categoria




“Outros” também é um pouco maior do que a porcentagem de mulheres na mesma categoria (41 versus 35%).

Essas grandes diferenças nas porcentagens indicam que o gênero e a afiliação política estão relacionados na amostra. Porém, essas tendências são seguidas pela população de todos os eleitores americanos? Para responder a essa pergunta, precisamos de um teste de hipótese. Uma vez que gênero e afiliação política são duas variáveis categóricas, o teste de hipótese necessário para essa situação é o teste Qui-quadrado. (No Capítulo 14, discuto os testes Qui-quadrado com mais detalhes.)



Para fazer uma tabela 2x2 a partir do conjunto de dados usando o Minitab, primeiro, insira os dados em duas colunas, onde a coluna 1 é a linha variável (neste caso, o gênero) e a coluna 2 é a coluna variável (neste caso, a afiliação política). Por exemplo, suponha que a primeira pessoa seja um homem democrata. Na linha 1 do Minitab, insira *M* (masculino) na coluna 1 e *D* (democrata) na coluna 2. Depois, vá ao menu Stat>Tables>Cross Tabulation e Chi-square. Selecione a coluna um e clique em Select para inserir essa variável no campo For Rows. Selecione a coluna dois e clique em Select para inserir essa variável no campo For Columns. Por último, clique em OK.



Muitas vezes, a palavra *correlação* é usada para discutir as relações entre variáveis, mas em Estatística, o termo correlação apenas relata a relação entre duas variáveis quantitativas (numéricas), e não entre duas variáveis categóricas. A *correlação* mede a proximidade da relação entre duas variáveis quantitativas, tais como altura e peso, acompanha uma reta e também mostra sua direção. Em suma, para quaisquer duas variáveis,  $x$  e  $y$ , a correlação mede a força e a direção de uma relação linear. À medida que uma diminui, o que a outra faz?

Já que as variáveis categóricas não possuem ordem numérica, seus valores não aumentam nem diminuem. Por exemplo, só porque masculino = 1 e feminino = 2, isso não significa que uma mulher vale duas vezes mais do que um homem (embora algumas mulheres possam discordar). Portanto, você não deve usar a palavra *correlação* para descrever a relação entre, digamos, gênero e afiliação política. (O Capítulo 4 abrange o termo correlação em detalhes.)

O termo apropriado para descrever as relações entre variáveis categóricas é *associação*. Você pode dizer que a afiliação política está associada ao gênero e, então, explicar como. (Para mais detalhes a respeito da associação, veja o Capítulo 13.)

## ***Construindo modelos para fazer previsões***

Você pode construir modelos para prever o valor de uma variável categórica com base em outra informação relacionada. Nesse caso, a construção de modelos é mais do que um monte de pecinhas plásticas e cola pegajosa.

Quando se constrói um modelo estatístico, se está buscando variáveis que ajudem a

explicar, estimar ou prever alguma resposta pela qual você se interessa; as variáveis que fazem isso são denominadas *variáveis explicativas*. Busque as variáveis explicativas e descubra as que melhor preveem a resposta. Em seguida, coloque-as em uma equação do tipo  $y = 2x + 4$ , em que  $x$  = tamanho do calçado e  $y$  = comprimento da panturrilha. Essa equação é um *modelo*.

Por exemplo, suponha que você queira saber quais fatores e variáveis podem ajudá-lo a prever a afiliação política de uma pessoa. Uma mulher sem filhos teria mais chances de ser republicana ou democrata? E um homem de meia idade que proclama o hinduísmo como sua religião?

Para comparar essas relações complexas, você deve construir um modelo para avaliar o impacto de cada grupo sobre a afiliação política (ou sobre alguma outra variável categórica). Esse tipo de construção de modelo é explorado com mais profundidade no Capítulo 8, onde discuto a regressão logística.

*A regressão logística* constrói modelos para prever o resultado de uma variável categórica, como a afiliação política. Caso você queira fazer previsões sobre uma variável quantitativa, tais como rendimentos financeiros, é preciso utilizar o tipo padrão de regressão (consulte os Capítulos 4 e 5).



# *Estatísticas para Variáveis Quantitativas*

As variáveis quantitativas, ao contrário das variáveis categóricas, possuem uma ampla variedade de estatísticas que podem ser aplicadas, dependendo das perguntas a serem respondidas. A principal razão para essa ampla variedade é que os *dados quantitativos* são números que representam medidas ou contagens, portanto, faz sentido que eles possam ser ordenados, somados, subtraídos, multiplicados ou divididos — e todos os resultados terão significado numérico. Nesta seção, apresento a principal técnica de análise para dados quantitativos. Falarei mais sobre cada técnica em capítulos mais adiante.

## *Fazendo estimativas*

As variáveis quantitativas tomam a forma de valores numéricos que envolvem contagens ou medidas, portanto, possuem médias, medianas, desvios padrão e todas essas coisas maravilhosas que as variáveis categóricas não têm. Muitas vezes, os pesquisadores querem saber qual é a média ou mediana para uma população (chamadas parâmetros). Isso requer a coleta de uma amostra e a elaboração de um bom palpite, também conhecido como estimativa desse parâmetro.

Para estimar qualquer parâmetro populacional, é preciso ter um intervalo de confiança. No caso das variáveis quantitativas, você teria que encontrar um intervalo de confiança para estimar a média, a mediana ou o desvio padrão de uma população, porém, o parâmetro de maior interesse é, de longe, a média populacional.

O intervalo de confiança para a média populacional é a média amostral mais ou menos a margem de erro. (Para calcular a margem de erro nesse caso, vá ao Capítulo 3.) O resultado será um conjunto de possíveis valores que você terá produzido para a média populacional real. Já que a variável é quantitativa, o intervalo de confiança tomará as mesmas unidades da variável. Por exemplo, os rendimentos domésticos serão expressos em milhares de dólares.

Não existe uma regra para definir o tamanho da margem de erro para uma variável quantitativa; isso depende do que a variável esteja contando ou medindo. Por exemplo, se você quiser saber a renda média de uma família do estado de Nova York, uma margem de erro de mais ou menos \$5.000 parece razoável. Mas, se a variável é a média dos número de passos dados do primeiro ao segundo andar de um sobrado nos Estados Unidos, a margem de erro será muito menor. As estimativas para variáveis categóricas, por outro lado, são porcentagens; e a maioria das pessoas quer que esses intervalos de confiança fiquem dentro de mais ou menos 3%.

## *Fazendo comparações*

Suponha que você queira observar a renda (uma variável quantitativa) e como ela se relaciona com uma variável categórica, tal como o gênero ou uma região do país. Sua

primeira pergunta poderia ser: “Os homens ainda ganham mais do que as mulheres?” Nesse caso, você pode comparar as rendas médias de duas populações — homens e mulheres. Essa avaliação requer um teste de hipótese de duas médias (mais conhecido como teste-*t* para amostras independentes). No Capítulo 3, trago mais informações sobre essa técnica.



Quando comparar as médias de *mais* do que dois grupos, não se apegue a todos os possíveis testes-*t* que podem ser feitos para os pares de médias, pois você deve controlar a taxa geral de erro em sua análise. Análises em excesso podem resultar em erros — e acabar provocando um desastre. Por exemplo, se você conduzir cem testes de hipótese, cada um com uma taxa de erro de 5%, então, em média, cinco desses cem testes terão resultados estatisticamente significativos, simplesmente por acaso, ainda que não exista nenhuma relação real.

Caso queira comparar a média salarial em diferentes regiões do país (Leste, Meio-oeste, Sul e Oeste, por exemplo), você vai precisar de uma análise mais sofisticada, uma vez que estará observando quatro grupos, em vez de apenas dois. O procedimento para comparar mais do que duas médias é chamado de *análise de variância* (abreviado pela sigla ANOVA), e trato disso com mais detalhes nos Capítulos 9 e 10.

## Explorando relações

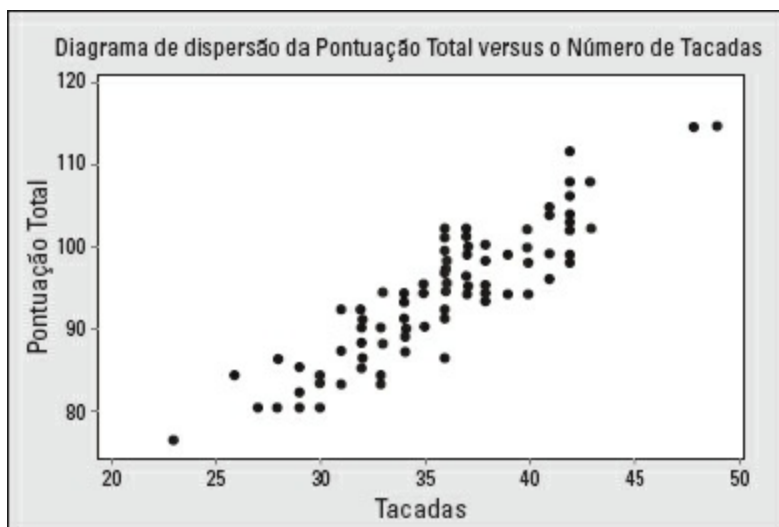
Um das principais razões para a coleta de dados é a observação de relações entre variáveis. No caso das variáveis quantitativas, o tipo mais comum de relação buscada é a linear; ou seja, à medida que uma variável aumenta, a outra aumenta ou diminui de forma semelhante e proporcionalmente? As relações entre quaisquer variáveis são examinadas através de estatísticas e gráficos especializados. Uma vez que a relação linear é muito comum, ela possui sua própria estatística, a chamada correlação. Nesta seção, você vai ficar sabendo como os estatísticos constroem gráficos e fazem estatísticas para investigar relações, dando atenção especial às lineares.

Suponha que você seja um ávido golfista e queira saber quanto tempo deveria gastar treinando suas tacadas. A pergunta é: o número de tacadas se relaciona com a sua pontuação geral? Se a resposta for sim, então faz sentido que você passe mais tempo treinando suas tacadas. Se a resposta for não, então você pode relaxar um pouquinho. Essas duas variáveis são quantitativas, e você está buscando uma ligação entre elas. Para isso, você coleta dados em cem rodadas jogadas por golfistas em seu curso favorito durante um fim de semana. A seguir, veja as primeiras linhas de seu conjunto de dados.

Rodada	Número de Tacadas	Pontuação Total
1	23	76
2	27	80
3	28	80
4	29	80
5	30	80

6	29	82
7	30	83
8	31	83
9	33	83
10	26	84

O primeiro passo para verificar uma ligação entre as tacadas e a pontuação total (ou qualquer outra variável) é fazer um diagrama de dispersão dos dados. O *diagrama de dispersão* é um gráfico em duas dimensões que utiliza um plano de coordenadas X e Y. Veja o diagrama de dispersão dos dados coletados nas rodadas de golfe na Figura 2-1. Aqui,  $x$  representa o número de tacadas e  $y$ , a pontuação geral. Por exemplo, o ponto no canto inferior esquerdo do gráfico representa alguém que deu apenas 23 tacadas e obteve uma pontuação total de 75. (Para instruções sobre como fazer um diagrama de dispersão usando o Minitab, veja o Capítulo 4.)



**Figura 2-1:** O diagrama de dispersão de duas dimensões o ajuda a procurar as possíveis relações dentro de um conjunto de dados.

De acordo com a Figura 2-1, aparentemente, à medida que o número de tacadas aumenta, a pontuação total do golfista também aumenta. A figura também mostra que as variáveis aumentam de forma linear, ou seja, os dados formam um padrão parecido com uma reta. A relação parece ser muito forte — o número de tacadas influencia muito a pontuação total.

Agora, é preciso saber o quão forte é a relação entre  $x$  e  $y$  e se ela é ascendente ou descendente. Medidas diferentes são usadas para diferentes tipos de padrões vistos em um diagrama de dispersão. Já que a relação que vemos nesse caso nos lembra uma reta, a correlação é a medida que utilizamos para quantificá-la. A correlação é o número que mede a proximidade dos pontos em relação à reta. Esse número está sempre entre  $-1,0$  e  $+1,0$ , e, quanto mais próximos os pontos estiverem da reta, mais próxima está a correlação de  $-1,0$  ou  $+1,0$ .

✓ Uma correlação positiva significa que, à medida que  $x$  aumenta no eixo  $x$ ,  $y$

**também aumenta no eixo y.** Os estatísticos chamam esse tipo de *relação ascendente*.

- ✓ **Uma correlação negativa significa que, à medida que x aumenta no eixo x, y diminui.** Os estatísticos chamam esse tipo de relação de...Adivinha? *Relação descendente*.

Para o conjunto de dados do jogo de golfe, a correlação é  $0,896 = 0,90$ , valor extremamente alto, uma vez que o valor de uma correlação é de, no máximo, 1,0. O sinal da correlação é positivo, portanto, à medida que seu número de tacadas aumenta, sua pontuação também aumenta (relação ascendente). Para instruções sobre como calcular a correlação no Minitab, veja o Capítulo 4.

## ***Previendo y através de x***

Caso queira prever alguma variável de resposta (y) usando uma variável explicativa (x) e quiser usar uma reta para fazê-lo, você pode utilizar a *regressão linear simples* (veja o Capítulo 4 para todos os detalhes sobre esse assunto). A regressão linear encontra a reta que melhor representa a relação entre as variáveis — denominada *reta de regressão* — que representa o conjunto de dados. Depois de obter a reta de regressão, você pode atribuir um valor a x e prever o valor de y. (Para instruções sobre como encontrar a reta que melhor representa a relação entre duas variáveis usando o Minitab, veja o Capítulo 4.)

Para usar o exemplo do jogo de golfe da seção anterior, suponha que você queira prever a sua pontuação total obtida a partir de um certo número de tacadas. Nesse caso, você terá que calcular a reta de regressão linear. Ao conduzir uma análise de regressão no conjunto de dados, o computador lhe diz que a melhor reta para prever sua pontuação total a partir do número de tacadas é a seguinte:

$$\text{Pontuação Total} = 39,6 + 1,52 * \text{Número de Tacadas}$$

Sendo assim, se você der 35 tacadas em um curso com 18 buracos, a previsão para sua pontuação total é de cerca de  $39,6 + 1,52 * 35 = 92,8$  ou 93. (Nada mal para 18 buracos!)

Não tente prever y para valores de x que estiverem fora do âmbito em que os dados foram coletados, pois você não terá garantias de que a reta ainda funcionará fora desse âmbito ou, até mesmo, de que continuará fazendo sentido. Para o exemplo do golfe, você não pode dizer que para x (o número de tacadas) = 1, a pontuação total seria de  $39,6 + 1,52 * 1 = 41,12$ . Esse erro é o que chamamos de *extrapolação*.

Se quiser saber mais sobre regressão linear simples e suas expansões, vá aos Capítulos 4 e 5.



# Evitando o Viés

O viés é uma desgraça na vida de um estatístico; é fácil de ser criado e muito difícil (se não impossível) de se lidar na maioria das situações. A definição estatística para *viés* é a super ou a subestimação sistemática de um valor real. Traduzindo: isso significa que os resultados estão sempre desviados do verdadeiro valor do parâmetro segundo determinada quantia em certa direção.

Por exemplo, uma balança de banheiro pode pesar sempre cinco quilos a mais do que deveria (e tenho certeza de que a balança do consultório do meu médico tem esse problema).

O viés pode aparecer em um conjunto de dados de várias formas diferentes. Veja as formas mais comuns que o viés utiliza para sabotar seus dados:

- ✓ **Selecionando uma amostra a partir da população:** O viés ocorre tanto quando alguns grupos que deveriam ser incluídos no processo de amostragem são deixados de fora, como quando determinados grupos recebem importância demais.

Por exemplo, as enquetes realizadas por programas de TV que pedem aos telespectadores que telefonem e deem suas opiniões são enviesadas, pois ninguém realizou uma seleção prévia de uma amostra de pessoas que representem a população — são os próprios telespectadores que se selecionam e participam através de um telefonema. Os estatísticos descobriram que as pessoas que decidem participar de enquetes por telefone ou em sites têm, muito provavelmente, opiniões mais fortes do que as que foram selecionadas aleatoriamente, mas que preferiram não se envolver em tais pesquisas. Tais amostras são denominadas *amostras autosselecionadas* e, em geral, são enviesadas.

- ✓ **Projetando o instrumento para a coleta de dados:** Os instrumentos mal projetados, incluindo as pesquisas de opinião e seus questionários, podem resultar em dados inconsistentes ou até mesmo incorretos. A redação de uma questão desempenha um papel decisivo sobre o enviesamento ou não dos resultados. Uma questão tendenciosa pode fazer com que as pessoas sintam que devem responder de determinada forma. Por exemplo: “Você não acha que o presidente deveria ter o poder de vetar apenas alguns itens de uma lei a fim de evitar os gastos do governo?” Quem pensaria em dizer *não* a essa pergunta?
- ✓ **Coletando dados:** Nesse caso, o viés pode se infiltrar nos resultados se alguém cometer erros durante o registro dos dados ou se os entrevistados saírem do roteiro.
- ✓ **Decidindo como e quando os dados serão coletados:** A hora e o lugar em que os dados são coletados podem influenciar o enviesamento ou não dos dados. Por exemplo, se você conduzir uma enquete por telefone durante o meio do dia, as pessoas que trabalham das 9h às 17h não poderão participar. Dependendo do assunto, o horário da realização da pesquisa pode provocar resultados enviesados.

A melhor forma de lidar com o viés é, em primeiro lugar, evitá-lo, mas também é possível tentar minimizá-lo:

- ✓ **Utilizando um processo aleatório para selecionar amostra a partir da população:** A única situação em que uma amostra é realmente aleatória é quando todos os membros de uma população têm chances iguais de serem selecionados. As amostras autosselecionadas não são aleatórias.
- ✓ **Garantindo que os dados sejam coletados de forma justa e consistente:** Tenha certeza de que suas questões estejam redigidas de forma neutra e de que o horário de realização da pesquisa esteja adequado.

## Uma andorinha não faz verão!

Uma vez, um veterinário pesquisador veio até mim com um conjunto de dados do qual estava muito orgulhoso. Ele estava estudando bovinos e as variáveis envolvidas na determinação de sua longevidade. Seu super conjunto de dados continha mais de 100.000 observações. Ele estava pensando: “Uau, isso vai ser uma maravilha! Tenho coletado esses dados há anos e, finalmente, chegou a hora de eles serem analisados. Vou poder tirar muitas informações daqui. Vou escrever muitos artigos, serei convidado para muitas palestras e vou conseguir aquele aumento!” Esperançoso, ele me apresentou seus preciosos dados com um sorriso e brilho no olhar.

Mas, depois de observar seus dados por alguns minutos, fiz uma terrível constatação — todos os dados vinham exatamente do mesmo animal. Sem outros animais para comparar e com um tamanho amostral de apenas um, ele não tinha nem como medir o quanto aqueles resultados iriam variar caso fossem aplicados a outro animal. Os resultados estavam tão enviesados àquele único animal, que não pude fazer absolutamente nada com os dados. Depois de juntar toda minha coragem para contar a ele a verdade, ele precisou de algum tempo para se recompor. Moral da história: fale sobre seus planos com um estatístico antes de a vaca ir para o brejo.



# Medindo a Precisão Através da Margem de Erro

*Precisão* é a quantia de movimento esperada em seus resultados amostrais caso você repita todos os seus estudos com uma nova amostra. A precisão possui duas formas:

- ✓ **Baixa precisão:** Significa que seus resultados amostrais possuem grande variabilidade (o que não é bom).
- ✓ **Alta precisão:** Significa que seus resultados amostrais permanecem razoavelmente próximos em amostras repetidas (o que é muito bom).

Nesta seção, você vai descobrir o que a precisão mede e o que ela não mede, além de ver como medir a precisão de uma estatística de maneira geral.

Antes de reportar ou tentar interpretar quaisquer resultados estatísticos, você precisa ter uma medida do quanto tais resultados podem variar de amostra para amostra. Essa medição é denominada *margem de erro*. Sempre esperamos, e até mesmo presumimos, que os resultados estatísticos não deveriam mudar muito com outra amostra, mas nem sempre é isso o que acontece.

## Íntimo e pessoal: resultados de pesquisas

A Gallup Organization relata os resultados de suas pesquisas em um formato universal, estatisticamente correto. Usando um exemplo específico de uma pesquisa realizada por eles recentemente, veja a linguagem usada para relatar os resultados:

“Esses resultados se baseiam em entrevistas telefônicas com uma amostra nacional aleatória de 1.002 adultos com idade de 18 anos ou mais, realizadas no período de setembro a novembro de 2006. Para os resultados baseados nessa amostra, é possível dizer, com 95% de confiança, que o erro máximo atribuído à amostragem e a outros efeitos aleatórios é de  $\pm 3$  pontos percentuais. Além do erro de amostragem, a redação das questões e as dificuldades práticas na realização das pesquisas podem introduzir erro ou viés às conclusões feitas a partir de pesquisas de opinião pública.”

A primeira frase da citação se refere a como a Gallup Organization coletou os dados, bem como ao tamanho da amostra. Assim como você já imaginava, a precisão se relaciona ao tamanho amostral, como visto na seção “Medindo a Precisão Através da Margem de Erro”.

A segunda frase da citação se refere à medição da precisão: o quanto a Gallup acha que esses resultados amostrais possam variar. O fato da Gallup ter 95% de confiança significa que se esse processo for repetido várias vezes, em 5% delas, os resultados estariam errados simplesmente por acaso. Essa inconsistência ocorre quando a amostra selecionada para a análise não representa a população — não devido a razões enviesadas, mas, sim, ao acaso. Consulte a seção “Evitando viés” para obter informações sobre o porquê de a terceira frase se incluir nessa citação.

A margem de erro é influenciada por dois elementos:

- ✓ O tamanho da amostra
- ✓ A diversidade na população (também conhecida como desvio padrão populacional)

Você pode ler mais sobre esses elementos no Capítulo 3, mas aqui vai uma explicação geral: à medida que o tamanho amostral aumenta, você tem mais dados com os quais trabalhar, e seus resultados se tornam mais precisos. Como resultado, a margem de erro diminui.

Por outro lado, uma diversidade muito alta na população reduz o nível de precisão, uma vez que a diversidade dificulta a compreensão do que está acontecendo. Como resultado, a margem de erro aumenta. (Para compensar esse problema, basta aumentar o tamanho amostral para recuperar a precisão.)

Para interpretar a margem de erro, pense nela como sendo uma folga que você dá a seus resultados com a finalidade de abranger a maioria das outras amostras que você poderia ter retirado.

Suponha que você esteja tentando estimar, com 95% de confiança em seus resultados, a proporção de pessoas na população que é a favor de uma determinada questão. Você, então, retira uma amostra de 1.002 indivíduos e descobre que 65% deles apoiam a causa. A margem de erro para essa pesquisa acaba sendo de mais ou menos três pontos percentuais (veja os detalhes sobre esse cálculo no Capítulo 3). Esse resultado significa que a proporção amostral de 65% pode expandir-se em três pontos percentuais em ambas direções caso você retire uma amostra diferente de 1.002 indivíduos. Ou seja, acredita-se que a real proporção populacional se encontra em algum ponto entre  $65 - 3 = 62\%$  e  $65 + 3 = 68\%$ . Isso é o máximo que podemos dizer.

Toda margem de erro relatada é calculada sobre a premissa da absoluta ausência de viés nos dados. Entretanto, essa premissa raramente é verdadeira. Antes de interpretar qualquer margem de erro, primeiro verifique se o processo de amostragem e de coleta de dados não contém nenhuma fonte visível de viés. Ignore os resultados que se baseiam em dados enviesados ou, pelo menos, encare-os com uma alta dose de ceticismo.

Para mais detalhes sobre como calcular a margem de erro para várias técnicas estatísticas, vá ao Capítulo 3.



# Conhecendo Seus Limites

O objetivo mais importante de qualquer analista de dados é o de permanecer focado no todo — na pergunta para a qual se está buscando a resposta — e assegurar que a análise de dados usada seja adequada e abrangente o suficiente para respondê-la de forma correta e justa.



Aqui vão algumas dicas de como analisar dados e interpretar resultados em termos de procedimentos e técnicas estatísticas que você pode usar — na escola, no trabalho ou em seu dia a dia. Estas dicas são implantadas e reforçadas ao longo de todo o livro:

- ✓ **Certifique-se de que a pergunta de pesquisa é clara e definitiva.** Alguns pesquisadores não se prendem a nenhum conjunto de perguntas, pois têm a intenção de explorar os dados — buscar qualquer relação que puderem encontrar e, então, estabelecer seus resultados depois dos fatos. Tal prática pode levar ao excesso de análise dos dados, fazendo com que os resultados fiquem sujeitos ao ceticismo dos estatísticos.
- ✓ **Cheque duas vezes seu entendimento sobre o tipo de dado que está sendo coletado.** Dado categórico ou quantitativo? O tipo de dado direciona a abordagem usada na análise.
- ✓ **Tenha certeza de que a técnica usada é projetada para responder a sua pergunta de pesquisa.** Se quiser fazer comparações entre dois grupos e seus dados forem quantitativos, utilize o teste de hipótese para duas médias. Caso queira comparar cinco grupos, utilize a análise de variância (ANOVA). Use este livro como um recurso para ajudá-lo a determinar a técnica necessária.
- ✓ **Conheça os limites da análise de dados.** Por exemplo, várias limitações sérias são encontradas em um estudo que, baseado em um grupo de estudantes universitários, busca saber se anúncios políticos negativos influenciam a população geral de eleitores. Para começar, as reações estudantis aos anúncios negativos não necessariamente representam as reações de toda a população de eleitores. Nesse caso, é melhor limitar as conclusões à população de estudantes universitários (o que nenhum pesquisador jamais gostaria de fazer). O melhor é, antes de qualquer coisa, retirar uma amostra que represente a população pretendida de todos os eleitores (tarefa muito mais complexa, porém compensatória).

## Capítulo 3

# Reverendo Intervalos de Confiança e Testes de Hipótese

---

### *Neste Capítulo*

- ▶ Utilizando intervalos de confiança para estimar parâmetros
  - ▶ Testando modelos por meio dos testes de hipótese
  - ▶ Descobrindo a probabilidade de dar certo e de dar errado
  - ▶ Descobrindo o poder de uma amostra grande
- 

**U**m dos principais objetivos em estatística é usar a informação coletada em uma amostra para obter uma ideia melhor do que acontece com a população estudada (já que as populações são geralmente grandes, e a informação exata, muitas vezes, é desconhecida). Os valores desconhecidos que resumem a população são chamados de *parâmetros populacionais*. Normalmente, o que os pesquisadores querem é compreender tais parâmetros ou testar uma hipótese sobre eles.

Em Estatística I, você provavelmente viu algo sobre intervalos de confiança e testes de hipótese para uma e duas médias e proporções populacionais. Com certeza, seu professor enfatizou o fato de que, independentemente do parâmetro que esteja tentando estimar ou testar, o processo geral é o mesmo. Caso isso não tenha sido feito, não se preocupe; este capítulo esclarecerá esse ponto.

Aqui, serão revisados os conceitos básicos de intervalos de confiança e testes de hipótese, incluindo as probabilidades de cometer erros ao acaso. Também discuto como os estatísticos medem a capacidade de um procedimento estatístico em fazer um bom trabalho — como, por exemplo, o de detectar uma diferença real nas populações.

# *Estimando Parâmetros Usando os Intervalos de Confiança*

O *intervalo de confiança* é a forma com que o estatístico abrange seu palpite quando precisa estimar um parâmetro populacional. Por exemplo, em vez de estimar apenas um número para, digamos, a média da renda familiar nos Estados Unidos, o estatístico apresenta um conjunto de possíveis valores para esse número. E ele o faz, pois:

- ✓ Todo estatístico que se preze sabe que os resultados amostrais variam de amostra para amostra e, portanto, uma estimativa de um único número não é boa.
- ✓ Os estatísticos desenvolveram fórmulas incríveis para obterem o conjunto desses possíveis valores, sendo assim, por que não usá-las?

Nesta seção, você vai ver a fórmula geral para um intervalo de confiança, incluindo a margem de erro, além da abordagem mais comum para a construção dos intervalos de confiança. Também discuto a interpretação e as chances de cometer um erro.

## *Entendendo o básico: A forma geral de um intervalo de confiança*

A grande sacada do intervalo de confiança é a de apresentar um conjunto de possíveis valores para um parâmetro populacional. O *nível de confiança* representa a probabilidade de se obter um conjunto de possíveis valores que realmente contenha o parâmetro populacional real, caso você repetisse o processo de amostragem várias e várias vezes. Traduzindo: o nível de confiança é a probabilidade a longo prazo de tudo estar correto.

A fórmula geral de um intervalo de confiança é:

$$\text{Intervalo de confiança} = \text{Estatística amostral} \pm \text{Margem de erro}$$

O intervalo de confiança possui um determinado nível de precisão (medido pela margem de erro). A precisão mede o quão próximo da verdade você espera que seus resultados estejam.

Por exemplo, suponha que você queira saber o tempo médio que um aluno da Ohio State University (OSU) passa ouvindo música em seu MP3 por dia. O tempo médio para toda a população de alunos da OSU que utiliza MP3 é o parâmetro buscado. Uma amostra de mil alunos é então coletada e descobre-se que o tempo médio por dia que um aluno passa ouvindo música no MP3 é de 2,5 horas e o desvio padrão é de 0,5 horas. É correto afirmar que a população de todos os alunos da OSU que possuem MP3 passam, em média, 2,5 horas ouvindo música diariamente? Você espera e pode até assumir que a média para a população total se aproxime de 2,5 horas, mas esse valor provavelmente não é exato.

Qual seria a solução para esse problema? A solução é relatar, além da média obtida a partir da amostra, a medida de quanto essa média pode variar de uma amostra para outra,



com um determinado nível de confiança. O número utilizado para representar esse nível de precisão para os resultados se chama *margem de erro*.

## ***Encontrando o intervalo de confiança para uma média populacional***

A parte estatística da amostra para a fórmula do intervalo de confiança é bem direta.

- ✓ **Para estimar a média populacional**, utilize a média amostral mais ou menos a margem de erro, que se baseia no erro padrão. A média possui um erro padrão de  $\frac{\sigma}{\sqrt{n}}$ . Nessa fórmula, é possível ver o desvio padrão populacional ( $\sigma$ ) e o tamanho amostral ( $n$ ).
- ✓ **Para estimar a proporção populacional**, utilize a proporção amostral mais ou menos a margem de erro.

Em muitos casos, o desvio padrão da população,  $\sigma$ , é desconhecido. Para estimar a média populacional utilizando um intervalo de confiança quando  $\sigma$  é desconhecido, utilize a fórmula  $\bar{x} \pm t_{n-1} \left( \frac{s}{\sqrt{n}} \right)$ . Esta fórmula contém o desvio padrão amostral ( $s$ ), o tamanho amostral ( $n$ ) e um valor- $t$  que representa a quantidade de erros padrão que você quer adicionar e subtrair a fim de conseguir a confiança necessária. Para obter a margem de erro para a média, observe que o erro padrão,  $\frac{s}{\sqrt{n}}$ , está sendo multiplicado por um fator  $t$ . Note, também, que  $t$  tem o subscrito  $n - 1$  para indicar qual das várias distribuições- $t$  está sendo utilizada para seu intervalo de confiança. O subscrito  $n - 1$  é chamado de *graus de liberdade*.

O valor de  $t$ , nesse caso, representa o número de erros padrão adicionados ou subtraídos da média amostral a fim de se obter a confiança desejada. Caso você queira ter 95% de confiança, por exemplo, adicione e subtraia 1,96 desses erros padrão. Mas, se quiser ter 99,7% de confiança, adicione e subtraia cerca de 3 erros padrão. (Veja a Tabela A-1 no apêndice para descobrir os valores- $t$  para os diferentes níveis de confiança; use  $\left( \frac{1 - \text{nível de confiança}}{2} \right)$  para a área à direita e encontre o valor  $t$  adequado a ela.)

Caso o desvio padrão populacional seja conhecido, não hesite em usá-lo. Neste caso, utilize o número correspondente na distribuição- $Z$  (distribuição normal padrão) na fórmula do intervalo de confiança. (A distribuição- $Z$  em seu livro de Estatística I lhe dará os números de que precisa.) No entanto, seria negligência de minha parte não avisar que, embora os livros didáticos e os professores sempre incluam problemas em que  $\sigma$  é conhecido, raramente ele o é no mundo real. Por que então ensinar dessa forma? Esse assunto está em debate; portanto, vamos deixar as coisas como estão, mas posso mantê-lo atualizado.

Para o exemplo do MP3 da seção anterior, uma amostra aleatória de mil estudantes da OSU demonstrou que esses alunos passam em média 2,5 horas por dia ouvindo música. O desvio padrão é de 0,5 horas. Colocando essa informação na fórmula para o intervalo de



confiança, temos:  $2,5 \pm 1,96 \left( \frac{0,5}{\sqrt{1.000}} \right)$ . E, assim, concluímos que *todos* os alunos da OSU que possuem MP3 passam, em média, entre 2,47 e 2,53 horas ouvindo música.

## ***O que altera a margem de erro?***

O que é preciso saber para calcular a margem de erro? A margem de erro, de modo geral, depende de três elementos:

- ✓ Do desvio padrão da população,  $\sigma$  (ou uma estimativa, denotada por  $s$  do desvio padrão amostral)
- ✓ Do tamanho amostral,  $n$
- ✓ Do nível de confiança necessário

Você vê esses elementos em ação na fórmula para a margem de erro da média amostral:

$\pm t_{n-1} * \frac{s}{\sqrt{n}}$ . Aqui presumo que  $\sigma$  é desconhecido;  $t_{n-1}$  representa o valor na tabela da distribuição- $t$  (ver Tabela A-1 no apêndice) com  $n - 1$  graus de liberdade.

Cada um desses três elementos desempenha um papel importante na definição do tamanho que a margem de erro terá quando a média de uma população for estimada. Nas seções a seguir, mostro como cada um dos elementos da margem de erro funciona separadamente e juntos na fórmula para influenciar seu tamanho.

### ***Desvio padrão populacional***

O desvio padrão da população normalmente se combina ao tamanho amostral na fórmula da margem de erro, ocupando a parte superior da fração enquanto  $n$  ocupa a inferior. (Nesse caso, o erro padrão da população,  $\sigma$ , é estimado pelo desvio padrão da amostra,  $s$ , pois  $\sigma$  normalmente é desconhecido.)

A combinação entre o desvio padrão e o tamanho amostral é conhecida como *erro padrão* da estatística e mede o quanto a estatística amostral se desvia de sua média a longo prazo.

Como o desvio padrão da população ( $\sigma$ ) influencia a margem de erro? À medida que ele aumenta, a margem de erro também aumenta e, assim, o conjunto de possíveis valores fica mais amplo.

Suponha que você tenha dois postos de gasolina, um em uma esquina movimentada (posto de gasolina nº 1) e outro mais afastado das principais vias. Você quer estimar o tempo médio entre os clientes em cada posto. No posto nº 1, o mais movimentado, o uso das bombas é constante e, portanto, não há quase tempo ocioso entre um cliente e outro. No posto nº 2, às vezes os clientes chegam todos ao mesmo tempo e, em outras, não aparece ninguém por até uma hora ou mais. Sendo assim, o tempo entre os clientes varia bastante.

Qual dos postos de gasolina conseguiria estimar mais facilmente o tempo total médio entre





os clientes como um todo? O posto nº 1 tem muito mais consistência, o que representa um desvio padrão de tempo menor entre os clientes. O posto nº 2 possui muito mais variabilidade no tempo entre os clientes. Isso significa que  $\sigma$  para o posto nº 1 é menor do que  $\sigma$  para o posto nº 2. Assim, é mais fácil estimar o tempo médio entre clientes no posto nº 1.

## ***Tamanho amostral***

O tamanho amostral influencia a margem de erro de forma bem intuitiva. Suponha que você esteja tentando estimar o número médio de animais de estimação por residência em sua cidade. Qual tamanho amostral concederia a você melhores informações: dez ou cem residências? Espero que você tenha concordado que cem residências lhe dariam informações mais precisas (desde que os dados dessas cem residências sejam coletados de forma adequada).

Seus resultados são mais precisos quando suas conclusões são tiradas a partir de uma grande quantidade de dados coletados de maneira apropriada. A precisão é medida pela margem de erro, portanto, conforme o tamanho amostral aumenta, a margem de erro de sua estimativa diminui.

Tratando-se de tamanho amostral, uma amostra maior é melhor quando os dados são coletados de forma adequada — ou seja, com o mínimo de viés. Se a qualidade dos dados não puder ser mantida em uma amostra maior, é melhor nem tê-la.

## ***Nível de confiança***

Para cada situação, é preciso definir a confiança necessária nos resultados e, claro, quanto maior a confiança maior a soma na fórmula da margem de erro. Esse nível de confiança em seus resultados ao longo do tempo se reflete em um número denominado *nível de confiança*, demonstrado em forma de porcentagem. De modo geral, quanto maior a confiança, maior o conjunto de possíveis valores. Portanto, à medida que o nível de confiança aumenta, a margem de erro também aumenta.

Toda margem de erro é interpretada com um certo número de erros padrão, os quais serão somados e subtraídos e determinados pelo nível de confiança. Caso precise de mais confiança, adicione e subtraia mais erros padrão. O número que representa a quantidade de erros padrão a serem somados e subtraídos muda de situação para situação. Para uma média populacional, você vai usar um valor na distribuição  $t$ , representado por  $t_{n-1}$ , em que  $n$  é o tamanho amostral (ver Tabela A-1 no apêndice).

Suponha que você tenha um tamanho amostral de 20 e queira estimar a média de uma população com 90% de confiança. O número de erros padrão que você soma e subtrai é representado por  $t_{n-1}$ , que, nesse caso, é  $t_{19} = 1,73$ . (Para encontrar esses valores de  $t$ , veja a Tabela A-1 no apêndice, com  $n - 1$  graus de liberdade para a linha e





$\frac{(1 - \text{nível de confiança})}{2}$  para a coluna.) Agora, suponha que você queira ter 95% de confiança em seus resultados, com o mesmo tamanho amostral de  $n = 20$ . Os graus de liberdade são  $20 - 1 = 19$  (linha), e a coluna é para  $\frac{(1 - 0,95)}{2} = 0,025$ . A tabela- $t$  lhe dá o valor de  $t_{19} = 2,09$ .

Observe que esse valor de  $t$  é maior do que o valor de  $t$  para 90% de confiança, pois, para obter mais confiança, é preciso buscar mais desvios padrão na tabela da distribuição- $t$  para abranger mais resultados possíveis.

### ***Nível de confiança alto, intervalo estreito — a combinação perfeita***

Um intervalo estreito é muito mais desejado do que um muito amplo. Por exemplo, dizer que o custo médio de uma casa nova é de \$150.000 mais ou menos \$100.000 não ajuda muito, pois sua estimativa ficaria entre \$50.000 e \$250.000. (E quem é a pessoa que tem \$100.000 sobrando por aí?). Mas você *realmente* quer um nível de confiança alto e, portanto, seu estatístico tem que somar e subtrair mais erros padrão para consegui-lo, fazendo com que o intervalo fique muito extenso (uma pena).

Mas, espere, não entre em pânico — você pode ter os dois! Se você sabe que precisa de um nível de confiança alto, mas não quer um intervalo de confiança amplo, aumente o tamanho amostral para satisfazer o nível de confiança desejado.

Suponha que o desvio padrão dos preços de casas do estudo anterior seja  $s = \$15.000$  e que você queira ter 95% de confiança em sua estimativa para o preço médio de casas. Usando um tamanho amostral grande, seu valor de  $t$  (da Tabela A-1 no apêndice) é 1,96.

Com uma amostra de cem casas, sua margem de erro será  $\pm 1,96 * \frac{15.000}{\sqrt{100}} = \$2.940$ . Se esse valor for alto demais, mas você ainda quiser 95% de confiança, aumente o valor de  $n$ . Se você coletar uma amostra com 500 casas, a margem de erro diminui para  $\pm 1,96 * \frac{15.000}{\sqrt{500}}$ , reduzindo para \$1.314,81.

Você pode usar uma fórmula para encontrar o tamanho amostral necessário para obter a margem de erro desejada. A fórmula é  $\left( \frac{t_{\text{margem}} s}{MOE} \right)^2$ , onde  $MOE$  é a margem de erro desejada (em forma de proporção),  $s$  é o desvio padrão amostral e  $t$  é o valor da distribuição- $t$  que corresponde ao nível de confiança requerido. (Para tamanhos amostrais grandes, a distribuição- $t$  é aproximadamente igual à distribuição- $Z$ ; você pode usar a última linha da Tabela A-1 do apêndice para obter os valores- $t$  adequados, ou utilizar uma tabela com os valores- $Z$ , encontrada em qualquer livro de Estatística I.)

### ***Interpretando um intervalo de confiança***

A interpretação de um intervalo de confiança envolve uma série de questões sutis, mas importantes. A grande sacada do *intervalo de confiança* é a de representar um conjunto de



possíveis valores para um parâmetro populacional, com base em sua amostra. Entretanto, ele não é interpretado apenas com relação à sua amostra, mas com relação a um número infinito de outras amostras que poderiam ter sido selecionadas. Por exemplo, suponha que mil pessoas colem uma amostra cada uma e calculem um intervalo de confiança de 95% para a média de suas amostras. O termo "confiança de 95%" significa que, desses mil intervalos de confiança, cerca de 950 podem realmente acertar a média. (Acertar a média significa que o intervalo de confiança realmente contém o verdadeiro valor do parâmetro.)



Um intervalo de 95% de confiança não significa que o seu intervalo em particular tenha 95% de chance de capturar o verdadeiro valor do parâmetro; depois que a amostra já tiver sido coletada, o parâmetro pode ou não estar dentro do intervalo. Um intervalo de confiança representa as chances de captura do verdadeiro valor do parâmetro populacional ao longo de várias amostragens diferentes.

Suponha que uma empresa de pesquisa queira estimar, com 95% de confiança em seus resultados, a porcentagem de pessoas nos Estados Unidos que dirigem um carro com mais de 161 mil quilômetros rodados. Para isso, a empresa coleta uma amostra aleatória de 1.200 pessoas e descobre que 420 delas (35%) possuem um carro com essa quilometragem; a margem de erro é, então, de mais ou menos 3%. (Veja em seu livro de Estatística I como determinar a margem de erro para porcentagens.)

Boa parte da interpretação está no nível de confiança — que, neste caso, é de 95%. Já que a empresa coletou uma amostra de 1.200 americanos, perguntou a cada um deles se a quilometragem de seus carros estava acima de 161 mil quilômetros e calculou um intervalo de confiança para os resultados, a empresa está, em essência, levando em conta todas as outras amostras que poderiam ter sido selecionadas ao construir a margem de erro ( $\pm 3\%$ ). A empresa quer abranger 95% dessas outras situações, condição satisfeita por uma margem de erro de  $\pm 3\%$ .

Outra forma de entender o intervalo de confiança é imaginar que, se a empresa repetisse várias e várias vezes a coleta de amostras com 1.200 pessoas e calculasse um intervalo de confiança para cada resultado, 95% desses intervalos de confiança estariam corretos. (Você só precisa torcer para que o seu fosse um desses resultados corretos.)



Usando a notação estatística, é possível escrever os níveis de confiança como  $(1 - \alpha)\%$ . Sendo assim, se você quiser 95% de confiança, escreva-o como  $1 - 0,05$ . Aqui,  $\alpha$  representa a chance de que seu intervalo de confiança seja um dos errados. Esse número,  $\alpha$ , também se relaciona à chance aleatória de que um erro seja cometido em um teste de hipótese, o qual explico na seção "Alarmes falsos e oportunidades perdidas: erros Tipo I e Tipo II".

# *O que é que os Testes de Hipótese Têm?*

Imagine que uma transportadora afirme que suas entregas são realizadas pontualmente em 92% das vezes, ou que um funcionário da universidade diga que 75% dos estudantes moram fora do campus. Caso você queira questionar tais afirmações, como usar a estatística para investigá-las?

Nessa seção, você vai ver as grandes sacadas dos testes de hipótese que servem como base para as técnicas de análise de dados apresentadas neste livro. Vamos revisar e expandir os conceitos envolvidos em um teste de hipótese, incluindo as hipóteses, a estatística de teste e o valor- $p$ .

## *O que $H_0$ e $H_a$ realmente representam?*

O teste de hipótese é utilizado em situações em que você possui um determinado modelo em mente e quer saber se tal modelo se adequa a seus dados. Esse modelo pode ser um que gire em torno apenas da média populacional (testando, por exemplo, se essa média é igual a dez). Ou ele também pode testar o coeficiente angular de uma regressão linear (se ele é ou não igual a zero, por exemplo, com zero significando a ausência de relação entre  $x$  e  $y$ ). Ou, ainda, você pode estar tentando usar muitas variáveis diferentes para prever a viabilidade de um produto no mercado e, por acreditar que um modelo que utilize a idade do cliente, o preço do produto e seu posicionamento na prateleira pode ajudá-lo a fazer essa previsão, vai precisar conduzir um ou mais testes de hipótese para averiguar o funcionamento de tal modelo. (Esse processo em particular é denominado regressão múltipla, e você pode encontrar mais informações sobre ele no Capítulo 5.)

Um teste de hipótese se divide em:

- ✓ **A hipótese nula,  $H_0$ :**  $H_0$  simboliza a situação atual — a que todos assumem como sendo verdadeira até que se prove o contrário.
- ✓ **A hipótese alternativa,  $H_a$ :**  $H_a$  representa o modelo alternativo a ser considerado. Essa é a hipótese do pesquisador, a quem cabe o ônus da prova.

$H_0$  é o modelo que está sendo julgado. Caso você reúna evidências suficientes contra ele, vai poder concluir que  $H_a$ , o modelo defendido, é o correto. No entanto, caso você não reúna evidências suficientes contra  $H_0$ , não poderá dizer que seu modelo ( $H_a$ ) é o correto.

## *Reunindo evidências em uma estatística de teste*

A *estatística de teste* é, basicamente, a estatística obtida a partir de uma amostra e padronizada de forma que possa ser analisada em uma tabela. Embora cada teste de hipótese seja um pouco diferente, a ideia principal é a mesma. Calcule sua estatística e padronize-a de modo que você possa usar a tabela correspondente a ela. Em seguida, procure sua estatística de teste em uma tabela para ver sua posição. Essa tabela pode ser a



tabela- $t$  (Tabela A-1 no apêndice), a tabela do Qui-quadrado (Tabela A-3 no apêndice) ou outra. É o tipo de teste necessário a seus dados que define a tabela a ser usada.

No caso do teste de hipótese para uma média populacional,  $\mu$ , utilize a média amostral,  $\bar{x}$ , como sua estatística. Para padronizá-la, calcule  $\bar{x}$  e converta-a em um valor de  $t$  usando a

fórmula  $t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ , em que  $\mu_0$  é o valor de  $H_0$ . Esse valor vai ser sua estatística de teste, a qual será comparada à distribuição- $t$ .

## ***Determinando a força da evidência através do valor- $p$ .***

Se quiser saber se seus dados possuem a força necessária para enfrentar a  $H_0$ , você precisará descobrir o valor- $p$  e compará-lo a um valor de corte predeterminado,  $\alpha$  (que, normalmente, é 0,05). O *valor- $p$*  é a medida da força de sua evidência contra a  $H_0$ . O valor- $p$  é calculado da seguinte forma:

- 1. Calcule a estatística de teste (consulte a seção anterior para mais informações sobre esse assunto).**
- 2. Procure a estatística de teste na tabela adequada (tal como a tabela- $t$ , a Tabela A-1 no apêndice).**
- 3. Encontre a porcentagem de valores na tabela que esteja além de sua estatística de teste. Essa porcentagem é o valor- $p$ .**
- 4. Se sua  $H_a$  "não for igual a", dobre a porcentagem obtida no terceiro passo, pois sua estatística de teste pode ter tomado qualquer uma das direções antes dos dados serem coletados. (Consulte seu livro de Estatística I ou o *Estatística Para Leigos, da Alta Books*, para uma explicação mais completa sobre como obter o valor- $p$  para testes de hipótese.)**

Seu amigo  $\alpha$  é o valor de corte para o valor- $p$  (normalmente, o valor estabelecido para  $\alpha$  é 0,05, mas, às vezes, também pode ser 0,10). Se seu valor- $p$  for menor do que o valor predeterminado para  $\alpha$ , rejeite a  $H_0$ , pois você terá evidência suficiente contra ela. Se seu valor- $p$  for maior ou igual a  $\alpha$ ,  $H_0$  não poderá ser rejeitada.

Por exemplo, se seu valor- $p$  for 0,002, sua estatística de teste está tão distante da  $H_0$ , que as chances de obter esse resultado ao acaso é de apenas 2 em 1.000. Assim, é possível concluir que a  $H_0$  é provavelmente falsa. Porém, se seu valor- $p$  for 0,30, esse mesmo resultado poderá aparecer em 30% das vezes, portanto, jogue a toalha, pois a  $H_0$  não poderá ser rejeitada. Você não tem evidência suficiente contra ela. Se seu valor- $p$  for muito próximo ao valor de corte, digamos  $p = 0,049$  ou 0,51, pode-se dizer que o resultado é marginal, e você deve deixar que o leitor tire suas próprias conclusões. Esta é a maior vantagem do valor- $p$ : ele permite que outras pessoas determinem se sua evidência é forte o bastante para rejeitar a  $H_0$  em suas mentes.



# *Alarmes falsos e oportunidades perdidas: Erros Tipo I e Tipo II*

Qualquer técnica estatística que você utilize para tirar conclusões sobre uma população, com base em uma amostra de dados, tem a chance de conter um erro. Os erros a que me refiro, os erros Tipo I e Tipo II, ocorrem ao acaso.

A forma como o teste é montado pode ajudar a reduzir esses tipos de erro, mas eles sempre serão um perigo. Você, como analista de dados, precisa saber como medir e entender o impacto dos erros que podem ocorrer com um teste de hipótese e o que pode ser feito para, na medida do possível, minimizá-los. Nas seções a seguir, vou lhe mostrar como fazer isso.

## *Alarmes falsos em virtude de erros Tipo I*

O *erro Tipo I* é a probabilidade condicional de rejeitar a  $H_0$  quando ela é de fato verdadeira. Vejo o erro Tipo I como um alarme falso: você apertou o botão quando não deveria.

A chance de cometer um erro do Tipo I é igual a  $\alpha$ , valor predeterminado antes do início da coleta de dados. Esse  $\alpha$  é o mesmo  $\alpha$  que representa a probabilidade de cometer um erro ao estabelecer um intervalo de confiança. Faz algum sentido que essas duas probabilidades sejam iguais, pois a probabilidade de rejeitar a  $H_0$  quando esta é de fato verdadeira (erro do Tipo I) é igual à probabilidade de que o verdadeiro parâmetro populacional fique de fora do conjunto para seus possíveis valores, quando não deveria. Essa probabilidade é representada por  $\alpha$ .

Suponha que uma transportadora afirme que o tempo médio de suas entregas é de 3,0 dias (portanto,  $H_0$  é  $\mu = 3,0$ ), porém, você acredita que a média é diferente desse valor (então,  $H_a$  é  $\mu \neq 3,0$ ). Seu nível  $\alpha$  é 0,05 e, já que este é um teste bilateral, você tem 0,025 em cada lado. Sua amostra com 100 encomendas apresenta uma média de 3,5 dias com desvio

padrão de 1,5 dias. A estatística de teste é igual a  $\frac{3,5 - 3,0}{\frac{1,5}{\sqrt{100}}} = 3,33$ , maior do que 1,96 (valor encontrado no cruzamento entre a última linha e a coluna 0,025 da tabela da distribuição- $t$  — veja a Tabela A-1 no apêndice). Assim, 3,0 não é um possível valor para o tempo médio de entrega de todas as encomendas, e você rejeita  $H_0$ .

Mas suponha que, simplesmente por obra do acaso, sua amostra continha alguns tempos de entrega mais longos do que os normais e que, na realidade, a afirmação da empresa estava correta. Você simplesmente cometeu um erro do Tipo I. Fez um alarme falso sobre a afirmação da empresa.

Para reduzir as chances de um erro Tipo I, reduza o valor de  $\alpha$ . Entretanto, não recomendo uma redução brusca. O lado positivo é que essa redução dificulta a rejeição de  $H_0$ , pois, para isso, seus dados vão precisar de mais evidências. O lado negativo é que, ao reduzir sua probabilidade para um alarme falso (erro Tipo I), você aumenta as chances de perder uma oportunidade (erro Tipo II).

## ***Perdendo oportunidades em virtude de um erro Tipo II***

O *erro Tipo II* é a probabilidade condicional de não rejeitar a  $H_0$  quando ela é realmente falsa. Eu o chamo de oportunidade perdida, pois o esperado era que você encontrasse algo de errado com a  $H_0$  e a rejeitasse, mas não foi o que aconteceu. Você não apertou o botão quando deveria.

A probabilidade de cometer um erro do Tipo II depende de uma série de fatores:

- ✓ *Tamanho amostral*: Quanto mais dados tiver, menores são suas chances de perder essa oportunidade. Por exemplo, se uma moeda realmente estiver viciada, lançá-la apenas dez vezes pode não revelar o problema. Porém, se lançá-la mil vezes, terá uma boa chance de observar um padrão que favorece a face cara em detrimento da coroa, ou vice-versa.
- ✓ *O valor real do parâmetro*: O erro do Tipo II também está relacionado ao tamanho do problema que você está tentando resolver. Por exemplo, suponha que uma transportadora afirme que o tempo médio de suas entregas é de 3,5 dias. Se o tempo real de entrega for de 5,0 dias, você não terá muito trabalho para descobrir isso usando sua amostra (mesmo que ela seja pequena). Mas se o tempo real de entrega for de 4,0 dias, você precisará se esforçar mais para realmente detectar o problema.



Para reduzir as chances de um erro Tipo II, utilize uma amostra grande, pois facilita a rejeição da  $H_0$ , embora aumente a chance de um erro Tipo I.



Os erros Tipo I e Tipo II ocupam os extremos de uma gangorra — conforme um sobe, o outro desce. Tente encontrar o equilíbrio selecionando uma amostra grande (quanto maior, melhor; veja as Figuras 3-1 e 3-2) e um nível  $\sigma$  baixo (0,05 ou menos) para seu teste de hipótese.

## ***O poder de um teste de hipótese***

Os erros do Tipo II, os quais explico na seção anterior, mostram a desvantagem de um teste de hipótese. Mas os estatísticos, ao contrário do que muita gente pensa, realmente tentam ver o lado bom das coisas de vez em quando; portanto, em vez de ver a probabilidade de *perder* uma diferença da  $H_0$  que de fato está ali, eles veem a probabilidade de *detectar* uma diferença que realmente existe. Essa detecção é o *poder de um teste de hipótese*.



O poder de um teste de hipótese é 1 — a probabilidade de cometer um erro Tipo II. Sendo assim, o *poder* é um número entre 0 e 1 que representa a probabilidade de rejeitar a  $H_0$  quando esta é falsa. (Isso dá até letra de música: "Se você sabe que a  $H_0$  é falsa, bata palmas...") Lembre-se de que esse poder (assim como o erro Tipo II) depende de dois elementos: o tamanho amostral e o valor real do parâmetro (veja a descrição desses elementos na seção anterior).

Na seções a seguir, você descobrirá o que significa poder em estatística (mas sem ser mandão); além disso, também ficará sabendo como quantificá-lo através de uma curva de poder.

## ***Traçando uma curva de poder***

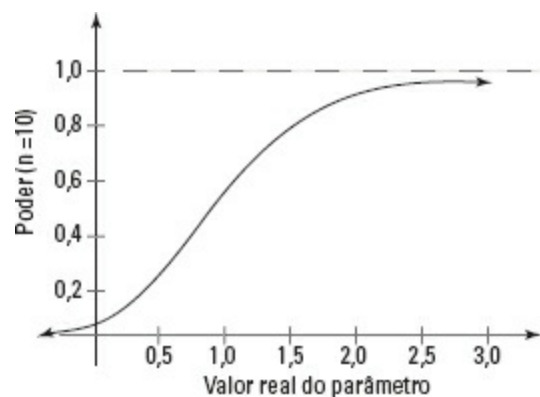
Os cálculos específicos para o poder de um teste de hipótese estão além do escopo deste livro (portanto, pode respirar aliviado), mas existem programas de computador e gráficos disponíveis na internet que lhe mostram qual é o poder para os diferentes testes de hipóteses e vários tamanhos amostrais (basta digitar "curva de poder para o teste [blá-blá-blá]" em um site de busca).

Esses gráficos são chamados de *curvas de poder* para um teste de hipótese. A curva de poder é um tipo especial de gráfico que lhe dá uma ideia do quanto de uma diferença da  $H_0$  você conseguirá detectar com o tamanho de sua amostra. Uma vez que a precisão de sua estatística de teste aumenta à medida que o tamanho de sua amostra também aumenta, o tamanho amostral está diretamente relacionado ao poder. Entretanto, ele também depende do quanto da diferença da  $H_0$  você está tentando detectar. Por exemplo, se uma empresa de entregas afirma que suas encomendas chegam em 2 dias ou menos, você acionaria o alarme se, na verdade, elas chegam em 2,1 dias? Ou esperaria até 3 dias? É preciso ter uma amostra muito maior para detectar a situação de 2,1 dias versus a de 3 dias só por causa do nível de precisão necessário.

Na Figura 3-1, vemos a curva de poder para um determinado teste da  $H_0: \mu = 0$  versus  $H_a: \mu > 0$ . Podemos assumir que  $\sigma$  (o desvio padrão da população) é igual a 2 (vou lhe dar esse valor a cada problema) e não se altera. Ao longo do livro, estabeleço o tamanho amostral em 10.

O eixo horizontal ( $x$ ) na curva de poder mostra um conjunto de valores reais de  $\mu$ . Por exemplo, você formula a hipótese de que  $\mu$  é igual a 0, mas, na verdade, ele pode ser 0,5; 1,0; 2,0; 3,0; ou qualquer outro valor possível. Se  $\mu$  é igual a 0, então  $H_0$  é verdadeira. A chance de detectar isso — e, portanto, rejeitar  $H_0$  — é igual a 0,05, o valor estabelecido para  $\alpha$ . Trabalhe a partir dessa linha de base. (Observe que faz sentido possuir um poder baixo nessa situação, pois não há nada para detectar os valores de  $\mu$  que estejam próximos de 0.) Sendo assim, no gráfico da Figura 3-1, quando  $x = 0$ , o valor de  $y$  é 0,05.





---

**Figura 3-1:** Curva de poder para  $H_0: \mu = 0$  versus  $H_a: \mu > 0$ , para  $n = 10$  e  $\sigma = 2$ .

---

Suponha que  $\mu$  seja realmente 0,5, e não 0 como em sua hipótese. O computador lhe diz que a probabilidade de rejeitar  $H_0$  (ação que você deve tomar aqui) é  $0,197 = 0,20$ , que é o poder. Então, você tem cerca de 20% de chance de detectar essa diferença com um tamanho amostral igual a 10. Conforme você se movimentar para a direita, afastando-se de 0 no eixo horizontal ( $x$ ), poderá ver que o poder aumenta e os valores de  $y$  aproximam-se cada vez mais de 1,0.

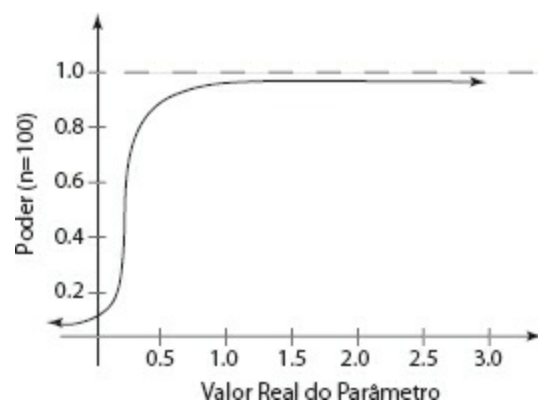
Por exemplo, se o valor real de  $\mu$  for 1,0, a diferença a partir de 0 é mais fácil de ser detectada do que se  $\mu$  fosse 0,50. Na verdade, o poder em 1,0 é igual a  $0,475 = 0,48$ , portanto, você tem quase 50% de chance de pegar a diferença da  $H_0$  nesse caso. E, à medida que os valores da média aumentam, o poder se aproxima cada vez mais de 1,0. O poder nunca alcança 1,0; pois a estatística nunca consegue provar nada com 100% de certeza, mas é possível chegar bem próximo de 1,0 se o valor real estiver longe o bastante de sua hipótese.

### ***Controlando o tamanho amostral***

Como aumentar o poder de seu teste de hipótese? Você não tem controle algum sobre o valor real de um parâmetro, pois este é um número desconhecido. Então, o que é possível controlar? O tamanho amostral. À medida que o tamanho amostral aumenta, torna-se mais fácil detectar uma real diferença na  $H_0$ .

A Figura 3-2 mostra a curva de poder com os mesmos números da Figura 3-1, exceto o tamanho amostral ( $n$ ), que é 100, em vez de 10. Note que a curva aumenta muito mais rapidamente, se comparada à sua hipótese de média igual a 0, e se aproxima de 1,0 quando a média real é 1,0. Você deverá observar essa curva que ascende rapidamente em direção ao valor 1,0; à medida que os valores reais do parâmetro aumentam no eixo  $x$ .





**Figura 3-2:** Curva de poder para  $H_0: \mu = 0$  versus  $H_a: \mu > 0$ , para  $n = 100$  e  $\sigma = 2$ .

Se compararmos o poder do teste quando  $\mu$  é 1,0 para  $n = 10$  (na Figura 3-1) versus  $n = 100$  (na Figura 3-2), veremos que o poder aumenta de 0,475 para mais de 0,999. A Tabela 3-1 mostra diferentes valores de poder para  $n = 10$  versus  $n = 100$ , quando testamos  $H_0: \mu = 0$  versus  $H_a: \mu > 0$ , assumindo um valor  $\sigma = 2$ .

**Tabela 3-1** Comparando Valores de Poder para  $n = 10$  Versus  $n = 100$  ( $H_0$  é  $\mu = 0$ )

<i>Valor Real de <math>\mu</math></i>	<i>Poder Quando <math>n = 10</math></i>	<i>Poder Quando <math>n = 100</math></i>
0,00	0,050 = 0,05	0,050=0,05
0,50	0,197 = 0,20	0,804=0,81
1,00	0,475 = 0,48	aprox. 1,0
1,50	0,766 = 0,77	aprox. 1,0
2,00	0,935 = 0,94	aprox. 1,0
3,00	0,999 = aprox. 1,0	aprox. 1,0

É possível encontrar curvas de poder para uma variedade de testes de hipótese sob várias situações diferentes. Todas possuem a mesma aparência: começam no valor de  $\alpha$  quando  $H_0$  é verdadeira, tomam a forma de S à medida que você se move da esquerda para a direita no eixo  $x$  e, por fim, aproximam-se do valor 1,0 em algum ponto. Curvas de poder para amostras grandes se aproximam de 1,0 mais rapidamente do que curvas para amostras pequenas.

Muito poder pode ser ruim. Por exemplo, se você construir a curva de poder para  $n = 10.000$  e compará-la às Figuras 3-1 e 3-2, verá que ela praticamente alcança o valor 1,0 para qualquer média diferente de 0,0. Ou seja, a média real poderia ser 0,05 e, com a hipótese  $H_0: \mu = 0,00$ , você rejeitaria  $H_0$  em virtude desse enorme tamanho amostral. A menos que se queira detectar diferenças muito pequenas na  $H_0$  (como em estudos médicos ou em situações de controle de qualidade), os valores muito grandes para  $n$  são, de modo geral, vistos com suspeita. Às vezes, as pessoas aumentam o  $n$  apenas para dizerem que encontraram uma diferença, mesmo que ela seja muito pequena, portanto, fique atento.

Quando ampliamos muito alguma coisa, sempre conseguimos detectar algo, mesmo que isso não tenha nenhuma diferença prática. Tome cuidado com pesquisas e experimentos com tamanho amostral muito grande, como os na casa de dezenas de milhares. Esses resultados com certeza estarão inflados.

## **O poder na indústria e na Medicina**

O poder de um teste tem seu papel no processo industrial. As indústrias muitas vezes possuem especificações bastante rigorosas com relação ao tamanho, peso e/ou à qualidade de seus produtos. Durante o processo de fabricação, as indústrias querem ser capazes de detectar os desvios dessas especificações, mesmo os pequenos, e, portanto, devem determinar a diferença da  $H_0$  que desejam detectar para, assim, calcular o tamanho amostral necessário. Por exemplo, se uma barra de chocolate deve pesar 56 gramas, o fabricante pode acionar o alarme quando o peso real médio se alterar para 62 gramas. Os estatísticos conseguem trabalhar de trás para frente, partindo do cálculo do poder para descobrir o tamanho amostral de que precisam para saber quando parar o processo. Os médicos pesquisadores também pensam no poder quando têm que planejar seus estudos (chamados de ensaios clínicos).

Suponha que eles estejam verificando se um antidepressivo afeta de forma adversa a pressão arterial (como um efeito colateral). Esses cientistas têm que ser capazes de detectar as mais sutis diferenças na pressão, uma vez que, para alguns pacientes, qualquer alteração da pressão arterial deve ser percebida e tratada.

## Parte II:

# Usando Diferentes Tipos de Regressão para Fazer Previsões





## *Nesta parte...*

***E***sta parte leva você além do uso de uma variável para a previsão de outra utilizando uma reta (ou seja, a regressão linear simples). Em vez disso, você vai descobrir formas de prever uma variável usando várias outras, além de formas de fazer uma previsão através de curvas. Por fim, fará previsões para probabilidade, e não apenas para médias. Aqui é o lugar onde você vai encontrar tudo sobre regressão. Uma vez que esses métodos lhe permitem resolver problemas mais complexos, prestam-se muito bem a muitas aplicações do mundo real.

# Capítulo 4

## Em Linha com a Regressão Linear Simples

---

### *Neste Capítulo*

- ▶ Usando diagramas de dispersão e coeficientes de correlação para examinar relações
  - ▶ Construindo um modelo de regressão linear simples para estimar  $y$  a partir de  $x$
  - ▶ Testando a adequação do modelo
  - ▶ Interpretando os resultados e fazendo boas previsões
- 

**A** busca por relações e a realização de previsões é um dos princípios básicos da análise de dados. Todos querem responder perguntas como: “Posso prever quantas unidades vou vender se gastar  $x$  com propagandas?” ou “É verdade que beber refrigerante diet em demasia engorda?” ou ainda: “As mochilas escolares realmente ficam mais pesadas a cada ano ou é só impressão minha?”

A *regressão linear* tenta encontrar relações entre duas ou mais variáveis e nos fornece um modelo que tenta descrever essa relação de forma muito parecida como a equação de reta ( $y = 2x + 3$ ) explica a relação entre  $x$  e  $y$ . Mas, diferente da Matemática, em que funções como  $y = 2x + 3$  dizem tudo sobre as duas variáveis em questão, em Estatística, as coisas não são assim tão perfeitas; a variabilidade e o erro sempre estarão envolvidos (mas é isso que a torna tão divertida!).

Este capítulo é, em parte, uma revisão dos conceitos de regressão linear simples abrangidos por qualquer livro didático de Estatística I. Mas a diversão não para por aqui. Vou expandir as ideias sobre regressão vistas no curso de Estatística I e prepará-lo para alguns outros tipos de modelos de regressão que serão vistos dos Capítulos de 5 a 8.

Neste capítulo, você verá como construir um modelo de regressão linear simples que examine a relação entre duas variáveis. Além disso, você também verá como a regressão linear simples funciona a partir da perspectiva da construção de um modelo.

# *Investigando Relações com Diagramas de Dispersão e Correlações*

Antes de tentar prever o valor de  $y$  a partir de  $x$  usando uma reta, é preciso:

- ✓ Estabelecer uma razão legítima para fazê-lo com o uso de uma reta.
- ✓ Ter certeza de que o uso de uma reta para esse fim realmente vai funcionar.

Para cumprir esses dois passos importantes, você vai precisar, primeiro, fazer um gráfico dos dados em pares para que, assim, possa visualizar uma possível relação; em seguida, terá que, de algum modo, quantificar a relação através da forma com que esses pontos acompanham a reta. Nesta seção, você vai fazer justamente isso, usando diagramas de dispersão e correlações.

Aqui vai um exemplo perfeito de uma situação em que a regressão linear simples é usada: em 2004, a Secretaria de Educação do Estado da Califórnia escreveu um relatório intitulado “O peso do livro didático na Califórnia: análise e recomendações”. Esse relatório discutia a grande preocupação a respeito do peso dos livros didáticos nas mochilas escolares e os problemas que isso poderia causar aos alunos. A secretaria conduziu um estudo em que foi pesada uma variedade de livros de cada uma das quatro principais áreas estudadas nas séries de 1 a 12 (Literatura, Matemática, Ciências e História — e Estatística?) de algumas editoras. Com isso, descobriu-se o peso total médio para os quatro livros de cada série.

Também foram consultados pediatras e ortopedistas, que recomendaram que o peso de uma mochila escolar não ultrapassasse 15% do peso do aluno. A partir daí, a secretaria formulou a hipótese de que o peso total dos livros didáticos dessas quatro áreas aumenta a cada série e quis saber se conseguiria encontrar uma relação entre o peso médio de uma criança em cada série e o peso médio de seus livros. Assim, com o peso médio dos livros das quatro principais áreas para cada série, os pesquisadores também registraram o peso médio dos alunos dessas séries. Os resultados estão na Tabela 4-1.

**Tabela 4-1      Peso Médio do Livro Didático e Peso Médio do Aluno (Séries 1 a 12)**

<i>Série</i>	<i>Peso Médio do Aluno (Em Quilos)</i>	<i>Peso Médio do Livro (Em Quilos)</i>
1	22,00	3,63
2	24,72	4,28
3	27,78	4,57
4	31,30	5,36
5	33,79	5,57
6	38,56	6,17
7	40,37	6,86
8	44,91	7,02

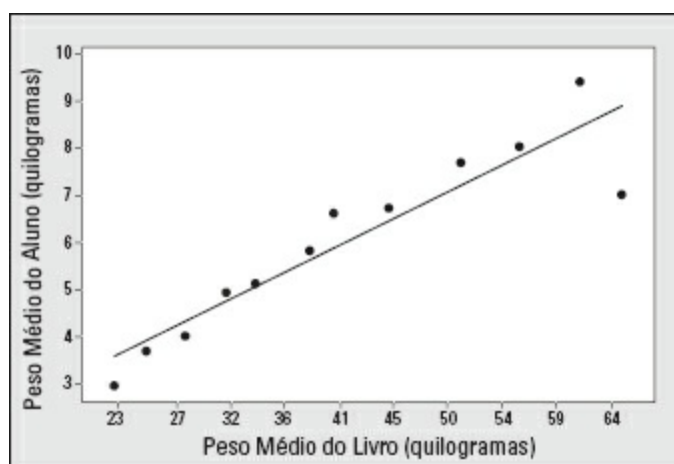
9	50,80	7,87
10	55,79	8,20
11	60,78	9,43
12	64,41	7,28

Nesta seção, você vai começar a investigar se existe ou não uma relação entre essas duas variáveis quantitativas. Comece dispondo os pares dos dados em um diagrama de dispersão de duas dimensões a fim de buscar um possível padrão e quantifique a força e a direção desse padrão por meio do coeficiente de correlação.

## Usando diagramas de dispersão para investigar relações

Para investigar uma possível relação entre duas variáveis, tais como o peso dos livros e o peso do aluno, primeiro, coloque os dados em um tipo especial de gráfico denominado *diagrama de dispersão*. O diagrama de dispersão é um gráfico bidimensional que exhibe pares de dados, um par por observação no formato  $(x, y)$ . A Figura 4-1 ilustra um diagrama de dispersão para os dados referentes ao peso do livro didático presentes na Tabela 4-1.

É possível notar que a relação parece acompanhar a reta incluída no gráfico, exceto, talvez, pelo último ponto, em que o peso do livro didático é 7,28 quilogramas e o peso do aluno é de 64,41 quilogramas (para a série 12). Esse ponto parece ser um *outlier* (valor discrepante) — o único ponto que não segue o padrão. Ainda assim, parece haver uma relação linear ascendente (ou *positiva*) entre o peso do livro didático e o peso do aluno; à medida que o peso do aluno aumenta, também aumenta o peso dos livros.



**Figura 4-1:** Diagrama de dispersão do peso médio do aluno versus peso médio do livro didático nas séries de 1 a 12.

Para fazer um diagrama de dispersão no Minitab, insira os dados nas colunas um e dois da planilha. Selecione no menu Graphs>Scatterplot. Clique em Simple e, depois, em OK. Selecione response variable ( $y$ ) na caixa à esquerda e clique em Select. Essa variável vai aparecer como a variável  $y$  no diagrama. Clique em explanatory variable ( $x$ ) na caixa à



esquerda e clique em Select. Ela vai aparecer como a variável  $x$  no diagrama. Clique em OK, e seu diagrama estará pronto.

## ***Comparando informações através do coeficiente de correlação***

Depois de colocar os dados em um diagrama de dispersão (veja a seção anterior), o próximo passo é encontrar uma estatística que, de alguma forma, quantifique a relação. O *coeficiente de correlação* (também conhecido como *coeficiente de correlação Pearson*, especialmente nos softwares estatísticos) mede a força e a direção da relação linear entre duas variáveis quantitativas  $x$  e  $y$ . É um número entre  $-1$  e  $+1$ , que *não possui unidade*, o que significa que, se você mudar de quilogramas para gramas, o coeficiente de correlação não vai se alterar. (Imagine a loucura que seria se não fosse assim!)

Se a relação entre  $x$  e  $y$  é ascendente ou positiva (à medida que  $x$  aumenta,  $y$  também faz o mesmo), a correlação é um número positivo. Se a relação é descendente ou negativa (à medida que  $x$  aumenta,  $y$  diminui), a correlação é um número negativo. A lista a seguir traduz os diferentes valores de correlação:

- ✓ **Uma correlação igual a zero indica a ausência de uma relação linear entre  $x$  e  $y$ .** (Pode ser que exista uma relação diferente, tal como uma curva; veja o Capítulo 7 para mais informações a respeito.)
- ✓ **Uma correlação igual a  $+1$  ou  $-1$  indica que os pontos se dispõem em uma reta perfeita.** (Valores negativos indicam uma relação descendente; valores positivos indicam uma relação ascendente.)
- ✓ **Uma correlação próxima a  $+1$  ou  $-1$  indica uma relação forte.** A regra diz que as correlações próximas a  $0,7$  ou  $-0,7$  ou maiores do que isso são consideradas fortes.
- ✓ **Uma correlação próxima a  $+0,5$  ou  $-0,5$  indica uma relação moderada.**

O coeficiente de correlação pode ser calculado por meio de uma fórmula que envolve o desvio padrão de  $x$  e  $y$ , além da covariância de  $x$  e  $y$ , a qual mede como  $x$  e  $y$  se movimentam juntos em relação às suas médias. Entretanto, o foco aqui não é a fórmula (você pode encontrá-la em qualquer livro didático de Estatística I ou no meu livro *Estatística Para Leigos*, publicado pela Alta Books); é o conceito que é importante. Qualquer software consegue calcular o coeficiente de correlação com apenas um clique.

Para que o Minitab calcule a correlação para você, faça o caminho Stat>Basic Statistics>Correlation. Selecione as variáveis para as quais quer a correlação e clique em Select. Depois, clique em OK.

A correlação para o exemplo do peso do livro didático é (tente adivinhar antes de olhar a resposta)  $0,926$ , valor muito próximo a  $1,0$ . Essa correlação indica a presença de uma relação linear muito forte entre o peso médio do livro didático e o peso médio dos alunos das séries de 1 a 12, além de nos informar que se trata de uma relação positiva e linear (ou seja, ela acompanha uma reta). Essa correlação é confirmada pelo diagrama de dispersão



mostrado na Figura 4-1.



Os analistas de dados nunca devem tirar nenhuma conclusão acerca de uma relação entre  $x$  e  $y$  baseados apenas na correlação ou no diagrama de dispersão; os dois elementos devem ser analisados juntos. É possível (mas, obviamente, não é uma boa ideia) manipular os gráficos para que pareçam melhores ou piores do que realmente são; para isso, basta alterar as escalas dos eixos. Por essa razão, os estatísticos nunca confiam somente no diagrama de dispersão para determinar a existência ou não de uma relação entre  $x$  e  $y$ . Usar a correlação sem um diagrama de dispersão também é perigoso, pois a relação entre  $x$  e  $y$  pode ser muito forte, mas não linear.

# ***Construindo um Modelo de Regressão Linear Simples***

Depois de saber que a variável  $x$  pode se relacionar com  $y$  de forma linear, é hora de encontrar a reta que melhor representa essa relação. Encontre o coeficiente angular e o intercepto  $y$  (ponto onde a reta corta o eixo  $y$ ), junte-os para formar uma reta e utilize a equação desta para fazer previsões para o  $y$ . Tudo isso faz parte da construção de um modelo de regressão linear simples.

Nesta seção, estabeleço os alicerces para a construção dos modelos de regressão em geral (incluindo os que você vai encontrar nos Capítulos de 5 a 8). Coloque os dados em um gráfico, consiga um modelo que faça sentido, avalie sua adequação aos dados e utilize-o para estimar o valor de  $y$  a partir de um valor dado a  $x$ .

## ***Encontrando a reta certa para modelar seus dados***

Depois de estabelecer que  $x$  e  $y$  possuem uma forte relação linear, evidenciada tanto pelo diagrama de dispersão quanto pelo coeficiente de correlação (próximo a 0,7 e -0,7 ou maior; veja as seções anteriores), você está pronto para construir um modelo que estime  $y$  a partir de  $x$ . No caso do peso dos livros didáticos, vamos estimar o peso médio dos livros usando o peso médio dos alunos.

O modelo de regressão mais básico de todos é o *modelo de regressão linear simples* cuja fórmula genérica é  $y = \alpha + \beta x + \varepsilon$ . Aqui,  $\alpha$  representa o intercepto  $y$  da reta,  $\beta$  representa o coeficiente angular e  $\varepsilon$  representa o erro casual presente no modelo.



A reta usada em uma regressão linear simples é apenas uma de uma família inteira de modelos (ou funções) que os estatísticos usam para expressar as relações entre variáveis. *Modelo* é somente um nome genérico dado a uma função usada para descrever um possível resultado com base em algumas informações sobre uma ou mais variáveis relacionadas.

Note que você nunca irá conhecer o verdadeiro modelo que descreve perfeitamente a relação. O melhor que pode fazer é estimá-lo com base nos dados.

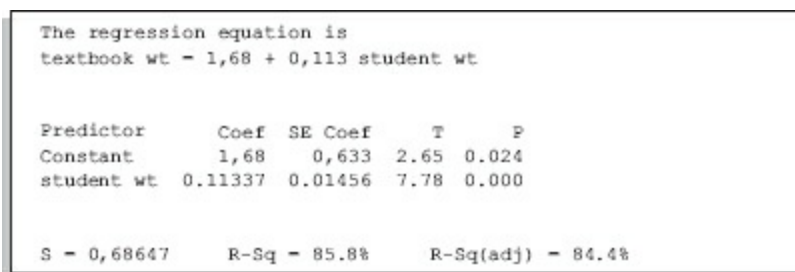
Para encontrar o modelo mais adequado a seus dados, o melhor é investigar todas as possíveis retas e escolher a que melhor se adequa aos dados. Felizmente, existe um algoritmo que faz isso por você (os computadores utilizam-no em seus cálculos). Também existem fórmulas para calcular à mão o coeficiente angular e o intercepto  $y$  da reta de regressão. A reta de regressão baseada em seus dados é  $y = a + bx$ , onde  $a$  estima o  $\alpha$  e  $b$  estima o  $\beta$  do modelo verdadeiro. (Essas fórmulas podem ser encontradas em qualquer livro didático de Estatística I ou no meu livro *Estatística Para Leigos*.)



Para conduzir uma análise de regressão no Minitab, faça o caminho Stat>Regression>Regression. Selecione sua variável de resposta ( $y$ ) na caixa à esquerda e clique em Select. Essa variável vai aparecer na caixa Response Variable. Em seguida, selecione sua variável explicativa ( $x$ ) e clique em Select. Essa variável vai aparecer na caixa Predictor Variable.

Por último, clique em OK.

A equação da reta que melhor representa a relação entre o peso médio dos livros e dos alunos é  $y = 1,68 + 0,113x$ , onde  $x$  é o peso médio dos alunos para uma determinada série, e  $y$ , o peso médio dos livros. A Figura 4-2 mostra o resultado da análise do Minitab.



The regression equation is textbook wt = 1,68 + 0,113 student wt				
Predictor	Coef	SE Coef	T	P
Constant	1,68	0,633	2.65	0.024
student wt	0.11337	0.01456	7.78	0.000
S = 0,68647      R-Sq = 85.8%      R-Sq(adj) = 84.4%				

**Figura 4-2:** Análise de regressão linear simples para o exemplo dos livros didáticos



Ao escrever  $y = 1,68 + 0,113x$ , você diz que essa equação representa seu valor estimado para  $y$ , dado o valor de  $x$  observado em seus dados. Os estatísticos escrevem essa equação usando um circunflexo, (ou chapéu, como é mais conhecido), como  $\bar{y}$  para que todos saibam que isso é só uma estimativa, e não o real valor de  $y$ . Esse  $y$  com chapéu é a sua estimativa para o valor médio de  $y$  a longo prazo, baseada nos valores observados para  $x$ . Entretanto, em muitos livros de Estatística I, o circunflexo é omitido, pois os estatísticos já entendem  $y$  dessa forma. Essa questão ainda será abordada novamente nos Capítulos de 5 a 8. A propósito, se você acha engraçado falar  $y$  com chapéu, imagina então no México, onde os estatísticos o chamam de  $y$  *con sombrero* — e não estou brincando!

## ***O intercepto y da reta de regressão***

Algumas partes daquele resultado do Minitab, mostrado na Figura 4-2, serão importantes nesse momento. Primeiro, você pode ver que, embaixo da coluna Coef, estão os valores numéricos do lado direito da equação da reta — ou seja, o coeficiente angular e o intercepto  $y$ . O número 1,68 representa o coeficiente de “Constant”, uma forma elegante de chamar o intercepto  $y$  (pois o intercepto  $y$  não passa de uma constante — ela nunca se altera). O intercepto  $y$  é o ponto onde a reta corta o eixo  $y$ ; ou seja, é o valor de  $y$  quando  $x$  é igual a 0.



O intercepto  $y$  de uma reta de regressão pode ou não ter um significado prático, dependendo da situação. Para saber se o intercepto  $y$  de uma reta de regressão possui ou não um significado prático, observe o seguinte:

- ✓ O intercepto  $y$  se encontra dentro dos valores reais do conjunto de dados? Se a resposta for sim, ele possui um significado prático.
- ✓ O intercepto  $y$  cai na área negativa quando os valores negativos de  $y$  não são possíveis? Por exemplo, se os valores de  $y$  representam pesos, eles não podem ser

negativos. Então o intercepto  $y$  não possui significado prático. No entanto, mesmo assim, precisamos dele na equação, só porque ele representa o local onde a reta, se estendida até o eixo  $y$ , cruza-o.

- ✓ O valor  $x = 0$  possui significado prático? Por exemplo, se  $x$  for a temperatura em uma partida de futebol realizada em Green Bay, então  $x = 0$  é o valor relevante. Se  $x = 0$  tiver significado prático, então o intercepto  $y$  também terá, pois representa o valor de  $y$  quando  $x = 0$ . Se  $x = 0$  não tiver significado prático por si só (por exemplo, quando  $x$  representa a altura de uma criança), então o intercepto  $y$  também não o terá.

No exemplo do livro didático, o intercepto  $y$  não tem significado prático, pois nenhum aluno pesa zero quilos, assim, você não precisa se importar com a estimativa do peso dos livros nessa situação. Mas precisa encontrar a reta que se ajusta a seus dados (onde os pesos médios vão de 22 a 64 quilos). E essa reta de regressão deve incluir o intercepto  $y$  que, para esse problema, é 1,68 quilos.

## ***O coeficiente angular da reta de regressão***

O valor 0,113 da Figura 4-2 indica o coeficiente (ou o número na frente) da variável que representa o peso dos alunos. Esse número também é conhecido como *coeficiente angular*. Ele indica que a mudança em  $y$  (peso dos livros) está associada ao aumento de uma unidade em  $x$  (peso dos alunos). À medida que o peso dos alunos aumenta em 1 quilo, o peso dos livros aumenta em aproximadamente 0,113 quilo em média. Para tornar essa relação mais significativa, multiplique os dois valores por 10 para dizer que, à medida que o peso do aluno aumenta em 10 quilos, o peso dos livros sobe, em média, aproximadamente 1,13 quilos.

Sempre que obtiver o coeficiente angular, coloque-o sobre 1, para ajudar em sua interpretação. Por exemplo, um coeficiente igual a 0,113 será reescrito como  $\frac{0,113}{1}$ . Usando a ideia de que o coeficiente angular é igual à mudança em  $y$  sobre a mudança em  $x$ , você vai interpretar o valor de 0,113 da seguinte forma: à medida que  $x$  aumenta, em média, 1 quilo,  $y$  aumenta 0,113 quilo.

## ***Estimando pontos através da regressão linear***

Quando você tem uma reta que estima  $y$  através de  $x$ , pode usá-la para fazer uma estimativa de um único número para o valor (médio) de  $y$  para um dado valor de  $x$ . Esse processo se chama *estimativa de ponto*. A ideia básica é pegar um valor de  $x$  razoável, aplicá-lo na equação da reta de regressão e ver o que você obtém para o valor de  $y$ .

No exemplo do peso dos livros, a reta de regressão (ou modelo) é a reta  $y = 1,68 + 0,113x$ . Para um aluno que pese 27 quilos, por exemplo, a estimativa de ponto do peso médio dos livros é  $1,68 + (0,113 * 27) = 4,73$  quilos (coitadinho!). Se o aluno pesar 45 quilos, o peso



médio estimado para os livros é  $1,68 + (0,113 * 45) = 6,77$  quilos, ou quase 7 quilos, mais ou menos alguma coisa. (Vamos descobrir o que é esse mais ou menos alguma coisa na seção a seguir.)

# ***Sem Deixar Nenhuma Conclusão para Trás: Testes e Intervalos de Confiança para a Regressão***

Depois de encontrar o coeficiente angular da reta de regressão para seus dados (veja as seções anteriores), você precisa dar um passo atrás e levar em consideração o fato de que os resultados amostrais variam. Você não deve simplesmente dizer, “Ok, o coeficiente angular dessa reta é 2. Acabei!” Ele não vai ser exatamente 2 da próxima vez. Essa variabilidade é o motivo pelo qual os professores de estatísticas ficam batendo na mesma tecla sobre a adição de uma margem de erro em seus resultados amostrais; você deve se garantir através da adição daquele mais ou menos alguma coisa.

No teste de hipótese, o que você faz não é simplesmente comparar sua média amostral com a média populacional e dizer: “Opa, elas são diferentes!” É preciso padronizar seu resultado amostral usando o erro padrão para que você possa colocar seus resultados na perspectiva correta (veja no Capítulo 3 a revisão sobre intervalos de confiança e testes de hipótese).

A mesma ideia é aplicada à regressão. Os dados foram usados para descobrir a reta de regressão, e você já sabe que ela se adequa bem a esses dados. Nem preciso dizer que a reta de regressão vai funcionar perfeitamente bem para o novo conjunto de dados retirado da mesma população. Portanto, em regressão, todos os seus resultados devem envolver o erro padrão a fim de considerar o fato de que os resultados amostrais variam. Isso também é válido para estimativa e teste do coeficiente angular e do intercepto  $y$  e de qualquer outra previsão que você fizer.

Muitas vezes, nos cursos de Estatística I, o conceito da margem de erro é esquecido depois que a reta de regressão é encontrada, mas esse é um conceito muito importante e deve sempre ser incluído. (Ok, chega de falação. Vamos colocar tudo isso em prática!)

## ***Analizando o coeficiente angular***

Recorde que o *coeficiente angular* da reta de regressão é a quantidade pela qual você espera que a variável  $y$  se altere, em média, à medida que a variável  $x$  aumenta em 1 unidade. (Veja a seção “O coeficiente angular da reta de regressão” mais no início deste capítulo.) Agora, como lidar com a ideia de saber que a reta de regressão irá mudar com o novo conjunto de dados? Basta aplicar as ideias básicas dos intervalos de confiança e dos testes de hipótese (veja o Capítulo 3).

### ***Um intervalo de confiança para o coeficiente angular***

O *intervalo de confiança*, de modo geral, possui esta fórmula: sua estatística mais ou menos uma margem de erro. A margem de erro inclui um determinado número de desvios padrão (ou erros padrão) a partir de sua estatística. O número de erros padrão adicionados e subtraídos depende do nível de confiança,  $1 - \alpha$ , desejado. O tamanho do erro padrão

depende do tamanho amostral e de outros fatores.

A equação da reta de regressão linear simples,  $y = a + bx$ , inclui um coeficiente angular ( $b$ ) e uma intercepção  $y$  ( $a$ ). Uma vez que estes foram encontrados através dos dados, são apenas estimativas do que realmente acontece com a população e, portanto, devem ser acompanhados por uma margem de erro.

A fórmula para um intervalo de confiança de nível  $1 - \alpha$  para o coeficiente angular de uma

reta de regressão é  $b \pm t_{n-2}^* SE_b$ , onde o erro padrão é denotado por  $SE_b = \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}$ , onde  $s = \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2}$ . O valor de  $t^*$  vem da distribuição- $t$  com  $n - 2$  graus de liberdade e área à sua direita igual a  $\alpha \div 2$ . (Consulte o Capítulo 3 para informações referentes ao conceito de  $\alpha$ .)

Caso queira saber por que você vê  $n - 2$  graus de liberdade aqui, em oposição a  $n - 1$  graus de liberdade usado nos testes  $t$  para a média populacional em Estatística I, eis uma explicação: em Estatística I, você aprendeu que um *parâmetro* é um número que descreve a população; geralmente, é conhecido e pode mudar de uma situação para outra. Para cada parâmetro em um modelo, você perde 1 grau de liberdade. A reta de regressão contém dois parâmetros — o coeficiente e o intercepto  $y$  — e, portanto, você vai perder 1 grau de liberdade para cada um. Com o teste- $t$  da Estatística I, você tem que se preocupar apenas com um parâmetro, a média populacional, por isso, utiliza  $n - 1$  graus de liberdade.

O valor  $t^*$  pode ser encontrado em qualquer tabela da distribuição- $t$  (consulte a do seu livro didático). Agora, suponha que você queira encontrar um intervalo de confiança de 95%, com base em um tamanho amostral de  $n = 10$ . O valor de  $t^*$  está na Tabela A-1 no apêndice, na linha  $10 - 2 = 8$  graus de liberdade, na coluna 0,025 (pois  $\alpha \div 2 = 0,05 \div 2 = 0,025$ ). O valor de  $t^*$  é 2,306. (Em meu livro *Estatística Para Leigos*, você encontra muito mais detalhes sobre a distribuição- $t$  e a tabela- $t$ .)

Para montar um intervalo de confiança de 95% para a estimativa do coeficiente angular usando o computador, você deve encontrar as peças necessárias. Para o exemplo dos livros didáticos, na Figura 4-2 vemos que o coeficiente angular é 0,113. (Lembre-se de que o coeficiente angular é o número que vem na frente da variável  $x$  na equação e, por isso, vemos a abreviação Coef no resultado do Minitab.)

Uma vez que o coeficiente angular varia de amostra para amostra, ele é uma variável aleatória com sua própria distribuição, sua própria média e seu próprio erro padrão. (Lembre-se de que em Estatística I você aprendeu que o erro padrão de uma estatística é semelhante ao desvio padrão de uma variável aleatória.) Se olharmos à direita do coeficiente angular na Figura 4-2, veremos SE Coef, que se refere ao erro padrão do coeficiente angular (que, nesse caso, é 0,01456).

Agora, tudo o que você precisa é do valor  $t^*$  na tabela- $t$  (Tabela A-1 no apêndice). Já que  $n = 12$ , consulte a linha em que o grau de liberdade é  $12 - 2 = 10$ . Você também quer um intervalo de confiança de 95%, portanto, consulte a coluna para  $(1 - 0,95) \div 2 = 0,25$ . O





valor obtido para  $t^*$  é 2,228.

Juntando essas peças, o intervalo de confiança de 95% para o coeficiente angular da reta de regressão para o exemplo dos livros didáticos é  $0,11337 \pm 2,228 * 0,01456$ , ou seja, vai de 0,0809 a 0,1458. As unidades estão em quilos (peso dos livros) por quilos (peso das crianças). Observe que esse intervalo é grande em virtude do pequeno tamanho amostral, responsável por aumentar o erro padrão.

### ***Um teste de hipótese para o coeficiente angular***

Talvez você possa querer realizar um teste de hipótese para o coeficiente angular de uma reta de regressão como forma de avaliar sua adequação aos dados. Caso o coeficiente seja igual a zero ou próximo disso, a reta de regressão é basicamente plana, significando que, independentemente do valor de  $x$ , você sempre irá estimar  $y$  usando sua média. Isso quer dizer que  $x$  e  $y$  não se relacionam e, portanto, um valor específico de  $x$  não o ajudará a prever um valor específico para  $y$ . Você também pode fazer um teste para ver se o coeficiente angular é diferente de zero, mas isso não é algo comum. Assim, para todas as intenções e todos os propósitos, utilizo as hipóteses  $H_0: \beta = 0$  versus  $H_a: \beta \neq 0$ , onde  $\beta$  é o coeficiente angular do verdadeiro modelo.

Para conduzir um teste de hipótese para o coeficiente angular de uma reta de regressão linear simples, siga os procedimentos básicos para qualquer teste de hipótese. Pegue a estatística ( $b$ ) de seus dados, subtraia o valor na  $H_0$  (nesse caso, esse valor é 0) e padronize o resultado dividindo-o pelo erro padrão (veja o Capítulo 3 para mais informações sobre esse processo).

Usando a fórmula do erro padrão para  $b$ , a estatística de teste para testar a hipótese de que o coeficiente angular é ou não igual a zero é  $\frac{b-0}{SE_b}$ , onde  $SE_b = \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}$ , e  $s = \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2}$ .

No resultado do Minitab, ilustrado na Figura 4-2, a estatística de teste está bem à direita da coluna SE Coef; sabiamente marcada como T. Nesse caso,  $T = 7,78$ . Compare esse valor ao  $t^* = 2,228$  obtido na tabela- $t$ . Uma vez que  $T > t^*$ , você tem evidência suficiente para rejeitar  $H_0$  e concluir que o coeficiente angular da reta de regressão para os dados sobre peso dos livros didáticos não é zero. (Na verdade, ele tem que ser maior do que isso, de acordo com seus dados.)

Você também pode encontrar o valor- $p$  exato no resultado do Minitab. Ele aparece à direita da coluna T, e sua coluna está marcada como P. O valor- $p$  para o teste do coeficiente nesse caso é 0,000, o que significa que ele é menor do que 0,001. Pode-se concluir, então, que o coeficiente angular dessa reta não é zero e, portanto, o peso dos livros possui uma relação significativa com o peso dos alunos. (Vá ao Capítulo 3 para fazer uma revisão sobre o valor- $p$ .)

Para ver se o coeficiente angular tem um valor diferente de zero, basta substituir esse valor em  $b_0$  na fórmula para a estatística de teste. Além disso, você pode conduzir um teste de



hipótese unilateral para ver se o coeficiente angular é bem maior ou menor do que zero. Nesses casos, você encontra a mesma estatística de teste, mas compara-a ao valor  $t^*$ , área à direita (ou à esquerda, respectivamente) onde está  $\alpha$ .

## *Inspecionando o intercepto $y$*

O intercepto  $y$  é o local onde a reta de regressão  $y = a + bx$  cruza o eixo  $y$  e é representada pela letra  $a$  na equação (veja a seção “O intercepto  $y$  da reta de regressão”). Às vezes, o intercepto  $y$  pode ser interpretado de forma significativa, mas, em outras, não. Ele se diferencia do coeficiente angular, que é sempre interpretável. Na verdade, entre os dois elementos, o coeficiente angular é a estrela do show, e o intercepto  $y$  é um coadjuvante menos famoso, mas ainda assim notável.

Há momentos em que o intercepto  $y$  não faz sentido. Por exemplo, suponha que você utilize a chuva para prever a quantidade de quilogramas de milho por hectare. Se você não tiver nada de chuva, não vai ter nada de milho, mas se a reta de regressão cruzar o eixo  $y$  em algum ponto próximo a zero (e isso provavelmente vai acontecer), o intercepto  $y$  não terá sentido.

Outra situação é quando não há dados coletados próximos para o valor de  $x = 0$ ; a interpretação do intercepto  $y$ , nesses casos, é inapropriada. Por exemplo, se usarmos as notas obtidas por uma aluna no primeiro semestre para prever suas notas no segundo, a não ser que a aluna não tenha feito nenhuma prova (e, nesse caso, isso não conta), ela vai conseguir pelo menos alguns pontos.

Muitas vezes, no entanto, o intercepto  $y$  pode lhe interessar e possuir um valor passível de interpretação, como quando falamos da previsão de vendas de café usando a temperatura do dia da partida de futebol americano. Algumas partidas podem ser realizadas em dias em que a temperatura pode chegar a zero ou abaixo de zero (como as partidas do Packers<sup>1</sup>, por exemplo. (Vai, Pack!))

Suponha que eu colete os dados de dez de meus alunos que registraram seu tempo de estudo (em minutos) para um teste valendo 10 pontos, com suas respectivas notas. Através de todos os métodos usados neste capítulo, cheguei à conclusão de que os dados possuem uma forte relação linear (consulte a seção “Investigando relações através de Diagramas de Dispersão e Correlações”). Segui em frente e fiz uma análise de regressão, cujos resultados estão na Figura 4-3.

Como tive alunos que não estudaram nada para o teste (que Deus os proteja!), o intercepto  $y$  igual a 3,29 pontos (onde o tempo de estudo é  $x = 0$ ) pode ser interpretado tranquilamente. Seu valor pode ser visto na coluna Coef, na linha Constant. (Veja a seção “O intercepto  $y$  da reta de regressão” para mais informações.) O próximo passo é obter um intervalo de confiança para o intercepto  $y$  da reta de regressão, onde você vai poder tirar conclusões além dessa amostra de dez alunos.

A fórmula para um intervalo de confiança de nível  $1 - \alpha$  para o intercepto  $y$  ( $a$ ) de uma reta de regressão linear simples é  $a \pm t_{n-2}^* SE_a$ . O erro padrão,  $SE_a$ , é igual a  $SE_a = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$ , onde  $s = \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2}$ . Mais uma vez, o valor de  $t^*$  vem da distribuição- $t$  com  $n - 2$  graus de liberdade, cuja área à direita é igual a  $\alpha \div 2$ . Usando o resultado mostrado na Figura 4-3 e a tabela- $t$ , tenho 95% de certeza de que as notas para o teste ( $y$ ) para um aluno com tempo de estudo de  $x = 0$  minutos é  $3,29 \pm 2,306 * 0,4864$ , ou seja, algo entre 2,17 e 4,41, em média. Observe que 2,306 vem da tabela- $t$  com  $10 - 2 = 8$  graus de liberdade, e 0,4864 é o SE (erro padrão) para o intercepto  $y$  na Figura 4-3. (Sendo assim, não estudar nada para meus testes não é uma boa ideia.)

A propósito, para encontrar o quanto o tempo de estudo influenciou a nota do teste desses alunos, você pode fazer uma estimativa do coeficiente angular do resultado mostrado na Figura 4-3, onde esse valor é 0,1793, o que quer dizer que cada minuto de estudo está relacionado a um aumento de 0,1793 na nota, mais ou menos a margem de erro, é claro. Ou 10 minutos a mais está relacionado a 1,793 pontos a mais. Em um teste valendo 10 pontos, tudo conta!

The regression equation is quiz score = 3.29 + 0.179 minutes studying				
Predictor	Coef	SE Coef	T	P
Constant	3.2931	0.4864	6.77	0.000
Minutes studying	0.17931	0.02103	8.53	0.000
S = 0.877153      R-Sq = 90.1%      R-Sq(adj) = 88.8%				

**Figura 4-3:** Análise de regressão para tempo de estudo e nota de teste.

Realizar um teste de hipótese para o intercepto  $y$  realmente não é algo que você vai precisar fazer muito, pois, na maioria das vezes, você não tem uma noção preconcebida de qual seria o intercepto  $y$  (nem precisa se preocupar com isso antes do tempo). O intervalo de confiança é muito mais útil. Entretanto, caso você realmente tenha que conduzir um teste de hipótese para o intercepto  $y$ , subtraia o intercepto do valor da  $H_0$  e divida o resultado pelo erro padrão, encontrando o resultado do Minitab no cruzamento entre a linha Constant e a coluna SE Coef (o valor atribuído como padrão é para ver se o intercepto  $y$  é igual a zero ou não). O teste está na coluna T do resultado e seu valor- $p$  é mostrado na coluna P. No exemplo do tempo de estudo versus a nota do teste, o valor- $p$  é 0,000, portanto, o intercepto  $y$  é significativamente diferente de zero. Tudo isso significa que a reta cruza o eixo  $y$  em algum ponto.

## ***Construindo intervalos de confiança para a resposta média***

Quando você tiver o coeficiente angular e o intercepto  $y$  da reta de regressão, junte-os para



formar a reta  $y = a + bx$ . O valor de  $y$ , aqui, realmente representa o valor médio de  $y$  para um determinado valor de  $x$ . Por exemplo, nos dados do peso dos livros didáticos, a Figura 4-2 mostra a equação da reta de regressão  $y = 1,68 + 0,113x$ , onde  $x$  é o peso médio dos alunos, e  $y$ , o peso médio dos livros. Se você substituir  $x$  por 45 quilos, terá  $y = 1,68 + 0,113 * 45 = 6,77$  quilos para o peso do livro para o grupo cuja média é 45 quilos. Este número (6,77) é uma estimativa do peso médio dos livros para as crianças com esse peso.

Mas você não deve parar aqui. Uma vez que obteve uma estimativa do peso médio do livro usando  $y$ , também vai precisar de uma margem de erro que o acompanhe, a fim de criar um intervalo de confiança para o  $y$  médio obtido a partir de um dado  $x$  que generalize a população.

Pegue a estimativa para  $y$ , obtida a partir da substituição de um dado valor de  $x$  na equação da reta de regressão, adicione ao valor de  $y$  a margem de erro e a subtraia também. A fórmula para um intervalo de confiança  $1 - \alpha$  para a média de  $y$  a partir de um dado valor de  $x$  (chamado de  $x^*$ ) é igual a  $y \pm t_{n-2}^* SE_y$ , onde  $y$  é o valor da equação da reta quando você

substitui  $x$  por  $x^*$ . O erro padrão para  $y$  é igual a  $SE_y = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$ , onde  $s = \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2}$ . Felizmente, o Minitab realiza esses cálculos para você e relata um intervalo de confiança para a média de  $y$  para um dado  $x^*$ .

Para encontrar o intervalo de confiança para o valor médio de  $y$  usando o Minitab, solicite uma análise de regressão (veja as instruções na seção “Encontrando a reta certa para modelar seus dados”) e clique em Options. Você vai ver um campo chamado Prediction Intervals for New Observations; insira o valor  $x^*$  desejado e, no campo abaixo, coloque seu nível de confiança (o nível adotado como padrão é 95%). Embora esse campo se chame Prediction Intervals, ele encontra tanto o intervalo de confiança quanto o de previsão. (Os intervalos de previsão são diferentes dos intervalos de confiança, e vou discuti-los na próxima seção.) No resultado do Minitab, o intervalo de confiança é chamado de 95% CI.

Voltando ao exemplo do peso dos livros, o resultado do Minitab que encontra um intervalo de confiança de 95% para o peso médio dos livros para crianças com peso médio de 45 quilos é mostrado na Figura 4-4. O resultado é (6,312; 7,228) quilos. Tenho 95% de certeza de que o peso médio dos livros para o grupo de crianças cujo peso médio é 45 quilos fica entre 6,312 e 7,228 quilos. (Crianças, joguem fora essas mochilas tão pesadas!)

Apenas faça previsões para o valor médio de  $y$  a partir de valores de  $x$  que estejam dentro do conjunto de dados coletados. Se você não fizer isso, acabará causando uma das proibições estatísticas conhecida como *extrapolação*. (Veja mais adiante a seção “Conhecendo os Limites de Sua Análise de Regressão”).

## Previendo o futuro com os intervalos de previsão



Suponha que, em vez do valor médio de  $y$ , você queira estimar o valor de  $y$  para um futuro valor de  $x$ . Uma vez que estamos falando do futuro, você tem que fazer uma previsão, e, para isso, vai precisar de um conjunto de possíveis valores de  $y$  para um dado  $x^*$ . Isso é o que nós estatísticos chamamos de *intervalo de previsão*.

A fórmula para um intervalo de previsão de nível  $1 - \alpha$  para  $y$ , dado o valor de  $x^*$ , é

$SE_y = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$ , onde  $s = \sqrt{\frac{1}{n-2} \sum_i (y_i - \hat{y}_i)^2}$ . Mais uma vez, o Minitab faz esses cálculos por você.

Para encontrar um intervalo de previsão com nível  $1 - \alpha$  para um valor de  $y$  dado  $x^*$  usando o Minitab, solicite uma análise de regressão (veja as instruções na seção “Encontrando a reta certa para modelar seus dados”) e clique em Options. No campo Prediction Intervals for New Observations, insira o valor  $x$  desejado e, no campo logo abaixo, coloque seu nível de confiança (o nível adotado como padrão é 95%). No resultado do Minitab, o intervalo de previsão recebe o nome de 95% PI e aparece à direita do intervalo de confiança para a média de  $y$  para o mesmo  $x^*$ .

### ***Prevendo o peso dos livros por meio do peso do aluno***

Para os dados a respeito do peso do livro didático, suponha que você já tenha encontrado a reta de regressão e, agora, um novo aluno entre em cena. Você quer, então, prever o peso dos livros desse aluno. Isso significa que você quer o intervalo de previsão, e não o de confiança, pois quer prever o peso dos livros para uma pessoa, e não o peso médio de um grupo. Suponha que esse novo aluno pese 45 quilos. Para encontrar o intervalo de previsão para o peso dos livros para esse aluno, utilize  $x^* = 45$  quilos e deixe o resto com o Minitab.

O resultado do Minitab na Figura 4-4 mostra que um intervalo de previsão de 95% para o peso dos livros para uma única criança que pese 45 quilos é (5,173; 8,367) quilos. Observe que esse valor é mais amplo do que o intervalo de confiança (6,312; 7,228) para o peso médio dos livros para um grupo de crianças que pesam 45 quilos, encontrado na seção “Construindo intervalos de confiança para a resposta média”. Essa diferença se deve ao aumento da variabilidade no fato de se observar apenas uma criança e prever um peso para os livros.

Predicted Values for New Observations				
New				
Obs	Fit	SE Fit	95% CI	95% PI
1	6,77	0,205	(6,312; 7,228)	(5,173; 8,367)

**Figura 4-4:** Intervalo de previsão do peso dos livros para uma criança de 45 quilos.

## *Comparando os intervalos de previsão e de confiança*

Note que as fórmulas para os intervalos de previsão e de confiança são muito semelhantes. Na verdade, a fórmula do intervalo de previsão é exatamente igual à fórmula do intervalo de confiança, exceto pelo fato de que a primeira tem um número 1 embaixo do radical. Por causa dessa diferença nas fórmulas, a margem de erro para um intervalo de previsão é mais ampla do que a de um intervalo de confiança.

Essa diferença também faz sentido se a consideramos sob a perspectiva estatística. Um intervalo de previsão possui mais variabilidade do que um intervalo de confiança pois é mais difícil fazer uma previsão sobre  $y$  para um único valor de  $x^*$  do que estimar o valor médio de  $y$  para um dado  $x^*$  (por exemplo, as notas individuais de provas variam mais do que as notas médias). O intervalo de previsão será mais amplo do que o intervalo de confiança, ou seja, ele terá uma margem de erro maior.

A semelhança entre os intervalos de previsão e de confiança está no fato de que a fórmula para suas margens de erro contém  $x^*$ , o que significa que a margem de erro nos dois casos depende do valor de  $x^*$  usado. Nos dois casos, se você usar o valor médio de  $x$  como sendo seu  $x^*$ , a margem de erro para cada intervalo ficará em seu menor valor, pois há mais dados ao redor da média de  $x$  do que ao redor de qualquer outro valor. À medida que você se afasta da média de  $x$ , a margem de erro para cada intervalo aumenta.

# ***Checando a Adequação do Modelo (dos Dados, Não das Roupas!)***

Depois de ter estabelecido uma relação entre  $x$  e  $y$  e ter obtido a equação da reta que representa tal relação, você deve estar pensando que seu trabalho acabou (muitos pesquisadores erroneamente param aqui e, por isso, estou contando com você para acabarmos com esse ciclo!). A tarefa mais importante ainda tem que ser cumprida: verificar se as condições do modelo realmente estão satisfeitas e se o modelo se adequa bem a meios mais específicos do que diagramas de dispersão e medidas de correlação (sobre os quais falo na seção “Investigando Relações com Diagramas de Dispersão e Correlações”).

Esta seção apresenta métodos para definir e avaliar a adequação de um modelo de regressão linear simples.

## ***Definindo as condições***

As duas principais condições a serem atendidas antes da aplicação de um modelo de regressão linear simples a um conjunto de dados são:

- ✓ Os valores de  $y$  devem ter uma distribuição aproximadamente normal para cada valor de  $x$ .
- ✓ Os valores de  $y$  devem ter uma quantidade constante de dispersão (desvio padrão) para cada valor de  $x$ .

### ***Para cada $x$ , um $y$ normal***

Para qualquer valor de  $x$ , a população de possíveis valores de  $y$  deve ter uma distribuição normal. A média dessa distribuição é o valor de  $y$  que está na reta de regressão para um certo valor de  $x$ . Ou seja, alguns de seus dados se encontram acima da reta de regressão, outros abaixo e alguns podem estar bem na reta.

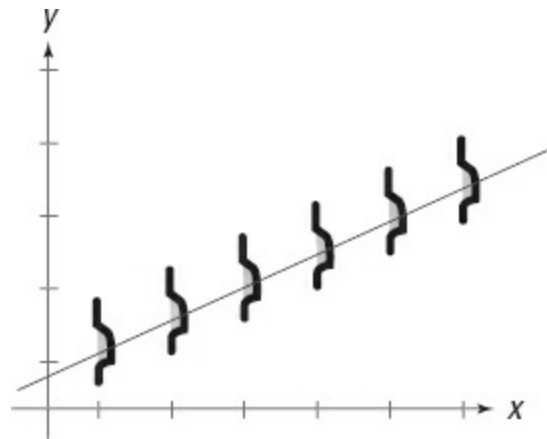


Se o modelo estiver bem adequado, os valores dos dados deverão estar espalhados ao redor da reta de regressão de modo que cerca de 68% deles estejam a menos de um desvio padrão da reta, cerca de 95% até dois desvios padrão, e cerca de 99,7% até três desvios padrão da reta. Essa especificação, como você deve se recordar de seu curso de Estatística I, é chamada *regra 68-95-99,7* (ou regra empírica), e se aplica a todos os dados distribuídos em uma curva em forma de sino (aos quais se aplica a distribuição normal).

Você pode notar na Figura 4-5 que, assim como para cada valor de  $x$ , os valores de  $y$  a serem observados tendem a se localizar próximos da reta de regressão em maior número e, à medida que você se afasta da reta tanto para cima quanto para baixo, vê cada vez menos valores  $y$ . Mais do que isso, eles estão espalhados ao redor da reta de forma a refletir uma curva em forma de sino: a distribuição normal. E isso indica uma boa adequação.



Por que essa condição faz sentido? Os dados coletados em  $y$  para qualquer valor específico de  $x$  variam de um indivíduo para outro. Por exemplo, nem todos os livros dos alunos têm o mesmo peso, mesmo para os alunos que pesam o mesmo. No entanto, tais valores não podem variar da maneira que quiserem. Para atender às condições de um modelo de regressão linear, para cada valor de  $x$ , os dados devem se espalhar ao redor da reta de acordo com uma distribuição normal. A maioria dos pontos deve estar próxima da reta e, à medida que você se distancia da reta, pode esperar a ocorrência de cada vez menos pontos. Assim, a condição número um é de que os dados tenham uma distribuição normal para cada valor de  $x$ .



---

**Figura 4-5:** Condições de um modelo de regressão linear simples.

---

### ***A mesma dispersão para cada $x$***

Para usar o modelo de regressão linear simples, à medida que você se move da esquerda para a direita no eixo  $x$ , a dispersão nos valores de  $y$  ao redor da reta deve ser a mesma, independentemente do valor de  $x$  que você esteja procurando. Esse requerimento recebe o nome de *condição de homocedasticidade*. (Como eles começaram a usar esse nome enorme só para descrever o fato de que os desvios padrão devem permanecer iguais ao longo dos valores de  $x$ , eu nunca vou saber.) Essa condição garante que a reta de regressão mais adequada funcione bem para todos os valores relevantes de  $x$ , e não apenas para certas áreas.

Na Figura 4-5, você pode ver que, independentemente do valor de  $x$ , a dispersão dos valores de  $y$  continua a mesma o tempo todo. Se a dispersão ficasse cada vez maior à medida que  $x$  aumentasse, por exemplo, a reta perderia sua capacidade de se adequar bem aos valores maiores de  $x$ .

### ***Encontrando e investigando os resíduos***

Para saber se os valores  $y$  vêm de uma distribuição normal ou não, é preciso medir a distância entre suas previsões e os dados reais. Essas diferenças são chamadas de *erros* ou *resíduos*. Para ver se um modelo se adequa bem, você precisa verificar esses erros e ver



como se sobrepõem.



Quando falamos da adequação de um modelo, a palavra *erro* não significa “engano”. Ela apenas se refere à diferença entre os dados e à previsão feita com base no modelo. Por isso, a palavra que prefiro usar para descrever essa diferença é *resíduo*. Pelo menos, soa de modo mais positivo.

As seções a seguir focam em como encontrar uma forma para medir esses resíduos gerados pelo modelo. Você também vai investigar os resíduos para identificar problemas ocorridos durante o processo de adequação de uma reta aos dados. Ou seja, você vai descobrir que observar os resíduos o ajudará a avaliar a adequação do modelo e a diagnosticar problemas que causaram uma adequação ruim, se for o caso.

## ***Encontrando os resíduos***

O *resíduo* é a diferença entre o valor previsto de  $y$  (a partir da reta de regressão mais adequada) e o valor de  $y$  observado, também conhecido como  $y$  (do conjunto de dados). Sua representação é  $(y - \bar{y})$ . De modo geral, para qualquer ponto de dados, você subtrai o valor  $y$  esperado (retirado da reta) do valor  $y$  observado (retirado dos dados). Se o resíduo for grande, a reta não se adequará de forma apropriada. Se o resíduo for pequeno, a reta se adequará de forma apropriada.

Por exemplo, suponha que você tenha o ponto (2,4) em seu conjunto de dados e a equação da reta de regressão seja  $y = 2x + 1$ . O valor esperado para  $y$ , nesse caso, é  $(2 * 2) + 1 = 5$ . O valor observado para  $y$  a partir do conjunto de dados é 4. Fazendo a diferença entre o valor observado e o valor esperado, você tem  $4 - 5 = -1$ . O resíduo para o ponto (2,4) em particular é  $-1$ . Se você observar um valor  $y$  igual a 6 e usar a mesma reta para estimar  $y$ , então o resíduo será  $6 - 5 = +1$ .



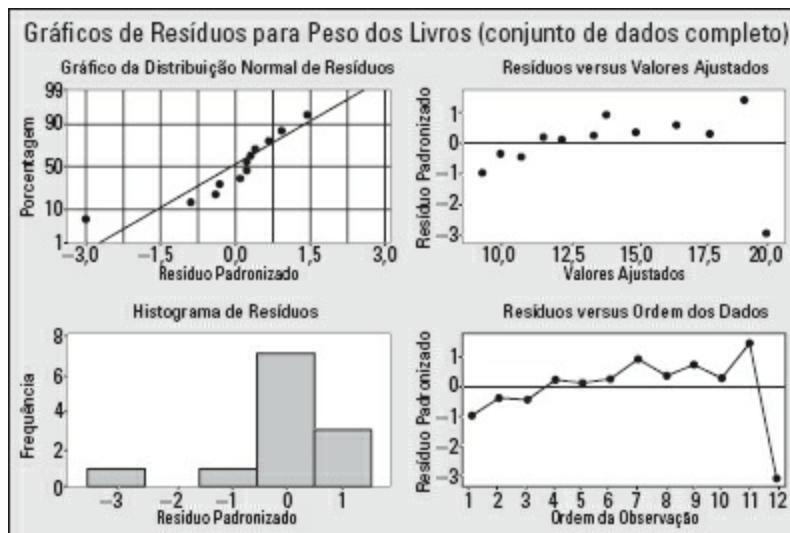
No geral, um resíduo positivo significa que você subestimou  $y$  para aquele ponto; a reta está abaixo dos dados. Um resíduo negativo significa que você superestimou  $y$  para aquele ponto; a reta está acima dos dados.

## ***Padronizando os resíduos***

Os resíduos, em sua forma pura, estão na mesma unidade que os dados originais, o que dificulta seu julgamento fora de um contexto. Para facilitar a interpretação dos resíduos, os estatísticos normalmente fazem a *padronização* — ou seja, subtraem a média dos resíduos (zero) e dividem-na pelo desvio padrão de todos eles. Os resíduos são um conjunto de dados como outro qualquer, portanto, é possível encontrar sua média e desvio padrão através dos mesmos procedimentos. Padronizar significa apenas converter em um escore  $Z$  para que você possa ver onde ele se encontra dentro da distribuição normal padrão. (Consulte um livro de Estatística I ou o *Estatística Para Leigos*, da Alta Books, para uma explicação mais completa sobre escores- $Z$ .)

## Plotando os resíduos

Podemos colocar os resíduos em um gráfico chamado *gráfico de resíduos* (se você já tiver padronizado os resíduos, o gráfico passa a se chamar *gráfico de resíduos padronizados*). A Figura 4-6 mostra o resultado do Minitab para uma variedade de gráficos de resíduos padronizados, todos com o mesmo propósito: verificar se as condições para o modelo de regressão linear simples são atendidas.



**Figura 4-6:** Gráficos de resíduos padronizados para os dados do peso dos livros didáticos

Gráficos de Resíduos para Peso dos Livros (conjunto de dados completo)

### Verificando a normalidade

Se a condição para a normalidade for satisfeita, você poderá ver no gráfico de vários resíduos (padronizados) próximos a zero; conforme você se afastar de zero, verá cada vez menos resíduos. **Observação:** não espere ver um resíduo padronizado em +3 ou -3 ou além desses valores. Caso isso aconteça, considere esse ponto um outlier (valor discrepante), além da necessidade de mais investigação. (Para mais informações sobre os outliers, veja a seção “Analisando outliers”, mais adiante neste capítulo.)

Os resíduos também devem aparecer aleatoriamente — alguns acima da reta, outros abaixo dela. Se um padrão para os resíduos for observado, significa que a reta pode não estar se ajustando corretamente.

Os gráficos na Figura 4-6 parecem ter um problema com a última observação, a que se refere aos alunos do 3º ano do Ensino Médio. Nessa observação, o peso médio de um aluno (64,41 quilos) parece seguir o padrão de aumentar a cada série, mas o peso do livro (7,28 quilos) foi menor do que o peso para os alunos do 2º ano do Ensino Médio (9,43 quilos) e é o primeiro ponto a quebrar o padrão.

Também é possível ver no canto superior direito do gráfico da Figura 4-6 que o último



dado tem um resíduo padronizado que se afasta dos demais e tem o valor de  $-3$  (algo que deveria ocorrer muito raramente). Assim, o valor esperado para  $y$ , com base em sua reta, estava fora por um fator de três desvios padrão. E, já que esse resíduo é negativo, o que foi observado para  $y$  foi muito mais baixo do que se esperava usando a reta de regressão.

Os outros resíduos parecem se encontrar na reta e estarem em uma distribuição normal, como você pode ver no gráfico superior direito da Figura 4-6. Os resíduos se concentram ao redor de zero, com poucos aparecendo à medida que você se afasta de zero. Esse mesmo padrão também pode ser observado no gráfico superior esquerdo da Figura 4-6, que demonstra o quão próximos da normalidade os resíduos estão. A reta nesse gráfico representa a reta normal. Se os resíduos se mantiverem próximos à reta, então a normalidade está ok; se não, você está enrascado (no sentido estatístico, é claro). Você vê que o resíduo com o valor mais alto é  $-3$ , e que esse número se encontra a uma boa distância da reta.

O gráfico na parte inferior esquerda da Figura 4-6 é um histograma dos resíduos padronizados, e você consegue notar que ele não se parece muito com uma distribuição em forma de sino. Ele nem mesmo é *simétrico* (um lado igual ao outro quando partido ao meio). O problema, mais uma vez, parece ser o resíduo  $-3$ , que distorce o histograma para a esquerda.

O gráfico na parte inferior direita da Figura 4-6 demonstra os resíduos na ordem em que se apresentam no conjunto de dados na Tabela 4-1. Uma vez que os dados já foram coletados, o gráfico residual da parte inferior direita se parece com o gráfico da parte superior direita, exceto pelo fato de os pontos estarem ligados. O gráfico inferior direito faz com que o resíduo  $-3$  apareça ainda mais.

### ***Verificando a dispersão dos valores de $y$ para cada $x$***

O gráfico do canto superior direito da Figura 4-6 também considera a condição de homocedasticidade. Se essa condição for satisfeita, então os resíduos para cada valor de  $x$  possuem mais ou menos a mesma dispersão. Se você passar uma reta vertical através de cada valor de  $x$ , os resíduos terão mais ou menos a mesma dispersão (desvio padrão), exceto o último, que, novamente, representa o 3º ano do Ensino Médio. Isso significa que a condição de haver igual dispersão para valores de  $y$  foi satisfeita para o exemplo do peso dos livros didáticos.

Se você tiver que observar apenas um gráfico de resíduos, escolha o do canto superior direito da Figura 4-6, o gráfico dos valores ajustados (os valores de  $y$  na reta) versus os resíduos padronizados. A maioria dos problemas com o ajuste do modelo aparecerá nesse gráfico, pois um resíduo é definido como a diferença entre o valor observado de  $y$  e o valor ajustado de  $y$ . No mundo perfeito, todos os valores ajustados não possuem resíduos; um resíduo grande (tal como o valor onde o peso estimado para o livro é 9,07 quilos para alunos com um peso médio de 64,41 quilos. Veja a Figura 4-1) é indicado por um ponto bem afastado de zero. Esse gráfico também mostra os desvios em relação ao padrão geral



da reta; por exemplo, se houver grandes resíduos nas extremidades desse gráfico (valores ajustados muito baixos ou muito altos), a reta não está se ajustando a essas áreas. No geral, podemos dizer que essa reta se ajusta bem pelo menos às séries de 1º ano do Ensino Fundamental ao 1º ano do Ensino Médio.

## *Usando $r^2$ para medir o ajuste do modelo*

Uma maneira importante de avaliar o ajuste de um modelo é usar a estatística chamada *coeficiente de determinação*, ou  $r^2$ . Essa estatística pega o valor da correlação,  $r$ , e o eleva ao quadrado para lhe dar uma porcentagem. Você interpreta o  $r^2$  como a porcentagem de variabilidade na variável  $y$  que é explicada por ou em virtude de sua relação com a variável  $x$ .

Os valores  $y$  dos dados que você coleta possuem muita variabilidade dentro e fora deles mesmos. Por isso, é preciso buscar outra variável ( $x$ ) para ajudá-lo a explicar a variabilidade nos valores  $y$ . Depois de colocar aquela variável  $x$  no modelo e descobrir que ela é altamente correlacionada a  $y$ , você deve descobrir como esse modelo se saiu ao explicar por que os valores de  $y$  são diferentes.

Observe que é preciso interpretar  $r^2$  usando padrões diferentes dos utilizados para interpretar  $r$ . Uma vez que o quadrado de um número entre  $-1$  e  $+1$  é menor do que o próprio número (exceto para  $-1$ ,  $+1$  e  $0$ , que permanecem inalterados ou mudam apenas o sinal), um  $r^2$  de  $0,49$  não é tão ruim, pois é o quadrado de  $r = 0,7$ , uma correlação bastante forte.

A seguir, veja algumas regras gerais para interpretar o valor de  $r^2$ :

- ✓ Se o modelo contendo  $x$  explicar muito da variabilidade nos valores  $y$ , então  $r^2$  é alto (entre  $80$  e  $90\%$  é considerado extremamente alto). No entanto, valores como  $0,70$  ainda são considerados bastante altos. Uma alta porcentagem de variabilidade significa que a reta se ajusta bem, pois não há muito o que explicar sobre o valor de  $y$ , a não ser usando  $x$  e a relação entre eles. Sendo assim, um valor alto para  $r^2$  é algo positivo.
- ✓ Se o modelo contendo  $x$  não explicar muito sobre a diferença nos valores  $y$ , então  $r^2$  é baixo (próximo a zero; entre, aproximadamente, digamos,  $0,00$  e  $0,30$ ). O modelo, nesse caso, não se ajustaria bem. Você precisaria de uma outra variável para explicar  $y$ , uma diferente da que você já tentou usar.
- ✓ Valores de  $r^2$  que se encontram no meio (entre, digamos,  $0,30$  e  $0,70$ ) significam que  $x$  ajuda a explicar um pouco de  $y$ , mas não o faz tão bem sozinho. Nesse caso, os estatísticos tentariam adicionar uma ou mais variáveis ao modelo para uma explicação mais completa de  $y$  como um grupo (leia mais sobre isso no Capítulo 5).

Para o exemplo do livro didático, o valor de  $r$  (o coeficiente de correlação) é 0,93. O quadrado desse valor é  $r^2 = 0,8649$ . Esse número quer dizer que aproximadamente 86% da variabilidade encontrada nos pesos médios dos livros para todos os alunos (valores  $y$ ) é explicada pelo peso médio dos alunos (valores  $x$ ). Essa porcentagem mostra que o modelo que usa a série escolar para estimar o peso da mochila é uma boa aposta.

No caso da regressão linear simples, você tem apenas uma variável  $x$ , mas, no Capítulo 5, vamos ver modelos que contêm mais do que uma variável. Nessa situação, você vai utilizar o  $r^2$  para organizar a contribuição que essas variáveis  $x$ , como um todo, trazem ao modelo.

## *Analizando outliers*

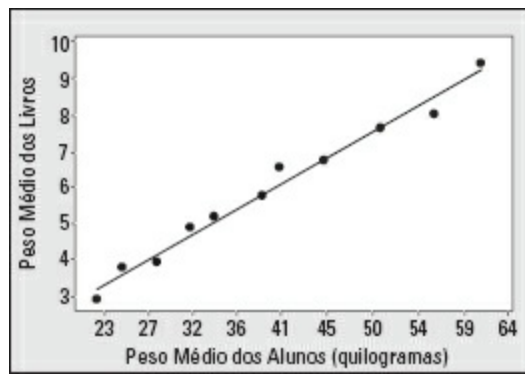
Às vezes, a vida não é perfeita (sério?), e você pode acabar encontrando um resíduo totalmente discrepante em seu conjunto de dados, aparentemente tão organizado. Esse valor se chama *outlier*, e possui um valor padronizado fixado em +3 ou -3 ou além destes. Ele é uma ameaça às condições de seu modelo de regressão e pode fazer você procurar seu professor aos prantos.

Antes de entrar em pânico, a melhor coisa a fazer é examiná-lo mais de perto. Primeiro, será que há um erro nos valores dos dados? Será que alguém escreveu a idade como 642, por exemplo? Caso você encontre um erro comprovável em seus dados (afinal de contas, errar é humano!), remova-o (ou conserte-o, se possível) e analise os dados sem ele. Entretanto, se você não conseguir explicar o problema através da descoberta de um engano, comece a pensar em outra abordagem.

Se não conseguir encontrar um engano que tenha causado o outlier, não precisa jogar seu modelo fora; afinal de contas, é só um ponto de dados. Analise os dados com e sem aquele ponto. Depois, faça os registros e compare as duas análises. Essa comparação lhe dá uma ideia do quanto aquele ponto de dados influencia a análise e pode levar outros pesquisadores a conduzirem mais estudos para focar o problema que você trouxe à tona.

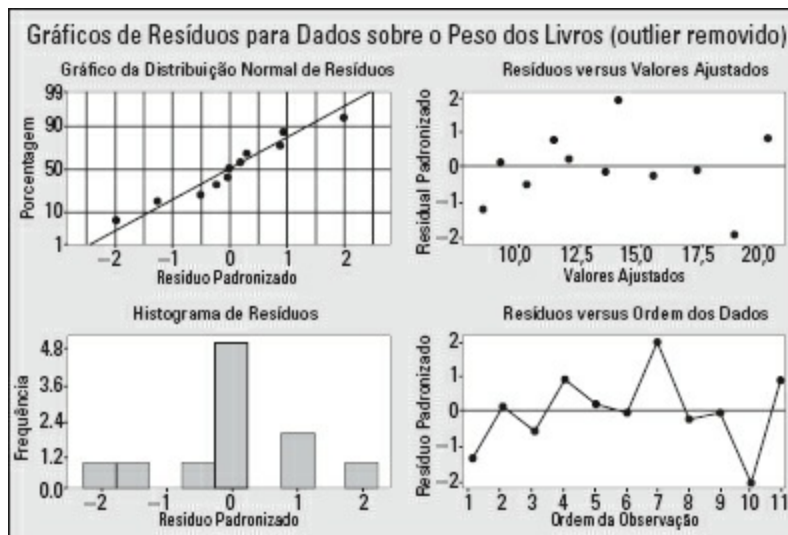
Na Figura 4-1, vemos o diagrama de dispersão do conjunto de dados completo para o exemplo do peso dos livros didáticos. A Figura 4-7 ilustra o diagrama de dispersão para o conjunto de dados sem o outlier. O diagrama ajusta os dados de forma melhor sem o outlier. A correlação aumenta para 0,993, e o valor de  $r^2$  aumenta para 0,986. A equação para a reta de regressão para esse conjunto de dados é  $y = 1,78 + 0,139x$ .

No entanto, o coeficiente angular da reta de regressão não se altera muito depois da remoção do outlier (compare-a com a Figura 4-2, onde o coeficiente angular é 0,113). Porém, o intercepto  $y$  se altera: ele passa a ser 1,78 sem o outlier, comparada a 3,69 com ele. Os coeficientes angulares da reta são quase iguais, mas a reta passa a cruzar o eixo  $y$  em lugares diferentes. Aparentemente, o outlier (o último ponto no conjunto de dados) possui um efeito considerável sobre a reta de regressão.



**Figura 4-7:** Diagrama de dispersão dos dados para o peso dos livros didáticos sem o outlier.

A Figura 4-8 ilustra o gráfico dos resíduos para a reta de regressão do conjunto de dados sem o outlier. Cada um desses gráficos mostra um ajuste muito melhor dos dados ao modelo, quando comparados à Figura 4-6. Esse resultado mostra que os dados para a 12ª série influenciam esse conjunto de dados e que o outlier deve ser observado e, talvez, mais investigado. Os alunos se esforçam mais quando estão nas primeiras séries? Ou, simplesmente, decidem não ficar carregando livros por aí quando se tornam adolescentes? (O trabalho do estatístico não é ficar imaginando o porquê, mas, sim, fazer e analisar.)



**Figura 4-8:** Gráficos de resíduos para os dados sobre o peso dos livros didáticos sem o outlier.

# *Conhecendo as Limitações de Sua Análise de Regressão*

O principal de qualquer análise de dados é tirar as conclusões corretas a partir de um determinado resultado. Quando se trabalha com um modelo de regressão simples, existe a possibilidade de se cometer três grandes erros. Esta seção vai mostrá-los a você e dizer o que você deve fazer para evitá-los.

## *Evitando cair no modo causa e efeito*

Em uma regressão linear simples, você investiga se  $x$  se relaciona com  $y$  e, caso consiga uma correlação forte e um diagrama de dispersão que mostre uma tendência linear, encontre a reta de regressão e utilize-a para estimar o valor de  $y$  para valores relevantes de  $x$ .



Existe, no entanto, uma linha tênue que sua interpretação dos resultados de regressão não pode ultrapassar. Tome cuidado para não fazer uma interpretação automática do coeficiente angular no modo de causa e efeito quando estiver usando a reta de regressão para estimar o valor de  $y$  através de  $x$ . Isso pode gerar um voto de confiança capaz de mandá-lo para a fogueira. A menos que tenha utilizado um experimento controlado para obter os dados, você pode apenas presumir que as variáveis se correlacionam; você não pode garantir por que elas se relacionam.

No exemplo do peso dos livros escolares, estimamos o peso médio dos livros usando o peso dos estudantes, mas isso não significa que, se aumentarmos o peso de uma criança, o peso de seu livro também vai aumentar. Por exemplo, em virtude da forte correlação positiva, você sabe que os alunos com pesos mais baixos estão associados a livros com pesos médios mais baixos, e alunos com pesos mais altos tendem a ter livros mais pesados. Porém, você não pode pegar um aluno da terceira série, aumentar seu peso e pronto — de repente, seus livros estão pesando mais.

A variável subjacente à relação entre o peso de uma criança e o peso de sua mochila é a série escolar do aluno sob uma perspectiva acadêmica; à medida que a série aumenta, também aumentam o tamanho e o número de livros, bem como a tarefa de casa. A série escolar do aluno é o que direciona tanto seu peso quanto o peso de seus livros. Nessa situação, a série escolar do aluno é o que os estatísticos chamam de *variável de confusão*; é uma variável que não foi incluída no modelo, mas se relaciona tanto com o resultado quanto com a variável de resposta. Uma variável de confusão dificulta a compreensão da relação de causa e efeito.



Se os dados coletados são o resultado de um experimento bem projetado, com controle para as possíveis variáveis de confusão, é possível estabelecer uma relação de causa e efeito entre  $x$  e  $y$ , caso eles possuam uma forte correlação. Caso contrário, você não pode estabelecer uma relação desse tipo. (Consulte um livro de Estatística I ou o *Estatística Para Leigos*, da Alta Books, para uma explicação mais completa sobre experimentos.)



## ***Extrapolação: N-A-O-Til, NUNCA!***

O uso de valores de  $x$  que estejam fora dos limites razoáveis de  $x$  é conhecido como *extrapolação*. E um de meus colegas resume muito bem essa ideia: “Amigos não deixam amigos extrapolar.”

Quando você determina a reta de regressão para seus dados, obtém uma equação que lhe permite substituir um valor para  $x$  e, assim, prever o valor de  $y$ . Em álgebra, quando você encontra a equação de uma reta e a coloca em um gráfico, a reta normalmente terá uma seta em ambas as pontas indicando que ela é infinita nas duas direções. No entanto, isso não funciona em estatística (pois a estatística representa o mundo *real*). Quando lidamos com unidades do mundo real, como altura, peso, QI, notas, preços de casas e o peso de seu livro de Estatística, apenas alguns números fazem sentido.

Então, o primeiro ponto é: não substitua na equação os valores de  $x$  que não fazem sentido. Por exemplo, se você estiver estimando o preço de uma casa ( $y$ ) usando metros quadrados ( $x$ ), nem pense em substituir na equação um valor de  $x$  igual a 10 metros quadrados ou 100 metros quadrados, pois não existe casa tão pequena assim.

Também não vá substituir  $x$  por valores como 1.000.000 metros quadrados (a não ser que sua “casa” seja o estádio de futebol americano do Estado de Ohio). Isso não faria o menor sentido. Da mesma forma, se estiver estimando a temperatura de amanhã usando a temperatura de hoje, os números negativos para  $x$  poderiam fazer sentido, mas, se estiver estimando a quantidade de chuva para amanhã dada a quantidade de chuva de hoje, números negativos para  $x$  (ou  $y$ , nesse caso) não fariam sentido.

Escolha somente os valores relevantes de  $x$  através dos quais vai estimar  $y$  — ou seja, atenha-se aos valores de  $x$  presentes em seus dados e se mantenha dentro desses limites quando fizer suas previsões. No exemplo do livro didático, o menor peso médio dos alunos é 22,00 quilos, e o maior é 64,41 quilos. Escolher pesos entre 22,00 e 64,41 para substituir  $x$  na equação está permitido, mas escolher valores menores do que 22,00 ou maiores que 64,41 não é uma boa ideia. Você não pode garantir que a mesma relação linear (ou qualquer relação linear nesse caso) continue fora desses limites.

Pense nisso: se a relação encontrada realmente continuar para qualquer valor de  $x$ , independentemente de seu tamanho, então um atacante da OSU que pesa 113,4 quilos teria que carregar  $1,68 + 0,113 * 113,4 = 14,5$  quilos de livros em sua mochila. É claro que isso não seria difícil para ele, mas, e quanto a nós?

## ***Às vezes é preciso ter mais do que uma variável***

Um modelo de regressão linear simples é justamente o que seu nome diz: simples. Não quero dizer que seja fácil trabalhar com ele, mas simples no sentido estrutural. Esse modelo tenta estimar o valor de  $y$  usando apenas uma variável,  $x$ . Entretanto, o número de situações reais que podem ser explicadas com o uso de uma regressão linear simples e de





uma única variável é pequeno. Muitas vezes, uma variável sozinha não consegue fazer todas as previsões.

Se uma variável não lhe proporciona um modelo que se ajuste bem, tente adicionar mais variáveis. São necessárias muitas variáveis para se obter uma boa estimativa de  $y$ , mas também é preciso ter muito cuidado ao escolhê-las. No caso dos preços na bolsa de valores, por exemplo, ainda hoje se busca um modelo de previsão realmente eficaz.

Outro exemplo são as empresas de plano de saúde que tentam estimar o quanto você vai viver através de uma série de perguntas (sendo que cada uma delas representa uma variável no modelo de regressão). Não é possível encontrar uma única variável que estime seu tempo de vida; muitos fatores devem ser considerados: sua saúde, seu peso, se você é ou não fumante, fatores genéticos, a frequência com que você faz exercícios físicos, e a lista ainda pode continuar infinitamente.

O ponto é que os modelos de regressão nem sempre usam apenas uma variável,  $x$ , para estimar  $y$ . Para fazer isso, alguns modelos utilizam duas, três e até mais variáveis. Tais modelos não são chamados de modelos de regressão linear simples, mas de *modelos de regressão linear múltipla*, em virtude do emprego de múltiplas variáveis. (Vamos investigar os modelos de regressão linear múltipla no Capítulo 5.)

---

<sup>1</sup> N.E.: Green Bay Packers: equipe integrante da Liga Nacional de Futebol Americano (NFL), com sede em Green Bay, Wisconsin, EUA.

# Capítulo 5

## Regressão Múltipla com Duas Variáveis X

---

### *Neste Capítulo*

- ▶ Conhecendo os conceitos por trás de um modelo de regressão múltipla
  - ▶ Encontrando, interpretando e testando coeficientes
  - ▶ Verificando o ajuste do modelo
- 

**O** conceito de regressão é o de construir um modelo que estime ou preveja uma variável quantitativa ( $y$ ) usando pelo menos uma outra variável quantitativa ( $x$ ). A regressão linear simples utiliza apenas uma variável  $x$  para estimar a variável  $y$ . (Consulte o Capítulo 4 para tudo o que você precisa sobre regressão linear simples.) A *regressão linear múltipla*, por outro lado, utiliza mais de uma variável  $x$  para estimar o valor de  $y$ .

Neste capítulo, você vai ver como a regressão múltipla funciona e como aplicá-la na construção de um modelo para  $y$ . Vou mostrar a você todos os passos necessários para o processo, inclusive a definição de quais variáveis  $x$  incluir, como estimar suas contribuições para o modelo, como encontrar o melhor modelo e usá-lo para estimar  $y$ , além de como avaliar seu ajuste aos dados. Pode parecer muita coisa, mas você não vai regredir no tópico sobre regressão se, ao ler esse capítulo, der um passo de cada vez.

# ***Conhecendo o Modelo de Regressão Múltipla***

Antes de irmos direto ao uso do modelo de regressão múltipla, vamos conhecê-lo um pouco melhor. Nesta seção, quero lhe mostrar a utilidade da regressão múltipla, bem como os elementos básicos de um modelo de regressão múltipla. Alguns dos conceitos são apenas uma extensão do modelo de regressão simples (veja o Capítulo 4). Outros são um pouco mais complexos, como era de se imaginar, uma vez que esse modelo é mais complexo. Porém, os conceitos e resultados devem ter um sentido intuitivo, o que é sempre uma boa notícia.

## ***Descobrendo os usos da regressão múltipla***

Um dos contextos em que a regressão múltipla é bastante útil é quando a variável  $y$  é difícil de ser encontrada — ou seja, seu valor não pode ser medido de forma direta e você precisa de mais informações para descobrir qual será esse valor. Por exemplo, suponha que você queira estimar o preço do ouro hoje. Seria difícil imaginar que alguém seria capaz de fazer isso usando apenas uma outra variável. Talvez seja possível basear sua estimativa nos preços mais recentes, no preço de outras *commodities*, cujos valores acompanham o preço do ouro e em outras possíveis condições econômicas associadas a seu preço.

Outro caso em que a regressão múltipla pode ser usada é quando você deseja saber quais fatores influenciam a determinação do valor de  $y$ . Por exemplo, você quer saber quais informações são importantes para que um corretor de imóveis estabeleça o preço de uma casa a ser vendida.

## ***A fórmula geral do modelo de regressão múltipla***

O conceito geral da regressão linear simples é o de ajustar a melhor reta aos dados que você conseguiu e usá-la para fazer estimativas para  $y$  com base em determinados valores de  $x$ . A equação da reta de regressão linear simples é  $y = b_0 + b_1x_1$ , onde  $b_0$  é o intercepto  $y$ , e  $b_1$  é o coeficiente angular da reta. (Essa equação também pode ser escrita sob a forma de  $y = a + bx$ ; veja o Capítulo 4.)

Na regressão linear múltipla, entretanto, temos mais do que uma variável  $x$  relacionada a  $y$ . Vamos chamar essas variáveis de  $x_1$  e  $x_2 \dots x_k$ . No modelos mais básico de regressão múltipla, algumas ou todas essas variáveis  $x$  são utilizadas para estimar  $y$ , sendo que, nesse modelo, cada variável  $x$  é elevada à primeira potência. Esse processo tem como objetivo encontrar a melhor função linear que se ajuste ao conjunto de dados. Essa função linear assume a seguinte forma:  $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$ , e você pode chamá-la de *modelo de regressão (linear) múltipla*. Esse modelo é utilizado para estimar o valor de  $y$  com base nos valores atribuídos às variáveis  $x$ .



A função *linear* é uma equação cujos termos  $x$  são elevados apenas à primeira potência. Por exemplo,  $y = 2x_1 + 3x_2 + 24x_3$  é uma equação linear que utiliza três variáveis  $x$ . Se um dos termos  $x$  for elevado ao quadrado, a função passa a ser *quadrática*; se um termo  $x$  for elevado ao cubo, ela passa a ser uma função *cúbica*, e assim por diante. Neste capítulo, vou considerar apenas as funções lineares.

## ***Seguindo os passos rumo a uma análise***

Seu papel na realização de uma análise de regressão múltipla é fazer o seguinte (o computador pode ajudá-lo do terceiro ao sexto passos):

- 1. Faça uma lista das possíveis variáveis  $x$  que podem ajudá-lo a estimar  $y$ .**
- 2. Colete dados sobre a variável  $y$  e para as variáveis  $x$  a que se refere o primeiro passo.**
- 3. Verifique as relações entre cada variável  $x$  e  $y$  (usando diagramas de dispersão e correlações) e utilize os resultados para eliminar aquelas que não possuem uma forte relação com  $y$ .**
- 4. Observe possíveis relações entre as variáveis  $x$  para ter certeza de que não está sendo redundante (em linguagem estatística, tente evitar a multicolinearidade).**

Se duas variáveis  $x$  se relacionam a  $y$  da mesma forma, você não precisará de ambas em seu modelo.

- 5. Utilize as variáveis  $x$  (do quarto passo) na análise de regressão múltipla para encontrar o modelo que melhor se ajuste a seus dados.**
- 6. Substitua os valores atribuídos a  $x$  no modelo (encontrado no quinto passo) para prever  $y$ .**

Nas seções a seguir, detalho cada um dos passos descritos acima.

# Observando $x$ 's e $y$ 's

O primeiro passo rumo a uma análise de regressão múltipla vem antes dos cálculos numéricos feitos pelo computador; vem antes mesmo da coleta dos dados. O primeiro passo é dado quando você sente e pensa sobre as variáveis que poderiam ser úteis na previsão de sua variável de resposta  $y$ . Esse passo provavelmente é o que levará mais tempo, talvez com exceção do processo de coleta de dados. A decisão sobre quais variáveis  $x$  seriam candidatas a seu modelo representa um passo primordial, pois você não poderá voltar e coletar mais dados depois que a análise tiver acabado.



Sempre certifique-se de que sua variável de resposta  $y$ , e pelo menos uma das variáveis  $x$ , são quantitativas. Por exemplo, caso  $y$  não seja quantitativo, mas pelo menos um  $x$  for, você poderá usar um modelo de regressão logística (veja o Capítulo 8).

Suponha que você trabalhe no marketing de uma grande empresa nacional que vende TVs de plasma. Seu trabalho é vender o máximo de aparelhos que puder e, para isso, deve descobrir quais fatores influenciam a venda de TVs de plasma. Por causa das conversas com o departamento de propaganda e do que aprendeu em suas aulas na faculdade, você sabe que uma das formas mais poderosas de obter vendas é a propaganda. Assim, você pensa nos tipos de propagandas que poderiam estar relacionadas às vendas de TVs, e sua equipe tem duas ideias:

- ✓ **Propagandas na televisão:** claro, o que pode ser melhor do que vender uma TV através de uma propaganda na TV?
- ✓ **Anúncios no jornal:** publicá-los aos domingos. Quando os leitores estiverem lendo o jornal antes de assistirem ao jogo na televisão — geralmente, com o aparelho que possuem, perdem os detalhes das boas jogadas e não tiram as dúvidas dos erros terríveis cometidos pelos juízes —, perceberão que precisam de uma TV melhor.

Ao fazer uma lista com as possíveis variáveis  $x$  para prever  $y$ , você terá completado o primeiro passo para uma análise de regressão múltipla, segundo a lista apresentada na seção anterior. Observe que as três variáveis que uso no exemplo da TV são quantitativas (a propaganda na TV, os anúncios em jornais e a venda de TVs de plasma), o que significa que podemos prosseguir e pensar em um modelo de regressão múltipla usando os dois tipos de propagandas para prever as vendas de aparelhos de TV.

## Coletando Dados

O segundo passo no processo de análise de regressão múltipla é a coleta de dados para as variáveis  $x$  e  $y$ . Para fazer isso, certifique-se de que, para cada indivíduo no conjunto de dados, você colete todos os dados referentes a esse indivíduo ao mesmo tempo (incluindo o valor de  $y$  e todos os valores de  $x$ ) e mantenha-os juntos para cada indivíduo, preservada qualquer relação que possa existir entre as variáveis. Em seguida, insira os dados em uma tabela usando o Minitab ou qualquer outro software (em que cada coluna represente uma variável em cada linha represente todos os dados retirados de um único indivíduo) para dar uma olhada nos dados e organizá-los para futuras análises.

Para continuar com o exemplo das vendas de TV da seção anterior, suponha que você comece a pensar no montão de dados disponíveis sobre o mercado de aparelhos de TV de plasma. Você se lembra de ter trabalhado com o departamento de propaganda antes de fazer uma campanha publicitária usando, entre outras coisas, propagandas na TV e em jornais. Portanto, possui dados sobre essas variáveis retirados de uma variedade de pontos de venda. Sua amostra é formada por 22 pontos de vendas em diferentes partes mais os dados referentes à quantidade investida em cada tipo de propaganda com as vendas de TV daquele local. Os resultados estão na Tabela 5-1.

**Tabela 5-1                      Vendas e Gastos com Propaganda de Tvs de Plasma**

<i>Local</i>	<i>Vendas (Em Milhões de Reais)</i>	<i>Anúncios em jornais (Em milhares de Reais)</i>	<i>Propagandas na TV (Em milhares de Reais)</i>
1	9,73	0	20
2	11,19	0	20
3	8,75	5	5
4	6,25	5	5
5	9,10	10	10
6	9,71	10	10
7	9,31	15	15
8	11,77	15	15
9	8,82	20	5
10	9,82	20	5
11	16,28	25	25
12	15,77	25	25
13	10,44	30	0
14	9,14	30	0
15	13,29	35	5
16	13,30	35	5

17	14,05	40	10
18	14,36	40	10
19	15,21	45	15
20	17,41	45	15
21	18,66	50	20
22	17,17	50	20

Ao rever esses dados, a pergunta é: a quantia investida nessas duas formas de propaganda podem fazer uma boa estimativa para as vendas, ou seja, vale investir nessas propagandas? E, se a resposta for sim, é preciso incluir o gasto com os dois tipos de propaganda para estimar as vendas, ou um deles já é suficiente? Olhando os números na Tabela 5-1, vemos que as vendas mais altas, pelo menos, parecem se relacionar com as quantidades mais altas gastas com propagandas na TV; a situação com os anúncios em jornais não parece tão clara. Assim, o modelo de regressão múltipla deve conter as duas variáveis  $x$  ou apenas uma delas? Nas seções a seguir, você vai descobrir.

# Identificando Possíveis Relações

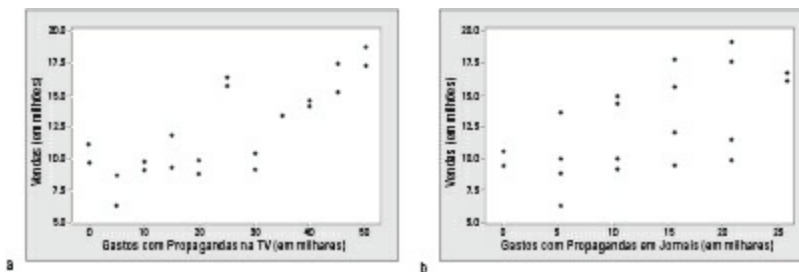
O terceiro passo para uma análise de regressão múltipla (veja a lista na seção “Seguindo os passos rumo a uma análise”) é descobrir quais (se houver alguma) de suas possíveis variáveis  $x$  realmente possui uma relação com  $y$ . Caso uma variável  $x$  não possua uma relação com  $y$ , sua inclusão no modelo não faz sentido algum. Os analistas de dados utilizam uma combinação de diagramas de dispersão e correlações para examinar as relações entre os pares de variáveis (como pode ser visto no Capítulo 4). Embora você veja essas duas técnicas como uma forma de procurar relações, nas próximas seções vou mostrar cada uma delas separadamente para discutir suas nuances.

## Construindo diagramas de dispersão

Em regressão linear múltipla, construímos diagramas de dispersão para entender se as possíveis variáveis  $x$  estão mesmo relacionadas à variável  $y$  estudada. Para investigar essas possíveis relações, você deve construir um diagrama de cada variável  $x$  com a variável  $y$ . Se tiver  $k$  variáveis  $x$  diferentes sendo consideradas para o modelo final, faça  $k$  diagramas diferentes.

Para fazer um diagrama de dispersão no Minitab, insira seus dados em colunas, sendo que cada coluna representa uma variável, e cada linha representa todos os dados de um indivíduo. Faça o caminho Graphs > Scatterplot > Simple. Selecione sua variável  $y$  à esquerda e clique em Select. Essa variável vai aparecer na caixa y-variable no lado direito. Em seguida, selecione sua variável  $x$  à esquerda e clique em Select. Essa variável vai aparecer na caixa x-variable no lado direito. Por fim, clique em OK.

Os diagramas para os gastos com propagandas de TV versus as vendas estão ilustrados na Figura 5-1.



**Figura 5-1:** Diagramas de dispersão para os gastos em propagandas na TV e em jornais versus as vendas de TVs de plasma.

É possível concluir, a partir da Figura 5-1a, que os gastos com TV realmente parecem ter uma relação bem forte com as vendas. Essa observação fornece evidências de que os gastos com anúncios na TV podem ser úteis para a estimativa das vendas de TV plasma. A Figura 5-1b demonstra uma relação linear entre os gastos com anúncios em jornais e as vendas, mas tal relação não é tão forte quanto a relação entre anúncios na TV e as vendas.



No entanto, ela ainda pode ser um pouco útil para a estimativa de vendas.

## Correlações: Examinando os vínculos

A segunda parte do terceiro passo envolve o cálculo e o exame das correlações entre as variáveis  $x$  e  $y$ . (É claro que, se o diagrama de dispersão de uma variável  $x$  em relação à variável  $y$  não demonstrar um padrão, você deve descartar essa variável  $x$  e não continuar o procedimento para encontrar a correlação).

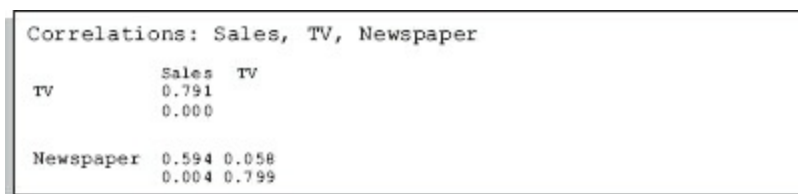
Sempre que empregar diagramas de dispersão para investigar possíveis relações lineares, as correlações normalmente não ficarão muito atrás. O *coeficiente de correlação* é o número que mede a força e a direção da relação linear entre duas variáveis quantitativas,  $x$  e  $y$ . (Veja no Capítulo 4 mais informações sobre correlação).

Este passo envolve duas partes:

- ✓ Encontrar e interpretar as correlações
- ✓ Testar as correlações para ver quais são estatisticamente significativas (ou seja, determinar quais variáveis  $x$  se relacionam de forma significativa com  $y$ )

### Encontrando e interpretando as correlações

Você pode calcular um conjunto de todas as possíveis correlações entre todos os pares de variáveis — processo chamado de *matriz de correlação* — no Minitab. Na Figura 5-2 você vê o resultado da matriz de correlação para os dados da TV organizados na Tabela 5-1. Observe as correlações entre a variável  $y$  (vendas, sales) e cada variável  $x$ , bem como a correlação entre as propagandas na TV e os anúncios em jornais (Newspaper).



Correlations: Sales, TV, Newspaper		
	Sales	TV
TV	0.791	0.000
Newspaper	0.594	0.058
	0.004	0.799

**Figura 5-2:** Correlação e valores- $p$  para o exemplo da venda de aparelhos de TV.

O Minitab pode encontrar uma matriz de correlação entre quaisquer pares de variáveis do modelo, incluindo a variável  $y$  e todas as variáveis  $x$ . Para calcular uma matriz de correlação para um grupo de variáveis no Minitab, primeiro insira os dados em colunas (uma para cada variável). Depois, siga o menu Stat>Basic Statistics>Descriptive Statistics>Correlation. Selecione, à esquerda, as variáveis para as quais quer a correlação e clique em Select.

Para encontrar os valores da matriz de correlação gerada pelo programa, cruze a linha e a colunas das variáveis para as quais quer encontrar a correlação: o número presente nessa

interseção representa a correlação entre essas duas variáveis. Por exemplo, a correlação entre as propagandas na TV e as vendas dos aparelhos de TV é 0,791, pois esse é o valor encontrado na interseção entre a linha das propagandas de TV e a coluna das vendas na matriz de correlação ilustrada na Figura 5-2.

## ***Testando a significância das correlações***

Como regra geral, estudada em Estatística I (e também revista no Capítulo 4), uma correlação próxima a 1 ou  $-1$  (começando por volta de  $\pm 0,75$ ) é forte; uma correlação próxima a 0 é muito fraca/inexistente; e uma ao redor de  $\pm 0,6$  a  $0,7$ , indica a presença de uma relação moderada. A correlação entre as propagandas na TV e as vendas dos aparelhos de TV é igual a 0,791, o que indica uma relação linear bastante forte entre essas duas variáveis, segundo a regra geral. A correlação entre os anúncios em jornais e as vendas dos aparelhos de TV vista na Figura 5-2 é 0,594, uma relação moderada.

Na maioria das vezes, em Estatística, essa abordagem é suficiente para interpretar o coeficiente de correlação. Porém, agora estamos falando de coisas mais sérias e você vai precisar de uma ferramenta mais precisa para determinar se o coeficiente de correlação é grande o bastante para ser significativo estatisticamente. Este é o verdadeiro teste para qualquer estatística: não se trata de saber se uma relação é forte ou moderada, mas se ela pode ou não ser generalizada para a população.

Nesta hora, o termo *estatisticamente significativo* deve entrar em cena. É o seu velho amigo, o teste de hipótese (veja o Capítulo 3 para uma revisão sobre testes de hipótese). Assim como existe o teste de hipótese para a média de uma população ou para a diferença entre as médias de duas populações, também existe um teste para a correlação entre duas variáveis dentro de uma população.


A hipótese nula para testar uma correlação é  $H_0: \rho = 0$  (ausência de relação) versus  $H_a: \rho \neq 0$  (existência de relação). A letra  $\rho$  é a versão grega de  $r$  e representa a verdadeira correlação entre  $x$  e  $y$  em toda a população;  $r$  é o coeficiente de correlação da amostra.

- ✓ **Se você não puder rejeitar  $H_0$  com base em seus dados**, não poderá concluir que a correlação entre  $x$  e  $y$  é diferente de zero, ou seja, não terá evidências suficientes para provar a existência de uma relação entre as duas variáveis e, portanto,  $x$  não deverá fazer parte do modelo de regressão múltipla.
- ✓ **Se você puder rejeitar  $H_0$  com base em seus dados**, concluirá que a correlação não é igual a zero e, portanto, as variáveis possuem uma relação. Mais do que isso, a relação entre elas é considerada estatisticamente significativa — ou seja, raramente a relação ocorreria em sua amostra por casualidade.

Qualquer software de estatística pode calcular o teste de hipótese para você. As fórmulas usadas nesse processo estão além do escopo deste livro. No entanto, a interpretação é a mesma para qualquer outro teste: se seu valor- $p$  for menor do que o valor predeterminado para  $\alpha$  (normalmente, 0,05), rejeite a  $H_0$  e conclua que existe uma relação entre  $x$  e  $y$ . Caso



contrário, a  $H_0$  não poderá ser rejeitada; com isso, conclui-se que não há evidências suficientes que indiquem uma relação entre as variáveis.



No Minitab, você pode conduzir um teste de hipótese para uma correlação clicando em Stat>Basic Statistics>Correlation e verificando o resultado no campo Display  $p$ -values. Selecione as variáveis para as quais quer as correlações e clique em Select. O resultado será apresentado na forma de uma pequena tabela que demonstra as correlações entre as variáveis para cada par com seu respectivo valor- $p$  em cada um. Na Figura 5-2, você vê o resultado da correlação para o exemplo das propagandas versus vendas.

Observando a Figura 5-2, vemos que a correlação de 0,791 entre as propagandas de TV e as vendas tem um valor- $p$  de 0,000 — ou seja, na verdade, ele é muito menor do que 0,001. Este é um resultado muito significativo, uma vez que é muito menor do que 0,05 (o nível predeterminado para  $\alpha$ ). Assim, o gasto com propagandas na TV possui uma relação muito forte com as vendas de aparelhos de TV. A correlação entre o gasto com anúncios em jornais e as vendas de aparelhos de TV foi 0,594, a qual também foi estatisticamente significativa, uma vez que seu valor- $p$  foi 0,004.

# Checando a Multicolinearidade

Você ainda tem um passo muito importante para completar no processo de investigação da relação antes de usar o modelo de regressão múltipla. Precisa completar o quarto passo: observar as relações entre as variáveis  $x$  e procurar alguma redundância. Se isso não for feito, você poderá ter problemas durante o processo de ajuste do modelo.



*Multicolinearidade* é o termo usado quando duas variáveis  $x$  (explicativas) são altamente correlacionadas. Incluir as duas variáveis relacionadas em um modelo de regressão múltipla não é apenas redundante, mas também problemático. A regra é a seguinte: se duas variáveis  $x$  possuírem uma correlação significativa, inclua apenas uma delas no modelo de regressão, e não as duas. Se você incluir as duas, o computador não saberá quais números fornecer como coeficientes para cada uma, já que ambas contribuem para determinar o valor de  $y$ . A multicolinearidade realmente pode bagunçar o processo de ajuste do modelo e fornecer respostas inconsistentes e que, com frequência, não se repetirão em estudos subsequentes.

Para evitar o problema da multicolinearidade, além de examinar as correlações entre as variáveis  $x$  e a variável de resposta  $y$ , encontre também as correlações entre todos os pares de variáveis  $x$ . Se duas variáveis  $x$  possuírem uma correlação alta, não deixe as duas no modelo, pois assim ocorrerá multicolinearidade. Para ver as correlações entre todas as variáveis  $x$ , faça com que o Minitab calcule a matriz de correlação de todas as variáveis. (Veja a seção “Encontrando e interpretando as correlações”.) Você pode ignorá-las entre a variável  $y$  e as variáveis  $x$  e escolher apenas as correlações entre as variáveis  $x$  mostradas na matriz de correlação. Encontre aquelas na interseção das linhas e colunas das variáveis  $x$  de interesse.



Se duas variáveis  $x_1$  e  $x_2$  possuírem uma correlação muito forte (ou seja, uma correlação maior do que  $+0,7$  ou  $-0,7$ ), então, qualquer uma delas pode estimar  $y$  tão bem quanto a outra, portanto, não é preciso incluir as duas no modelo. Se  $x_1$  e  $x_2$  não possuírem uma correlação forte, então as duas, trabalhando juntas, fariam estimativas melhores do que uma trabalhando sozinha.

Para o exemplo dos gastos com propagandas, você deve examinar a correlação entre as duas variáveis  $x$ , o gasto com propagandas na TV e com anúncios em jornais, para ter certeza de que não há multicolinearidade. A correlação entre essas duas variáveis (como pode ser visto na Figura 5-2) é apenas 0,058. Nem é preciso que um teste de hipótese lhe diga se essas variáveis se relacionam ou não; está evidente que elas não possuem uma relação.

O valor- $p$  para a correlação entre o gasto para os dois tipos de propagandas é 0,799 (veja a Figura 5-2), valor muito, mas muito maior do que 0,05, e, portanto, não é estatisticamente significativo. O alto valor- $p$  para a correlação entre o gasto com os dois tipos de propagandas confirma o fato de que as duas variáveis juntas podem ser úteis na estimativa de  $y$ , pois cada uma dá sua própria contribuição. Isso também mostra que manter as duas no

modelo não irá criar nenhum problema de multicolinearidade. (Isso encerra o quarto passo da análise de regressão múltipla, conforme listado na seção “Seguindo os passos rumo a uma análise”).)

# ***Encontrando o Modelo sob Medida para Duas Variáveis X***

Depois de obter um grupo de variáveis  $x$  que se relacionem a  $y$ , mas não se relacionem entre si (consulte a seção anterior), é hora de dar o quinto passo rumo à realização da análise de regressão múltipla (conforme listado na seção “Seguindo os passos rumo a uma análise”). Você está pronto para encontrar o modelo que melhor se ajusta a seus dados.

No modelo de regressão múltipla com duas variáveis  $x$ , você tem a equação geral  $y = b_0 + b_1x_1 + b_2x_2$  e já sabe quais variáveis  $x$  incluir no modelo (segundo o que foi realizado no quarto passo na seção anterior); a tarefa agora é descobrir os coeficientes (números) a serem colocados no lugar de  $b_0$ ,  $b_1$  e  $b_2$ , para que, assim, você possa usar a equação resultante para estimar  $y$ . Esse modelo em específico é o *modelo de regressão linear múltipla*. Esta seção vai ensiná-lo a obter, interpretar e testar esses coeficientes a fim de completar o quarto passo na jornada pela análise de regressão múltipla.



Encontrar a equação linear mais bem ajustada é como encontrar a reta de regressão na regressão linear simples, exceto pelo fato de que você não está procurando uma reta. Quando se tem duas variáveis  $x$  em uma regressão múltipla, por exemplo, estima-se o plano mais ajustado aos dados.

## ***Obtendo os coeficientes de regressão múltipla***

No modelo de regressão linear simples, temos a reta  $y = b_0 + b_1x$ , onde o coeficiente de  $x$  é a inclinação (coeficiente angular) e representa a mudança ocorrida em  $y$  por unidade alterada em  $x$ . No modelo de regressão linear múltipla, os coeficientes  $b_1$ ,  $b_2$ , ... e assim por diante, quantificam, de forma semelhante, a contribuição única de cada variável  $x$  correspondente ( $x_1$ ,  $x_2$ ) à previsão de  $y$ . O coeficiente  $b_0$  indica a quantidade pela qual todos esses valores devem ser ajustados a fim de obter um ajuste final aos dados (assim como o intercepto  $y$  faz na regressão linear simples).

O software faz todo o trabalho necessário para encontrar os coeficientes ( $b_0$ ,  $b_1$  etc.) que melhor se ajustam aos dados. Os coeficientes estabelecidos pelo Minitab para criar o modelo que mais se ajusta são os que, como um grupo, minimizam a soma dos resíduos ao quadrado (mais ou menos como a variância nos dados relativos ao modelo escolhido). As equações para obter esses coeficientes à mão são muito complexas para serem incluídas neste livro, e o computador pode fazer todo esse trabalho para você. No Minitab, os coeficientes aparecem no resultado para a regressão. Você encontra os coeficientes de regressão múltipla ( $b_0$ ,  $b_1$ ,  $b_2$ , ...,  $b_k$ ) sob a coluna nomeada COEF.



Para realizar uma análise de regressão múltipla no Minitab, clique em Stat>Regression>Regression. Em seguida, selecione a variável de resposta ( $y$ ) e clique em Select. Depois, selecione suas variáveis preditoras (variáveis  $x$ ) e clique em Select. Clique em OK, e o computador fará a análise.

Para o exemplo das vendas de TVs de plasma das seções anteriores, a Figura 5-3 mostra os coeficientes de regressão múltipla na coluna COEF para o modelo. O primeiro coeficiente (5,257) é simplesmente o termo constante (ou  $b_0$ ) no modelo e não está associado a nenhuma variável  $x$ . Essa constante passa brevemente pela análise; é o valor que você coloca no final para fazer com que os cálculos deem certo. O segundo coeficiente na coluna COEF é 0,162; esse valor é o coeficiente de  $x_1$  (gastos com propagandas na TV), também conhecido como  $b_1$ . O terceiro coeficiente na coluna COEF é 0,249; que é o valor para  $b_2$  no modelo de regressão múltipla e é o coeficiente que acompanha  $x_2$  (gasto com propagandas em jornais).

```

The regression equation is
Sales = 5.267 + 0.162 TV ads + 0.249 Newsp ads

Predictor    Coef    SE Coef    T    P
Constant     5.2574    0.4984    10.55 0.000
TV ads       0.16211    0.01319    12.29 0.000
Newsp ads    0.24887    0.02792     8.91 0.000
S = 0.976613    R-Sq = 92.8%    R-Sq(adj) = 92.0%

```

**Figura 5-3:** Resultado da regressão para o exemplo das propagandas versus a venda de aparelhos de TV.

Colocando esses coeficientes na equação de regressão múltipla, você vê que a equação de regressão é  $\text{Vendas} = 5,267 + 0,162 (\text{propaganda na TV}) + 0,249 (\text{propaganda em jornais})$ , onde as vendas estão em milhões de reais, e os gastos com propagandas em milhares de reais.

Bom, então você tem os coeficientes (nem foi tão difícil, não é?). Mas, e agora? O que tudo isso quer dizer? A próxima seção vai guiar sua interpretação.

## *Interpretando os coeficientes*

Na regressão linear simples (abrangida no Capítulo 4), os coeficientes representam a inclinação e o intercepto  $y$  da reta de regressão, e sua interpretação é bem simples. Esta inclinação (coeficiente angular) em particular representa a mudança em  $y$  ocorrida em virtude do aumento de uma unidade em  $x$ , pois o coeficiente angular pode ser escrito como um número sobre 1.

No modelo de regressão múltipla, a interpretação é um pouco mais complexa. Em virtude de todos os fundamentos matemáticos em que o modelo se apoia e como ele é finalizado (acredite em mim, você não vai querer saber isso, a não ser que esteja interessado em ser um PhD em Estatística), os coeficientes possuem um significado diferente.

O coeficiente de uma variável  $x$  em um modelo de regressão múltipla é o valor pelo qual  $y$  se altera se aquela variável  $x$  aumentar em uma unidade e os valores de todas as outras variáveis  $x$  do modelo *não se alterarem*. Assim, basicamente, o que se observa é a



contribuição marginal de cada variável  $x$  quando as outras variáveis  $x$  do modelo são mantidas constantes.

Na análise de regressão das vendas versus propagandas (veja a Figura 5-3), o coeficiente de  $x_1$  (gasto com propagandas na TV) é igual a 0,16211. Portanto,  $y$  (vendas de TVs de plasma) aumenta em 0,16211 milhões de reais quando os gastos com propagandas na TV aumentam em 1,0 mil reais, e os gastos com anúncios em jornais não se alteram. (Observe que manter um grande número de algarismos depois da vírgula decimal reduz o erro de arredondamento quando falamos em milhões.)



É mais fácil interpretar o número “0,16211 milhões de reais” se você convertê-lo para um valor em reais que não tenha a vírgula decimal: R\$0,16211 milhões é igual a R\$162.110. (Para chegar a esse valor, eu apenas multipliquei R\$0,16211 por 1.000.000). Sendo assim, as vendas de TVs de plasma aumentam em R\$162.110 a cada R\$1.000 de aumento nos gastos com as propagandas na TV e quando os gastos com anúncios em jornais permanecem iguais. Da mesma forma, o coeficiente de  $x_2$  (gastos com anúncios em jornais) é igual a 0,24887. Sendo assim, as vendas de TVs de plasma aumentam em 0,24887 milhões de reais (ou R\$248.87) a cada R\$1.000 de aumento no gasto com anúncios em jornais e quando os gastos com as propagandas na TV permanecem iguais.



Não se esqueça das unidades de cada variável em uma análise de regressão múltipla. Esse é um dos erros mais comuns em Estatística II. Se você esquecesse as unidades do exemplo, pensaria que as vendas aumentariam em 0,24887 reais a cada R\$1 gasto com anúncios em jornais!

Conhecendo os coeficientes de regressão múltipla ( $b_1$  e  $b_2$ , nesse caso) e sua interpretação, você pode responder a pergunta original: vale a pena gastar esse dinheiro com propagandas na TV e em jornais? A resposta é um *sim* em alto e bom som! E não é só isso, você também pode dizer o quanto espera que as vendas aumentem a cada R\$1.000 gastos com publicidade na TV e em jornais. Observe que essa conclusão presume que o modelo se ajusta bem aos dados. Você já tem algumas evidências a respeito disso, providenciadas por diagramas de dispersão e testes de correlação, porém, é preciso fazer mais algumas verificações antes de correr até sua gerente e lhe dar as boas notícias. A seção a seguir vai lhe dizer qual é o próximo passo.

## ***Testando os coeficientes***

Para determinar oficialmente se você tem as variáveis  $x$  corretas em seu modelo de regressão múltipla, faça um teste de hipótese formal para ter certeza de que os coeficientes não são iguais a zero. Observe que, se o coeficiente de uma variável  $x$  for igual a zero, ao colocá-lo no modelo, você terá zero vezes aquela variável  $x$ , o que é igual a zero. Tal resultado quer dizer que, caso o coeficiente de uma variável seja igual a zero, você não precisará dessa variável em seu modelo.





Com qualquer análise de regressão, o computador automaticamente realiza todos os testes de hipótese necessários para os coeficientes de regressão. Junto aos coeficientes encontrados na saída do software, você também vê a estatística de teste e os valores- $p$  para um teste de cada um desses coeficientes na mesma linha; sendo que cada um está testando  $H_0$ : Coeficiente = 0 versus  $H_a$ : Coeficiente  $\neq$  0.



A fórmula geral para encontrar uma estatística de teste na maioria das situações é pegar a estatística (nesse caso o coeficiente), subtrair o valor de  $H_0$  (zero) e dividi-lo pelo erro padrão daquela estatística (para esse exemplo, o erro padrão do coeficiente). (Para mais informações sobre a fórmula geral dos testes de hipótese, veja o Capítulo 3.)

Para testar um coeficiente de regressão, a estatística de teste (usando os nomes presentes na Figura 5-3), faça  $(\text{Coef} - 0)/\text{SE Coef}$ . Em linguagem não computacional, isso significa que você vai pegar o coeficiente, subtrair-lo de zero e dividir o resultado pelo erro padrão (SE) do coeficiente. O erro padrão do coeficiente é a medida da sua variação esperada quando uma nova amostra for coletada. (Consulte o Capítulo 3 para mais informações sobre erro padrão.)

A estatística de teste tem uma distribuição- $t$  com  $n - k - 1$  graus de liberdade, onde  $n$  é igual ao tamanho amostral, e  $k$  é o número de preditoras (variáveis  $x$ ) no modelo. Esse número de graus de liberdade funciona para qualquer coeficiente no modelo (você não precisa se incomodar em fazer um teste para a constante, pois não há nenhuma variável  $x$  associada a ela).

As estatísticas de teste para cada coeficiente estão listadas na coluna nomeada pela letra T (pois possui uma distribuição- $t$ ) na saída do Minitab. Você deve comparar o valor da estatística de teste com a distribuição- $t$  com  $n - k - 1$  graus de liberdade (usando a Tabela A-1 no apêndice) e, assim, obter seu valor- $p$ . Se o valor- $p$  for menor do que o valor predeterminado para  $\alpha$  (normalmente, 0,05), então, rejeite a  $H_0$  e conclua que o coeficiente dessa variável  $x$  não é zero e que sua contribuição para a estimativa de  $y$  é significativa (sabendo que as outras variáveis também estão incluídas no modelo). Se o valor- $p$  for maior que 0,05, você não pode rejeitar  $H_0$ , pois essa variável  $x$  não faz nenhuma contribuição significativa para a estimativa de  $y$  (quando as outras variáveis estão incluídas no modelo).

No caso do exemplo das propagandas e das vendas de TVs de plasma, a Figura 5-3 demonstra que o coeficiente para as propagandas na TV é 0,1621 (o segundo número na coluna dois). O erro padrão é dado como 0,0132 (o segundo número na coluna três). Para encontrar a estatística de teste para as propagandas na TV, subtraia 0,1621 de zero e divida o resultado pelo erro padrão, 0,0132. Você vai obter o valor de  $t = 12,29$ , que é o segundo número na coluna quatro. Comparando esse valor de  $t$  à distribuição- $t$  com  $n - k - 1 = 22 - 2 - 1 = 19$  graus de liberdade (Tabela A-1 no apêndice), você vai notar que o valor de  $t$  está fora da escala. Isso quer dizer que o valor- $p$  é menor do que o que pode ser medido pela tabela- $t$ . O Minitab lista o valor- $p$  na coluna cinco da Figura 5-3 como sendo 0,000 (quer dizer, menor do que 0,001). Esse resultado leva você a concluir que o coeficiente

para as propagandas na TV é estatisticamente significativo e, portanto, estas devem ser incluídas no modelo para a previsão das vendas de aparelhos de TV.

Seguindo o mesmo raciocínio, o coeficiente dos anúncios em jornais também é significativo, uma vez que apresenta um valor- $p$  de 0,000; esses resultados se encontram em toda a linha de anúncios de jornal da Figura 5-3. Com base em seus testes de coeficientes e na ausência de multicolinearidade entre as propagandas na TV e em jornais (veja a seção anterior “Evitando a Multicolinearidade”), você deve incluir tanto a variável das propagandas na TV quanto a dos anúncios em jornais no modelo que estima as vendas de aparelhos de TV de plasma.

# Previendo y Através das Variáveis x

Quando você tiver seu modelo de regressão múltipla, finalmente estará pronto para completar o sexto passo da análise de regressão múltipla: prever o valor de  $y$ , dado um conjunto de valores para as variáveis  $x$ . Para fazer essa previsão, pegue esses valores de  $x$ , para os quais quer prever  $y$ , substitua-os no modelo de regressão múltipla e simplifique.

No exemplo das vendas de TVs de plasma versus propagandas (veja análise na Figura 5-3), o modelo que mais se ajusta é  $y = 5,26 + 0,162x_1 + 0,249x_2$ . No contexto do problema, o modelo é Vendas =  $5,26 + 0,162$  propaganda na TV ( $x_1$ ) +  $0,249$  anúncios em jornais ( $x_2$ ).



Lembre-se de que a unidade para as vendas de TVs de plasma está em milhões de reais, e a unidade para os gastos com publicidade na TV e em jornais está em milhares de reais. Ou seja, R\$20.000 gastos em propagandas na TV significa que  $x_1 = 20$ , no modelo. Da mesma forma, R\$10.000 gastos em propagandas em jornais significa que  $x_2 = 10$ , no modelo. O esquecimento das unidades envolvidas pode causar erros graves de cálculos.

Suponha que você queira estimar as vendas de TVs de plasma gastando R\$20.000 em propagandas na TV e R\$10.000 em anúncios em jornais. Substitua  $x_1 = 20$  e  $x_2 = 10$  no modelo de regressão múltipla e você vai obter  $y = 5,26 + 0,162(20) + 0,249(10) = 10,99$ . Isto é, se você gastar R\$20.000 com propagandas na TV e R\$10.000 com anúncios em jornais, a estimativa de vendas é R\$10.990 milhões.

Ao menos, essa estimativa está de acordo com os dados retirados dos 22 pontos de venda mostrados na Tabela 5-1. O ponto de venda 10 gastou R\$20.000 com propaganda na TV e R\$5.000 com publicidade em jornais (menos do que estimamos) e vendeu R\$9.820 milhões. O ponto de venda 11 gastou um pouco mais com propagandas na TV e muito mais com publicidade em jornais e alcançou o valor de R\$16.280 milhões em vendas. Sua estimativa para as vendas dos pontos 10 e 11 são  $5,26 + 0,162 * 20 + 0,249 * 5 = R\$9.745$  milhões e  $5,26 + 0,162 * 25 + 0,249 * 25 = R\$15.535$  milhões, respectivamente. Essas estimativas se mostraram muito próximas das vendas reais alcançadas por esses dois pontos de venda (R\$9.820 milhões e R\$16.280 milhões, respectivamente, como mostrado na Tabela 5-1), o que, pelo menos, nos faz confiar que suas estimativas serão próximas, também, para os outros pontos de venda não escolhidos para o estudo.



Tenha o cuidado de atribuir às variáveis  $x$  apenas valores que estejam dentro do conjunto de dados. Isto é, a Tabela 5-1 mostra dados para o gasto com propagandas na TV entre R\$0 e R\$50.000; os gastos com anúncios em jornais vão de R\$0 a R\$25.000. Assim, não seria apropriado tentar estimar a venda de TVs usando valores de R\$75.000 para gastos com anúncios na TV e de R\$50.000 para anúncios em jornais, respectivamente, pois o modelo de regressão obtido se ajusta apenas aos dados coletados. Não há como saber se essa mesma relação continua fora dessa área. Essa proibição, que consiste na estimativa de  $y$  a partir de valores de  $x$  fora do conjunto de dados, se chama *extrapolação*. Como um de

meus amigos diz: “Amigos não deixam amigos extrapolarem.”

# *Verificando o Ajuste do Modelo de Regressão Múltipla*

Antes de correr para seu chefe dizendo que conseguiu resolver a questão de como estimar as vendas de TVs de plasma, primeiro, você tem que ter certeza de que colocou todos os pingos nos is e cortou todos os t's, da mesma forma que faria com qualquer outro procedimento estatístico. Nesse caso, você deve checar as condições do modelo de regressão múltipla. Tais condições estão, principalmente, focadas nos *resíduos* (a diferença entre os valores estimados para  $y$  e os valores de  $y$  observados a partir de seus dados). Se o modelo estiver próximo dos dados reais coletados, você pode estar relativamente seguro de que, se tiver que coletar mais dados, estes também se alinharão ao modelo e suas previsões também serão boas.

Nesta seção, você vai ver quais são as condições para a regressão múltipla e as técnicas específicas usadas pelos estatísticos para checar cada uma dessas condições. O personagem principal de todo esse processo de checagem de condições é o resíduo.

## *Observando as condições*

As condições para a regressão múltipla se concentram no termo erro, ou resíduos. Os resíduos são os valores que sobram depois que o modelo é ajustado. Eles representam a diferença entre o valor real de  $y$ , observado no conjunto de dados, e o valor de  $y$  estimado, com base no modelo. A seguir, veja as condições para os resíduos do modelo de regressão múltipla; observe que todas as condições precisam ser satisfeitas antes de se prosseguir com o modelo:

- ✓ Possuem uma distribuição normal com média zero.
- ✓ Possuem a mesma variância para cada valor (previsto) de  $y$  ajustado.
- ✓ São independentes (ou seja, não influenciam uns aos outros).

## *Traçando um plano para checar as condições*

Pode parecer que você tem uma tonelada de coisas para checar aqui e ali, mas, felizmente, o Minitab lhe dá toda a informação necessária em uma série de quatro gráficos, todos apresentados de uma vez só. Estes são os chamados *gráficos de resíduos* e mostram os resíduos de forma que você possa verificar se as condições mencionadas acima foram atendidas.

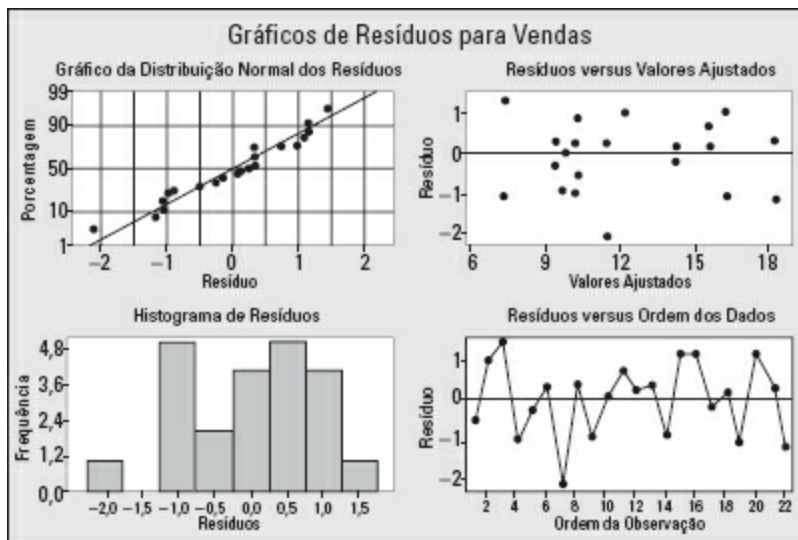
Você ainda pode ter o conjunto de gráficos de resíduos em dois sabores:

- ✓ **Resíduos regulares:** Os gráficos de resíduos regulares (os de sabor baunilha) lhe mostram exatamente quais são os resíduos para cada valor de  $y$ . Sua unidade depende das variáveis no modelo; utilize-os *só* se seu principal objetivo for procurar padrões nos dados. A Figura 5-4 ilustra o gráfico de resíduos regular para o exemplo da venda de TVs de plasma. A unidade desses resíduos é milhões de

reais.

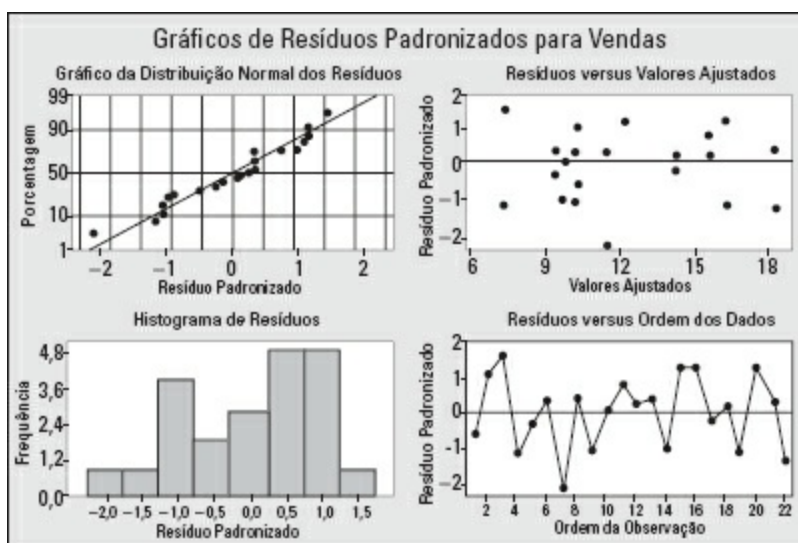
- ✓ **Resíduos padronizados:** Os gráficos residuais padronizados (os de sabor morango) convertem cada resíduo em um escore-Z, subtraindoos da média e dividindo o resultado pelo desvio padrão de todos os resíduos. A Figura 5-5 ilustra o gráfico de resíduos padronizado para o exemplo da venda de TVs de plasma. Utilize esses gráficos caso queira não apenas procurar padrões nos dados, mas também, avaliar os valores dos resíduos padronizados em relação à distribuição-Z, a fim de detectar a presença de outliers. (A maioria dos estatísticos utiliza os gráficos de resíduos padronizados.)

Note que os gráficos na Figura 5-5 são quase iguais aos da Figura 5-4. Não é uma surpresa que as formas de todos os gráficos sejam iguais para os dois tipos de resíduos. Observe, no entanto, que os valores dos resíduos regulares na Figura 5-4 estão em milhões de reais e os resíduos padronizados na Figura 5-5 vêm da distribuição normal padrão, que não possui unidades.



**Figura 5-4:** Gráficos de resíduos para o exemplo das propagandas e venda de aparelhos de TV.

Para construir gráficos de resíduos no Minitab, faça o caminho Stat>Regression>Regression. Selecione sua variável de resposta ( $y$ ) e suas variáveis preditoras ( $x$ ). Clique em Graphs e escolha entre Regular ou Standardized para os resíduos, dependendo do que deseja. Em seguida, clique em Four-in-one, o que indica que você quer os quatro gráficos mostrados na Figura 5-4 (usando os resíduos regulares) e na Figura 5-5 (usando os resíduos padronizados).



**Figura 5-5:** Gráficos de resíduos padronizados para o exemplo das propagandas e venda de aparelhos de TV.

## Verificando as três condições

As seções a seguir ensinam como verificar os resíduos para ver se seu conjunto de dados satisfaz as três condições do modelo de regressão múltipla.

### *Satisfazendo a primeira condição: Distribuição normal com média zero.*

A primeira condição a ser satisfeita é a de que os resíduos tenham uma distribuição normal com média zero. O gráfico no canto superior esquerdo da Figura 5-4 mostra o quão bem os resíduos se adequam a uma distribuição normal. Resíduos caindo sobre uma reta indicam que a condição de normalidade foi satisfeita. Pela aparência desse gráfico, eu diria que a condição foi satisfeita para o exemplo em questão.

O gráfico no canto superior direito da Figura 5-4 mostra a aparência dos resíduos para os vários valores estimados para  $y$ . Note a linha horizontal que corta o gráfico: ela tem o zero como marco. A média dos resíduos deve se encontrar nessa linha (zero). O gráfico dos resíduos versus valores ajustados verifica a condição da média zero e vale para os anúncios e exemplos de venda apresentados na Figura 5-4.

Uma alternativa para a verificação da normalidade, além do uso dos resíduos regulares, é observar os resíduos padronizados (veja a Figura 5-5) e analisar o gráfico no canto superior direito. Ele mostra como os resíduos estão distribuídos ao longo de vários valores estimados (ajustados) para  $y$ . Os resíduos padronizados devem acompanhar uma distribuição normal, ou seja, devem ter média zero e desvio padrão igual a um. Portanto, quando olhar os resíduos padrão, eles devem estar centralizados ao redor de zero de forma que não haja um padrão previsível, com a mesma quantidade de variabilidade em torno da linha horizontal que cruza o marco zero à medida que você se movimenta da esquerda para a direita.





Ao analisar o gráfico no canto superior direito da Figura 5-5, você também deve observar que a maioria dos resíduos padrão (95%) estão dentro de dois desvios padrão em relação à média, que, nesse caso, é  $-2$  a  $+2$  (graças à regra 68-95-99,7 — lembra dela?). Você deve ver mais resíduos rondando o zero (onde a massa estaria em uma distribuição normal padrão) e, à medida que se afastar do zero, deve ver cada vez menos resíduos. O gráfico no canto superior direito da Figura 5-5 confirma uma distribuição normal para o exemplo em questão sob todos os aspectos aqui mencionados.

Os gráficos na parte inferior esquerda das Figuras 5-4 e 5-5 mostram histogramas de resíduos regulares e padronizados, respectivamente. Esses histogramas devem refletir uma distribuição normal; sua forma deve ser quase simétrica e se parecer com uma curva em forma de sino. Caso o conjunto de dados seja pequeno (como no caso aqui, em que há apenas 22 observações), o histograma pode não ficar tão próximo da normalidade quanto você gostaria; nesse caso, considere-o como parte do corpo de evidências que os quatro gráficos lhe mostram. Os histogramas ilustrados nos cantos inferiores esquerdos das Figuras 5-4 e 5-5 não têm uma aparência muito normal; entretanto, uma vez que você não vê nenhum problema evidente nos gráficos da parte superior, não se preocupe!

### ***Satisfazendo a segunda condição: Variância***

A segunda condição a ser checada é a de que os resíduos tenham a mesma variância para cada valor ajustado (previsto) de  $y$ . Observe mais uma vez o gráfico no canto superior direito da Figura 5-4 (ou da Figura 5-5). Você não deve notar nenhuma mudança na quantidade de dispersão (variabilidade) nos resíduos ao redor daquela linha horizontal à medida que se movimenta da esquerda para a direita. Analisando o gráfico na parte superior direita da Figura 5-4, não encontramos razão para dizer que a condição número dois não foi satisfeita.

O que nos faria dar um cartão vermelho para a segunda condição é se os resíduos se difundissem ou se sua dispersão aumentasse à medida que você se movimentar da esquerda para a direita no gráfico ilustrado na parte superior direita. Essa difusão significaria que a variabilidade aumenta gradativamente para os valores cada vez mais altos previstos para  $y$  e, portanto, a condição de uma mesma variabilidade ao redor da reta ajustada não seria satisfeita e, neste caso, o modelo de regressão não se ajustaria.

### ***Verificando a terceira condição***

A terceira condição é a de que os resíduos sejam independentes, ou seja, não influenciem uns aos outros. Analisando o gráfico no canto inferior direito, tanto na Figura 5-4, quanto na 5-5, vemos os resíduos traçados de acordo com o número de observação, ou seja, na ordem em que os dados aparecem na amostra. Se você notar um padrão, temos um problema; por exemplo, se ligássemos os pontos, por assim dizer, poderíamos ver o padrão de uma reta, de uma curva ou de qualquer tendência ascendente ou descendente. Nos gráficos da parte inferior direita, não é possível observar nenhum tipo de padrão, sendo





assim, a condição de independência está satisfeita para esse exemplo.



Se os dados devem ser coletados ao longo de um período, assim como o preço de ações ao longo de um período de dez anos, a condição de independência pode se tornar um grande problema, pois os dados do período anterior podem estar relacionados aos dados do próximo período. Esse tipo de dado requer uma análise de série temporal, o que vai além do escopo deste livro.

## Capítulo 6

# Como Vou Sentir Sua Falta se Você Não Sair? Escolha do Modelo de Regressão

### *Neste Capítulo*

- ▶ Avaliando os diferentes métodos de escolha de um modelo de regressão múltipla
- ▶ Entendendo como funcionam os métodos de seleção forward e backward
- ▶ Usando os melhores métodos de subconjunto para encontrar um bom modelo

**S**uponha que você esteja tentando estimar uma variável quantitativa,  $y$ , e tem muitas variáveis  $x$  à sua disposição. São tantas as variáveis relacionadas a  $y$  que, na verdade, você se sente como eu me sinto todos os dias em meu trabalho — pressionada pela quantidade de oportunidades. Para onde ir? O que fazer? Não se preocupe mais, este capítulo foi feito para você.

Neste capítulo, você vai descobrir os critérios para determinar quando um modelo se ajusta bem. Discuto os diferentes procedimentos para a seleção de modelos e todos os detalhes dos métodos aprovados pelos estatísticos para a seleção do melhor deles. Além disso, você vai conhecer os fatores que entram em cena quando um punter chuta a bola. (Comece a pensar nisso enquanto estiver lendo.)



Observe que o termo *melhor* possui aqui muitas conotações. Não existe um único melhor modelo a ser obtido por todos como resposta final. Isto é, cada analista de dados pode optar por um modelo diferente e, ainda assim, cada um deles pode funcionar muito bem na hora de fazer as previsões para  $y$ .

# ***Dando o Pontapé Inicial na Estimativa para a Distância de um Punt<sup>1</sup>***

Antes de saltar sobre um procedimento de seleção de modelo para prever  $y$  usando um conjunto de variáveis  $x$ , é preciso trabalhar as pernas. A variável de interesse é  $y$ , e essa é dada de mão beijada. Mas de onde vêm as variáveis  $x$ ? Como escolher as que devem ser investigadas como possíveis candidatas para o modelo de previsão de  $y$ ? E como essas possíveis variáveis  $x$  interagem umas com as outras na hora de fazer a previsão?

Você deve responder a todas essas perguntas antes de usar qualquer procedimento de seleção de um modelo. Entretanto, essa é a parte mais desafiadora e mais divertida; afinal, o computador não pode pensar nas variáveis  $x$  por você!

Suponha que você esteja em um jogo de futebol americano e o time adversário vai executar um punt (chute de devolução). Você vê o jogador se alinhando e se preparando para chutar a bola e, então, algumas perguntas lhe ocorrem: “Ai, queria saber a que distância esse chute vai chegar. Quais serão os fatores que influenciam a distância alcançada por um punt? Será que posso usar esses fatores em um modelo de regressão múltipla para tentar estimar a distância de um punt? Hum, acho que vou consultar meu *Estatística II Para Leigos* sobre isso e analisar alguns dados durante o intervalo...”

Bom, talvez pareça meio forçado, mas, ainda assim, é uma linha de questionamento interessante para jogadores de futebol americano, golfe, futebol e, até mesmo, beisebol. Todos buscam mais distância e uma forma de consegui-la.

Nas seções a seguir, você vai ver como identificar e avaliar as diferentes variáveis  $x$  com relação a seu potencial de contribuição para prever  $y$ .

## ***Fazendo o brainstorm das variáveis e coletando os dados***

Começar do zero, tentando pensar em um conjunto de variáveis  $x$  que possam se relacionar a  $y$ , pode parecer uma tarefa árdua, mas, na realidade, não é tão ruim quanto você pode estar imaginando. A maioria dos pesquisadores interessados em prever uma variável  $y$ , em primeiro lugar, fazem uma ideia de quais variáveis podem se relacionar a  $y$ . Depois de montar um conjunto de possibilidades lógicas para  $x$ , você deve coletar dados sobre essas variáveis, bem como sobre  $y$ , para ver qual é a verdadeira relação entre elas.

O Virginia Polytechnic Institute, nos Estados Unidos, conduziu um estudo para tentar estimar a distância de um punt no futebol americano (coisa com a qual os torcedores do Ohio State não estão familiarizados). As possíveis variáveis que, segundo eles, poderiam estar relacionadas à distância alcançada por um punt incluíam:

- ✓ Hang time (tempo, em segundos que a bola fica “suspensa” no ar)
- ✓ A força da perna direita (medida em quilogramas de força)

- ✓ A força da perna esquerda (em quilogramas de força)
- ✓ A flexibilidade da perna direita (em graus)
- ✓ A flexibilidade da perna esquerda (em graus)
- ✓ A força total das pernas (em quilogramas)

Os dados coletados em uma amostra composta por 13 punts (executados por jogadores destros) são mostrados na Tabela 6-1.

**Tabela 6-1                      Dados Coletados para Estudo sobre a Distância de um Punt**

<i>Distância (Em metros)</i>	<i>Hang Time</i>	<i>Força Perna Direita</i>	<i>Força Perna Esquerda</i>	<i>Flexibilidade Perna Direita</i>	<i>Flexibilidade Perna Esquerda</i>	<i>Força Total das Pernas</i>
49,5	4,75	77	77	106	106	109,12
43,9	4,07	64	59	92	93	89,67
45,0	4,04	82	77	93	78	69,40
49,8	4,18	73	73	103	93	89,40
58,2	4,35	77	68	104	93	120,91
52,3	4,16	68	68	101	87	118,19
49,4	4,43	77	82	108	106	99,45
32,0	3,20	50	50	86	92	60,18
32,2	3,02	54	50	90	86	59,08
35,8	3,64	59	54	85	80	93,39
42,7	3,68	54	64	89	83	69,82
45,8	3,60	64	59	92	94	70,14
50,3	3,85	73	68	95	95	109,12

Outras variáveis que podem estar relacionadas à distância atingida em um punt podem incluir a direção e a velocidade do vento no momento do chute, o ângulo em que a bola foi posicionada, a distância média alcançada pelos punts já executados por um determinado punter, se o jogo está sendo em casa ou no campo do adversário, e outras. No entanto, esses pesquisadores parecem ter informações suficientes em mãos para construir um modelo que estime a distância de um punt.

Pelo bem da simplicidade, você pode supor que o chutador é destro, o que nem sempre é o caso, mas representa a maioria esmagadora de chutadores.

Analisando apenas o conjunto de dados brutos na Tabela 6-1, você não consegue saber quais variáveis, caso haja alguma, se relacionam à distância do punt ou como essas variáveis podem se relacionar a ela. Você vai precisar de mais análises para compreender isso melhor.

# *Examinando diagramas de dispersão e correlações*



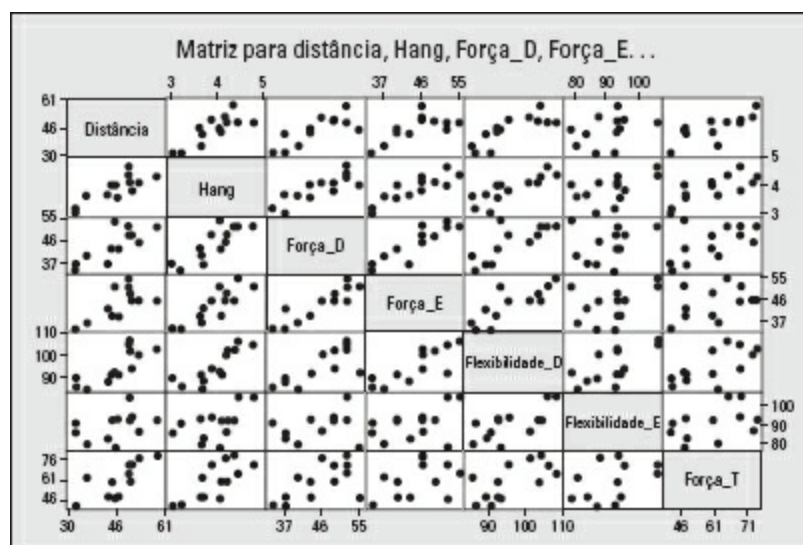
Depois de ter identificado um conjunto de possíveis variáveis  $x$ , o próximo passo é descobrir quais dessas variáveis estão altamente relacionadas a  $y$ , a fim de começar a peneirar o conjunto de possíveis candidatas ao modelo final. No exemplo da distância do punt, o objetivo é ver quais das seis variáveis presentes na Tabela 6-1 possuem uma relação realmente forte com a distância atingida em um punt. As duas formas de observar essas relações são:

- ✓ **Diagrama de dispersão:** uma técnica gráfica
- ✓ **Correlação:** medida da relação linear entre as duas variáveis compostas por um único número.

## *Observando relações através dos diagramas de dispersão*

Para começar a examinar as relações entre as variáveis  $x$  e  $y$ , utilize uma série de diagramas de dispersão. A Figura 6-1 mostra todos os diagramas de dispersão — não apenas para cada variável  $x$  com  $y$ , mas também para cada variável  $x$  com as outras variáveis  $x$ . Os diagramas estão representados na forma de uma *matriz*, uma tabela composta por linhas e colunas. Por exemplo, o primeiro diagrama, na linha dois da Figura 6-1, foca as variáveis de distância (que aparecem na coluna um) e o hang time (que aparece na linha dois). Esse diagrama demonstra uma possível relação linear positiva (ascendente) entre a distância e o hang time.

Observe que a Figura 6-1 é uma matriz simétrica ao longo da linha diagonal. Ou seja, o diagrama para distância e hang time é igual ao diagrama para hang time e distância; os eixos  $x$  e  $y$  estão apenas em posições trocadas, mas a relação essencial aparece de qualquer forma. Sendo assim, basta analisar todos os diagramas abaixo da diagonal (onde aparecem os nomes das variáveis) ou acima da diagonal. Não é preciso examinar ambos.



**Figura 6-1:** Matriz para todos os diagramas de dispersão entre os pares de variáveis envolvidos no exemplo da distância atingida por um punt.

Para obter uma matriz com todos os diagramas de dispersão para um conjunto de variáveis no Minitab, clique em Graph>Matrix Plot e escolha Matrix of Plots>Simple. Selecione todas as variáveis no campo à esquerda, para as quais você deseja obter um diagrama; clique em Select e, depois, em OK. Você vai ver a matriz de diagramas de dispersão em um formato semelhante ao da Figura 6-1.

Analisando a linha um da Figura 6-1, você pode ver que todas as variáveis parecem ter uma relação linear positiva com a distância do punt, exceto a flexibilidade da perna esquerda. Talvez, a razão para que a flexibilidade da perna esquerda não esteja tão relacionada à distância do punt seja o fato de que a perna esquerda fique plantada no solo no momento do chute — para um jogador destro, a perna esquerda não precisa ter a mesma flexibilidade que a direita, a perna que realiza o chute. Assim, a flexibilidade da perna esquerda por si só não parece contribuir muito para a estimativa da distância de um punt.

Você também pode notar na Figura 6-1 que os diagramas que mostram as relações entre os pares das variáveis  $x$  estão à direita da coluna um e abaixo da linha um (lembre-se, você precisa olhar apenas a parte inferior ou a parte superior da matriz para visualizar os diagramas relevantes). Aparentemente, o hang time se relaciona um pouco com cada uma das outras variáveis (exceto com a flexibilidade da perna esquerda, que não contribui para a estimativa de  $y$ ). Portanto, possivelmente, o hang time pode ser a variável mais importante para a estimativa da distância de um chute de devolução.

### ***Procurando vínculos por meio das correlações***

Os diagramas de dispersão podem lhe dar ideias gerais, como por exemplo, se duas variáveis se relacionam de forma linear. Entretanto, a localização dessa relação requer um valor numérico que lhe diga a força com que essas variáveis se relacionam (de modo linear), bem como a direção dessa relação. Esse valor numérico é a *correlação* (também

conhecida como *correlação de Pearson*, veja o Capítulo 4). Assim, o próximo passo para continuar peneirando as possíveis candidatas às variáveis  $x$  é calcular a correlação entre cada variável  $x$  e  $y$ .

Para conseguir o conjunto de todas as correlações entre qualquer conjunto de variáveis em seu modelo, usando o Minitab, clique em Stat>Basic Statistics>Correlation. Em seguida, selecione todas as variáveis para as quais quer a correlação e clique em Select (para incluir os valores- $p$  para cada correlação, clique no campo Display  $p$ -values). E, depois, clique em OK. Você vai ver uma listagem com todos os nomes das variáveis desde a linha superior à parte inferior da primeira coluna. Faça a interseção entre a linha que descreve a primeira variável com a coluna que descreve a segunda, a fim de encontrar a correlação para esse par.

A Tabela 6-2 mostra as correlações que podem ser calculadas entre  $y$  = distância do punt e cada uma das variáveis  $x$ . Esses resultados confirmam o que os diagramas de dispersão estavam lhe mostrando. A distância parece estar relacionada a todas as variáveis, exceto à flexibilidade da perna esquerda, pois essa foi a única variável que não teve uma correlação estatisticamente significativa com a distância, usando um nível  $\alpha$  de 0,05. Para mais sobre testes de correlação, veja o Capítulo 5.

**Tabela 6-2** Correlações entre Distância de um Punt e Outras Variáveis

<i>Variável <math>x</math></i>	<i>Correlação com Distância do punt</i>	<i>valor-<math>p</math></i>
Hang time	0,819	0,001*
Força perna direita	0,791	0,001*
Força perna esquerda	0,744	0,004*
Flexibilidade perna direita	0,806	0,001*
Flexibilidade perna esquerda	0,408	0,167
Força total das pernas	0,796	0,001*

\* Estatisticamente significativo em nível  $\alpha = 0,05$

Se você olhar a Figura 6-1, vai ver que o hang time se relaciona com outras variáveis  $x$ , assim como com a força das pernas direita e esquerda, flexibilidade da perna direita etc. E é aqui que as coisas começam a ficar complicadas. Vimos que o hang time se relaciona à distância e que muitas outras variáveis se relacionam ao hang time. Embora seja claramente a variável mais relacionada à distância, talvez o modelo de regressão múltipla final não possa incluí-lo.

Aqui vai uma possível situação: você encontra uma combinação de outras variáveis  $x$  que, trabalhadas juntas, podem fazer uma boa estimativa para  $y$ . Mas todas essas outras variáveis possuem uma relação muito forte com o hang time. Esse resultado pode então

significar que no final você não vai precisar incluir o hang time no modelo. É, coisas estranhas acontecem quanto você pode optar por muitas variáveis  $x$ .

Depois de estreitar ainda mais o conjunto de possíveis variáveis  $x$  a serem incluídas no modelo para prever a distância de um punt, o próximo passo é colocá-las em um processo de seleção que irá resumir essa lista a um conjunto de variáveis essenciais à previsão de  $y$ .



# *Igual a Comprar Sapatos: O Modelo É Lindo, Mas Serve?*

Quando você entra no processo de seleção de um modelo, descobre que existem muitos métodos diferentes para escolher o melhor deles, segundo um grande número de critérios. Cada um pode resultar em modelos que se diferenciam uns dos outros, mas isso é uma das coisas que amo em estatística: às vezes não existe uma única melhor resposta.

Os três processos de seleção de modelos abrangidos nesta seção são:

- ✓ **Best subsets**
- ✓ **Forward**
- ✓ **Backward**

De todos os processos de seleção de modelos por aí, o mais votado pelos estatísticos é o *best subsets*, que examina cada modelo possível e determina o que melhor se ajusta, usando alguns critérios.

Nesta seção, você vai ver os diferentes métodos usados pelos estatísticos para avaliar e comparar o ajuste de diferentes modelos. Vamos ver, passo a passo, como o procedimento *best subsets* funciona. Depois eu mostro como pegar toda a informação e esmiuçá-la de forma a chegar à resposta — o modelo mais ajustado com base em um subconjunto de variáveis  $x$  disponíveis. E, por fim, você vai ver como esse procedimento é aplicado para encontrar um modelo que preveja a distância atingida por um punt.

## *Avaliando o ajuste do modelo de regressão múltipla*

Para qualquer procedimento de seleção de modelo, a avaliação do ajuste de cada um dos que forem considerados está embutida no processo. Ou seja, à medida que você examina todos os possíveis modelos, estará sempre de olho em como ele se ajusta. Então, antes de entrar em uma discussão sobre como realizar o procedimento *best subsets*, é preciso ter os critérios para avaliar como um determinado modelo se ajusta a um conjunto de dados.

Embora existam milhares de estatísticas diferentes para avaliar o ajuste dos modelos de regressão, discutirei as mais comuns:  $R^2$  (apenas para a regressão linear simples),  $R^2$  ajustado e C-p de Mallows. Essas três aparecem na linha inferior da saída do Minitab quando você realiza qualquer tipo de processo de seleção de modelo. Aqui vai uma explicação de cada uma dessas técnicas de avaliação:

- ✓  **$R^2$ :**  $R^2$  é a porcentagem de variabilidade nos valores de  $y$  explicada pelo modelo. Essa porcentagem vai de 0 a 100% (0 e 1,0). Na regressão linear simples (veja o Capítulo 4), um valor  $R^2$  alto significa que a reta se ajusta bem, enquanto um valor  $R^2$  baixo significa o contrário. Quando se trata de uma regressão múltipla, no entanto, temos um probleminha aqui.

Conforme você adiciona cada vez mais variáveis (independentemente do quão significativas são), o valor de  $R^2$  aumenta ou permanece o mesmo — mas nunca diminui, o que resulta em uma medida inflada de como o modelo se ajusta. Mas é claro que os estatísticos têm a solução para o problema, o que me leva ao próximo item da lista.

- ✓  **$R^2$  ajustado:** Ajusta o valor de  $R^2$  para baixo, segundo o número de variáveis no modelo. Quanto maior o número de variáveis no modelo, menor será o valor do  $R^2$  ajustado, comparado ao seu valor original do  $R^2$ . Um alto  $R^2$  ajustado significa que o modelo está se ajustando muito bem aos dados (quanto mais próximo de 1, melhor). Para mim, o valor de 0,70 já pode ser considerado bom para o  $R^2$  ajustado, e, quanto mais alto, melhor.

Sempre utilize o  $R^2$  ajustado no lugar do  $R^2$  para avaliar o ajuste de um modelo de regressão múltipla. A cada adição de uma nova variável ao modelo de regressão múltipla, o valor de  $R^2$  permanece igual ou aumenta, mas nunca diminui, já que uma nova variável ou ajudaria a explicar a variabilidade nos valores de  $y$  (aumentando, assim, o valor do  $R^2$ ), ou não faria diferença alguma (deixando o valor de  $R^2$  exatamente igual ao que estava antes). Portanto, teoricamente, seria possível adicionar cada vez mais variáveis ao modelo só para obter valores mais altos de  $R^2$ .

O  $R^2$  ajustado é importante, pois faz com que você evite a adição de mais e mais variáveis ao levar em conta a quantidade de variáveis já presentes no modelo. Na verdade, o valor do  $R^2$  ajustado pode até diminuir se o valor agregado das variáveis adicionadas for superado pelo número de variáveis no modelo. Isso lhe dá uma ideia do valor agregado obtido a partir de um modelo maior (nem sempre mais é melhor).

- ✓ **C-p de Mallow:** O C-p de Mallow pega o erro não explicado por um modelo de  $p$  com as variáveis  $x$ , divide-o pela quantidade média de erro deixada pelo modelo completo (com todas as variáveis  $x$ ) e ajusta o resultado para o número de observações ( $n$ ) e o número de variáveis  $x$  usadas ( $p$ ). De modo geral, quanto menor for o C-p de Mallow, melhor, pois quando se trata da quantidade de erro em seu modelo, menos é sempre mais. Um C-p com valor próximo a  $p$  (o número de variáveis  $x$  no modelo) reflete um modelo que se ajusta bem.

## *Processo de seleção de modelo*

O processo para encontrar o “melhor” modelo não é fixo (que coisa! Aqui, nem a definição de “melhor” é fixa). Existem muitos processos diferentes para se chegar a diferentes modelos de forma sistematizada, avaliando cada um até encontrar o modelo certo. Os três processos mais comuns são a forward selection, a backward selection e o best subsets model. Nesta seção, faço uma breve explicação a respeito dos procedimentos da forward e



backward selection e, então, entro nos detalhes sobre o best subsets model, o mais usado pelos estatísticos.

## ***O processo de seleção forward***

O *processo de seleção forward* começa com um modelo sem variáveis, ao qual elas vão sendo adicionadas, uma de cada vez, de acordo com o quanto podem contribuir com o modelo.

Comece escolhendo um nível de entrada com valor  $\alpha$ . Depois, faça testes de hipótese (veja o Capítulo 3 para mais instruções) para cada variável  $x$  a fim de verificar como elas se relacionam com  $y$ . A variável  $x$  com o menor valor- $p$  vence e é adicionada ao modelo, desde que este valor seja menor do que o nível  $\alpha$ . Continue fazendo isso com as variáveis restantes até chegar à que tiver o menor valor- $p$  que não satisfaça o nível de entrada. Então pare.

A desvantagem do processo de seleção forward é que ele começa sem nenhuma variável e vai adicionando-as uma de cada vez e, depois que uma variável é adicionada, nunca mais é retirada. O melhor modelo pode até nem ser testado.

## ***Optando pelo processo de seleção “backward”***

O *processo de seleção backward* é o contrário do método forward. Ele começa com um modelo com todas as variáveis  $x$  e vai retirando uma de cada vez. As que menos contribuem com o modelo são as primeiras a serem retiradas. Primeiro, escolha um nível de retirada; depois, teste todas as variáveis  $x$  e encontre a com o maior valor- $p$ . Se o valor- $p$  dessa variável  $x$  for maior do que o nível de retirada, ela é retirada do modelo.

Continue retirando as variáveis do modelo até chegar à que tiver o maior valor- $p$  que não exceda o nível de retirada. Então pare.

A desvantagem do processo de seleção backward é que ele começa com todas as variáveis e retira-as uma de cada vez e, depois que uma variável é retirada, nunca mais ela volta. E, novamente, o melhor modelo pode até nem ser testado.

## ***Usando o procedimento best subsets***

O procedimento best subsets possui menos etapas do que os modelos de seleção forward e backward, pois o computador formula e analisa todos os possíveis modelos de uma só vez. Nesta seção, você vai aprender a obter os resultados e usá-los para chegar ao melhor modelo de regressão múltipla para a previsão de  $y$ .

Veja aqui os passos para a realização do processo de seleção *best subsets* para escolher o modelo de regressão múltipla; observe que o Minitab faz todo o trabalho de cálculo por você.

### **1. Realize o processo best subsets no Minitab, utilizando todos os possíveis**



## subconjuntos das variáveis $x$ a serem incluídas no modelo final.

Para realizar o processo de seleção best subsets no Minitab, clique em Stat>Regression>Best Subsets. Selecione a variável resposta ( $y$ ) e clique em Select. Selecione a variável preditora ( $x$ ), clique em Select e, depois, em OK.

A saída contém uma lista com todos os modelos que têm uma, duas, três variáveis  $x$ , e assim por diante, até chegar ao modelo completo (que contém todas as variáveis  $x$ ). Cada modelo é apresentado em uma linha da saída.

- 2. Para escolher o melhor de todos os modelos mostrados na saída do Minitab, encontre o modelo com o maior valor para  $R^2$  ajustado e o menor valor C-p de Mallow; se tiver que desempatar entre dois modelos, escolha o que tiver o menor número de variáveis.**

Se o modelo se ajusta bem, o valor para  $R^2$  ajustado é alto. Sendo assim, procure o menor modelo possível que tenha um  $R^2$  ajustado alto e um C-p de Mallow pequeno, comparado a seus concorrentes. Caso encontre dois modelos semelhantes, sempre escolha o modelo com menos variáveis, pois assim será mais fácil interpretar o modelo final.

### *O segredo para ser um punter de sucesso: um exemplo*

Voltando ao exemplo da distância atingida por um punt, suponha que você tenha analisado os dados sobre a distância alcançada por um punt usando o processo de seleção best subsets. Os resultados estão na Figura 6-2. Esta seção segue os passos do Minitab na obtenção desses resultados e serve como um guia para sua interpretação.

Assumindo que você já tenha usado o Minitab para realizar o processo de seleção best subsets, você pode agora analisar a saída ilustrada na Figura 6-2. Cada variável aparece como uma coluna no lado direito da saída. Cada linha representa os resultados de um modelo contendo o número de variáveis mostradas na coluna um. Os X no final de cada linha mostram quais variáveis foram incluídas naquele modelo. O número de variáveis no modelo começa em 1 e vai até 6, pois seis eram as variáveis  $x$  disponíveis no conjunto de dados.

Os modelos com o mesmo número de variáveis estão ordenados segundo seu valor para  $R^2$  ajustado e para o C-p de Mallow, do melhor para o pior. Os dois melhores modelos (para cada número de variáveis) estão incluídos na saída do programa.

Por exemplo, as linhas um e dois da Figura 6-2 (ambas marcadas como 1 na coluna Vars) mostram os dois melhores modelos que contêm uma variável  $x$ ; as linhas três e quatro mostram os dois melhores modelos que contêm duas variáveis, e assim por diante. Por fim, a última linha mostra os resultados para o modelo completo, o que contém as seis variáveis (apenas um modelo contém todas as seis variáveis, portanto, não existe um segundo melhor modelo nesse caso).

Analisando as duas primeiras linhas da Figura 6-2, o primeiro melhor modelo é o que inclui apenas a variável hang time. O segundo, é o que inclui apenas a flexibilidade da perna direita, que tem um  $R^2$  menor e um C-p Mallow mais alto do que o modelo com o hang time e, por isso, é melhor.

A linha três mostra que o melhor modelo com duas variáveis para estimar a distância do punt é o modelo que contém a força da perna direita e a força total das pernas. O melhor modelo com três variáveis está na linha cinco, que inclui a força da perna direita, a flexibilidade da perna direita e a força total das pernas. O melhor modelo com quatro variáveis encontra-se na linha sete e inclui a força da perna direita, a flexibilidade das pernas direita e esquerda, além da força total das pernas. O melhor modelo com cinco variáveis encontra-se na linha nove e inclui quase todas as variáveis, com exceção da força da perna esquerda. O único modelo que inclui todas variáveis é mostrado na última linha.

Entre os melhores modelos com uma, duas, três, quatro e cinco variáveis, qual você deve escolher como modelo final para a regressão múltipla? Qual é o melhor entre os melhores? Com todos esses resultados, é fácil pirar na hora de escolher um, mas não tenha medo — Mallow está aqui para ajudá-lo (com seu leal escudeiro, o  $R^2$  ajustado).

Analisando a coluna três da Figura 6-2, é possível notar que, à medida que o número de variáveis aumenta, o  $R^2$  ajustado chega ao máximo e, depois, começa a diminuir. Isso acontece porque o  $R^2$  ajustado considera o número de variáveis no modelo e corrige o  $R^2$ . Você pode ver que o  $R^2$  ajustado atinge seu ponto máximo em 74,1% para dois modelos. Os modelos correspondentes são o modelo com duas variáveis (força da perna direita e a força total das pernas) e o modelo com três variáveis (força da perna direita, flexibilidade da perna direita e a força total das pernas).

Agora, veja o C-p de Mallow para esses dois modelos. Observe que ele é zero para o modelo com duas variáveis e 1,3 para o modelo com três variáveis. Os dois valores são pequenos se comparados aos demais, mas, uma vez que o C-p de Mallow é menor para o modelo com duas variáveis e também porque ele tem uma variável a menos, você deve escolher o modelo com duas variáveis (força da perna direita e a força total das pernas) como seu modelo final, usando o procedimento best subsets.

Regressão Best Subsets: Distância versus Hang, Força\_D. . .

Resposta é a Distância

					R L									
					$\overline{F} \overline{F}$									
					R L 1 1 O									
					$\overline{S} \overline{S} x x \overline{S}$									
					t t i i t									
					r r b b r									
					e e i i e									
					H n n 1 1 n									
					a g g i i g									
					n t t t t t									
					g h h y y h									
Vars	R-Sq	R-Sq(adj)	Mallows		S									
			C-p											
1	67.1	64.1	1.7	15.570	x									
1	65.0	61.8	2.3	16.043					x					
2	78.5	74.1	-0.0	13.206		x							x	
2	78.2	73.8	0.1	13.294			x			x				x
3	80.6	74.1	1.3	13.214			x		x		x			x
3	79.5	72.7	1.6	13.581		x	x						x	
4	81.4	72.1	3.0	13.724			x		x	x	x			x
4	80.7	72.0	3.3	13.977			x	x	x			x		x
5	81.5	68.2	5.0	14.643		x	x		x	x	x	x		x
5	81.4	68.2	5.0	14.650		x	x	x	x	x			x	x
6	81.5	62.9	7.0	15.812		x	x	x	x	x	x			x

**Figura 6-2:** Resultados do procedimento best subsets para o exemplo da distância de um punt.

<sup>1</sup> N.E.: Punt é uma jogada de futebol americano onde um jogador (punter) recebe a bola lançada por um companheiro de equipe (long snapper) chutando-a em longa distância (metros) em direção ao adversário, objetivando avançar o máximo de posições de campo possível.

# Capítulo 7

## Subindo na Curva de Aprendizagem com a Regressão Não Linear

### *Neste Capítulo*

- ▶ Conhecendo a regressão não linear
- ▶ Usando os diagramas de dispersão
- ▶ Ajustando um polinômio a seu conjunto de dados
- ▶ Explorando modelos exponenciais para ajustar seus dados

**E**m Estatística I, você se concentra no *modelo de regressão linear simples*, onde procura uma variável quantitativa,  $x$ , que pode ser usada para estimar outra variável quantitativa,  $y$ , usando uma reta. Os exemplos estudados em Estatística I caem como uma luva para este tipo de modelo, tal como usar a altura para estimar o peso ou usar o tempo de estudo para estimar a nota na prova. (Para mais informações e exemplos de como usar os modelos de regressão linear simples, veja o Capítulo 4.)

Mas nem todas as situações caem nessa categoria. Por exemplo, pegue a relação entre o rendimento do combustível e a velocidade. A baixas velocidades, o rendimento da gasolina é mais baixo; a velocidades altas, também é mais baixo; porém, a velocidades médias, o rendimento da gasolina é mais alto. Essa relação baixa-alta-baixa entre velocidade e rendimento do combustível representa uma relação em curva. As relações que não são representadas por retas são chamadas de *relações não lineares* (muito esperto, não?). Simplificando, a regressão não linear entra em cena quando você quer prever uma variável quantitativa  $y$  usando outra variável quantitativa  $x$ , mas o padrão observado nos dados coletados lembra uma curva, e não uma reta.

Neste capítulo, você vê como percorrer a estrada cheia de curvas que nos leva aos modelos de regressão não linear. Mas a boa notícia vem em dobro: você pode usar muitas das técnicas vistas para a regressão linear e, no final, o Minitab faz toda a análise por você.

# *Antecipando a Regressão Não Linear*

A regressão não linear entra em cena nas situações em que você construiu para seus dados o *diagrama de dispersão* (gráfico bidimensional que mostra a variável  $x$  no eixo  $x$  e a variável  $y$  no eixo  $y$ ; veja a próxima seção "Começando pelos Diagramas de Dispersão") que mostra um padrão que se parece mais com um tipo de curva. Exemplos de dados que seguem uma curva incluem as mudanças no tamanho da população ao longo do tempo, a demanda por um produto em função de sua oferta ou o tempo de duração de uma bateria. Quando um conjunto de dados segue um padrão em curva, é hora de deixar os modelos de regressão linear de lado (abordados nos Capítulos 4 e 5) e partir em busca de um modelo de regressão não linear.

Suponha que a gerente de uma empresa esteja considerando a compra de um novo software de gestão, mas está hesitante. Ela quer saber quanto tempo uma pessoa normalmente leva para aprender a usar o software com agilidade.

Onde a estatística entra aqui? Ela quer um modelo que mostre como é a curva de aprendizagem (em média), que mostra a redução do tempo de realização de uma tarefa com o aumento da prática adquirida. Nesse contexto, você tem duas variáveis: o tempo para completar a tarefa e o número de tentativas (por exemplo, a primeira tentativa é designada pelo número 1, a segunda por 2, e assim por diante). As duas variáveis são *quantitativas* (numéricas), e o que você quer é encontrar um vínculo entre elas. Nesse ponto, você pode começar a pensar em uma regressão.

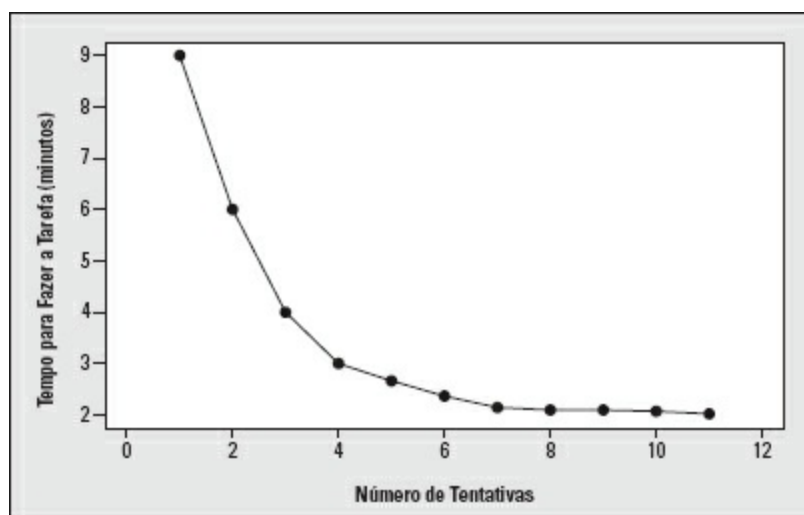
O *modelo de regressão* produz uma função (seja uma reta ou não) que descreve um padrão ou uma relação. A relação aqui é o tempo de realização de uma tarefa versus o número de vezes que a tarefa é praticada. Mas que tipo de modelo de regressão usar? Afinal de contas, neste livro, vemos quatro tipos: regressão linear simples, regressão linear múltipla, regressão não linear e regressão logística. Precisa de mais pistas?

A palavra "curva" no termo curva de aprendizagem é uma pista de que a relação a ser modelada aqui pode não ser linear. Essa palavra indica que se trata de um modelo de regressão não linear. Se pensarmos em como pode ser uma curva de aprendizagem, podemos imaginar que o tempo de realização da tarefa fica no eixo  $y$  e que o número de experimentos fica no eixo  $x$ .

Você pode imaginar que os valores de  $y$  serão altos de primeira, pois nas primeiras vezes que tentamos fazer uma tarefa nova, levamos mais tempo para realizá-la. Depois, à medida que a tarefa vai sendo repetida, o tempo para realizá-la diminui, mas em algum ponto o aumento da prática não vai reduzir muito mais o tempo de realização. Assim, a relação pode ser representada por algum tipo de curva, como a que simulo na Figura 7-1 (que pode ser ajustada por uma função exponencial).

Este exemplo ilustra o básico da regressão não linear; o restante do capítulo mostra como o modelo se divide.





---

**Figura 7-1:** Curva de aprendizagem para tempo de realização de uma nova tarefa.

---

# Começando com Diagramas de Dispersão

Como em qualquer análise de dados, antes de mergulhar de cabeça e escolher um modelo que você ache que se ajusta aos dados (ou que deveria se ajustar), dê um passo para trás e observe-os novamente para ver se existe algum padrão. Para isso, analise o diagrama de dispersão dos dados e veja se você pode ou não traçar uma curva ao longo deles e descubra se a maioria dos dados acompanha essa curva.

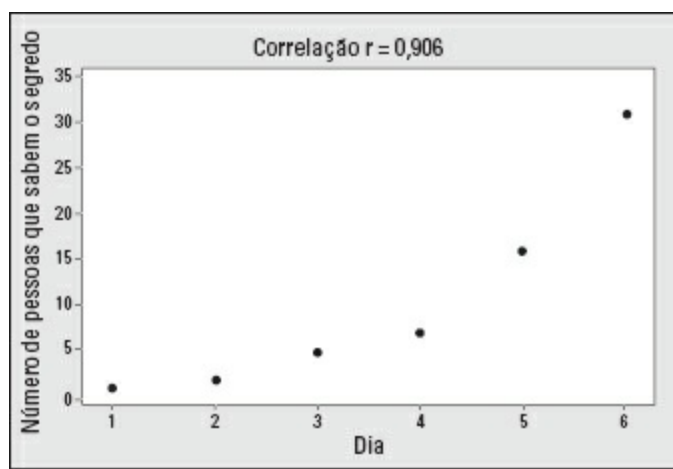
Suponha que você esteja interessado em modelar a rapidez com que um boato se espalha. Uma pessoa fica sabendo de um segredo e o conta para outra, e, agora, duas pessoas sabem sobre ele; cada uma delas conta a outra pessoa, e, agora, quatro já sabem o segredo. Algumas dessas pessoas podem passá-lo adiante e, assim, ele vai sendo contado de pessoa para pessoa. Logo, um grande número de pessoas saberá o segredo, que passa a não ser mais um segredo.

Para coletar seus dados, você pode contar o número de pessoas que saibam um segredo, rastreando, ao longo de um período de seis dias, quem contou para quem. Os dados estão na Tabela 7-1. Observe que a difusão do segredo pega fogo no quinto dia — é assim que um modelo exponencial funciona. Veja o diagrama de dispersão dos dados ilustrado na Figura 7-2.

**Tabela 7-1**                      **Número de Pessoas que Souberam de um Segredo em 6 Dias**

<i>x (Dia)</i>	<i>y (Número de pessoas)</i>
1	1
2	2
3	5
4	7
5	17
6	30

Nesta situação, a variável explicativa,  $x$ , é o dia, e a variável de resposta,  $y$ , é o número de pessoas que sabem do segredo. Observando a Figura 7-2, podemos visualizar um padrão entre os valores de  $x$  e  $y$ . Porém, esse padrão não é linear. Ele é uma curva ascendente. Se você tentasse ajustar uma reta a esse conjunto de dados, como ficaria esse ajuste?



**Figura 7-2:** Diagrama de dispersão mostrando a difusão de um segredo ao longo de seis dias.

Para solucionar essa questão, veja o coeficiente de correlação entre  $x$  e  $y$ , encontrado na Figura 7-2, como sendo 0,906 (veja o Capítulo 4 para mais informação sobre correlação). Você pode interpretar essa correlação como sendo uma relação linear forte e positiva (ascendente) entre  $x$  e  $y$ . No entanto, nesse caso, a correlação é enganosa, pois o diagrama de dispersão parece estar curvado.

Se a correlação parecer boa (próxima de  $+1$  ou  $-1$ ), não pare por aqui. Como em qualquer análise de regressão, é muito importante levar em consideração tanto o diagrama de dispersão quanto a correlação antes de se decidir pelo modelo que será ajustado aos dados. A contradição entre o diagrama e a correlação, neste exemplo, é um sinal de que o modelo linear não é uma boa opção para este caso.

O coeficiente de correlação é o número que mede a força e a direção da relação linear entre duas variáveis quantitativas,  $x$  e  $y$  (veja o Capítulo 4). Entretanto, você pode se deparar com situações (como a mostrada na Figura 7-2) em que a correlação é forte, embora o diagrama de dispersão nos diga que uma curva se ajustaria melhor. Não confie apenas no diagrama ou no coeficiente de correlação para decidir se deve continuar e ajustar uma reta a seus dados.

A lição aqui é que ajustar uma reta a dados que parecem ter um padrão de curva não é o caminho a ser seguido. Em vez disso, explore os modelos que possuem padrões em curva.

As seções a seguir abordam os dois principais modelos não lineares usados para delinear dados em curvas: polinômios (que não são retas — ou seja, são curvas como as funções quadráticas ou cúbicas) e os modelos exponenciais (que começam pequenos, mas aumentam rapidamente, ou vice-versa). Uma vez que o padrão dos dados na Figura 7-2 começa baixo e faz uma curva ascendente, o modelo correto para ajustá-lo é o modelo de regressão exponencial (esse modelo também é adequado a dados que começam altos e fazem uma curva descendente).

# Nas Curvas da Estrada com os Polinômios

A principal família de modelos não lineares é a dos *polinômios*. Esses modelos são usados quando uma função polinomial (que está além de uma reta) é a que melhor descreve a curva nos dados (por exemplo, os dados podem seguir o formato de uma parábola, que é uma função polinomial de segundo grau). Normalmente, os modelos polinomiais são usados quando os dados seguem um padrão de curvas que sobem e descem um certo número de vezes.

Por exemplo, suponha que uma médica examine a ocorrência de problemas cardíacos em pacientes em relação à sua pressão sanguínea. Ela, então, descobre que os pacientes com pressão sanguínea muito baixa ou muito alta tiveram maior ocorrência de problemas, enquanto os pacientes cujas pressões foram médias, ficando dentro do considerado normal, apresentaram poucas complicações. O padrão para esses dados possui o formato de um U, e a parábola se ajustaria muito bem aqui.

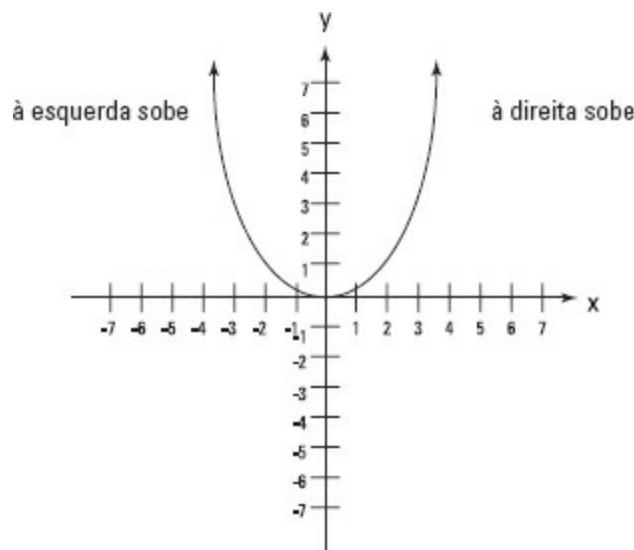
Nesta seção você vai ver o que é um modelo de regressão polinomial, como procurar um polinômio que se ajuste bem a seus dados e como avaliá-lo.

## Relembrando o que é um polinômio

Talvez você se lembre de que, em álgebra, um *polinômio* é a soma de termos  $x$  elevados a uma variedade de potências e que cada  $x$  é precedido por uma constante chamada *coeficiente*. Por exemplo, o modelo  $y = 2x + 3x^2 + 6x^3$  é um polinômio. A fórmula geral de um modelo de regressão polinomial é  $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_k x^k + \varepsilon$ . Aqui,  $k$  representa o número total de termos no modelo. O  $\varepsilon$  representa o erro que ocorre simplesmente em virtude da casualidade (não é um erro ruim, apenas uma flutuação aleatória de um modelo perfeito).

Veja aqui alguns dos polinômios mais comuns com os quais você pode se deparar enquanto estiver analisando dados e ajustando modelos. Lembrese, o modelo a ser usado é sempre o mais simples que se ajustar (não tente bancar o herói em estatística — deixe isso para Batman e Robin). Os modelos que discuto neste livro são seus velhos amigos da álgebra: polinômios de segundo, terceiro e quarto graus.

- ✓ **Polinômio de segundo grau (ou quadráticos):** Esse modelo chama-se *polinômio de segundo grau (ou quadrático)*, pois o expoente mais alto é 2. Um exemplo é o modelo  $y = 2x + 3x^2$ . Um polinômio de segundo grau forma uma parábola — de concavidade para cima ou para baixo; ela muda a direção uma vez (veja a Figura 7-3).



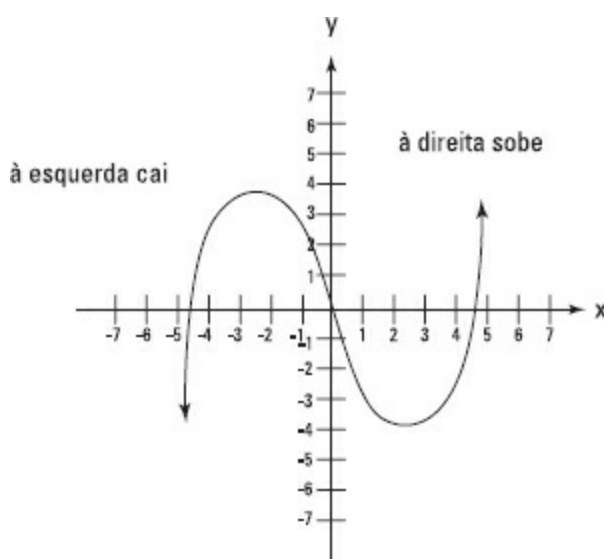

---

**Figura 7-3:** Exemplo de um polinômio de segundo grau.

---

- ✓ **Polinômio de terceiro grau:** Esse modelo tem o 3 como expoente mais alto de  $x$ . Geralmente, tem a forma de um S deitado e muda de direção duas vezes (veja a Figura 7-4).
- ✓ **Polinômio de quarto grau:** Os polinômios de quarto grau possuem  $x^4$ . Normalmente mudam de direção três vezes e se parecem com um W ou um M, dependendo do como estiverem (veja a Figura 7-5).

De forma geral, se o maior expoente no polinômio for  $n$ , o número de vezes que a curva muda de direção no gráfico é igual a  $n - 1$ . Para mais informações sobre os gráficos dos polinômios, consulte um livro de álgebra ou *Álgebra I Para Leigos*, de Mary Jane Sterling (Alta Books).




---

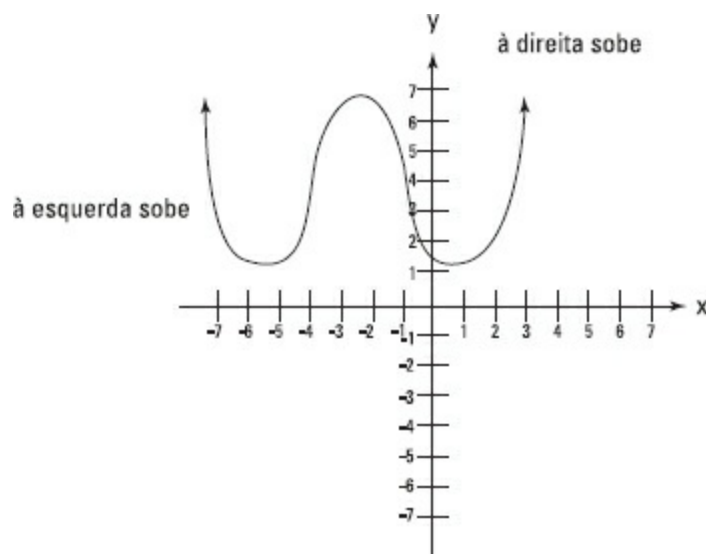
**Figura 7-4:** Exemplo de um polinômio de terceiro grau.

---

Os modelos não lineares neste capítulo envolvem apenas uma variável explicativa,  $x$ . Você



pode adicionar mais variáveis explicativas a um modelo de regressão não linear elevando cada variável a uma potência, mas esses modelos estão além do escopo deste livro. Veja algumas informações básicas sobre os modelos de regressão múltipla no Capítulo 5.



**Figura 7-5:** Exemplo de um polinômio de quarto grau.

## *Em busca do melhor modelo polinomial*



Ao ajustar um modelo de regressão polinomial aos seus dados, sempre comece por um diagrama de dispersão, pois, assim, poderá procurar padrões; o diagrama de dispersão lhe dá uma ideia do tipo de modelo que pode dar certo. Sempre comece com o modelo mais simples possível e vá incrementando-o, se necessário. Não comece logo de cara substituindo os valores em um modelo polinomial de grau alto. Aqui vão algumas das razões do porquê:

- ✓ **Os polinômios de alto grau são difíceis de serem interpretados e seus modelos são complexos.** Por exemplo, com uma reta você pode facilmente interpretar os valores do intercepto  $y$  e do coeficiente angular (a inclinação), mas a interpretação de um polinômio de décimo grau é complicada (e isso porque estou sendo boazinha).
- ✓ **Os polinômios de alto grau tendem a causar um sobreajustamento.** Se você ajustar seu modelo o mais próximo possível de cada ponto de dados, ele poderá não se ajustar a um novo conjunto de dados, o que quer dizer que suas estimativas para  $y$  podem estar totalmente erradas.



Para ajustar um polinômio a um conjunto de dados usando o Minitab, clique em Stat>Regression>Fitted Line Plot> e no tipo de modelo de regressão desejado: linear, quadrático ou cúbico (as opções no Minitab não vão além do polinômio de terceiro grau, mas essas opções são o suficiente para cobrir 90% dos casos em que o uso de um polinômio é adequado). Clique na variável  $y$  no campo à esquerda e depois em Select; essa

variável vai, então, aparecer no campo Response ( $y$ ). Clique na variável  $x$  no campo à esquerda e, depois, em Select. Essa variável vai, então, aparecer no campo Predictor ( $x$ ). Por fim, clique em OK.

A seguir, veja os passos a serem seguidos para verificar se um polinômio se ajusta ou não a seus dados (os softwares estatísticos podem assumir essa tarefa e ajustar os modelos depois que você lhes disser quais ajustar).

1. **Faça um diagrama de dispersão para seus dados e procure padrões, tais como retas ou curvas.**
2. **Se os dados formarem algo parecido com uma reta, primeiro tente ajustar um polinômio de primeiro grau (uma reta) aos dados:  $y = b_0 + b_1x$ .**

Mas se o diagrama não demonstrar um padrão linear, ou se a correlação não estiver próxima de  $+1$  ou  $-1$ , dê o terceiro passo.


3. **Se os dados formarem algo parecido com uma parábola, tente ajustar um polinômio de segundo grau:  $y = b_0 + b_1x + b_2x^2$ .**

Se os dados se ajustarem bem ao modelo, pare aqui e consulte a seção “Avaliando o ajuste de um modelo polinomial”. Se o modelo ainda não se ajustou bem, dê o quarto passo.

4. **Se você visualizar uma curva mais complexa do que uma parábola, tente o polinômio de terceiro grau:  $y = b_0 + b_1x + b_2x^2 + b_3x^3$ .**

Se os dados se ajustarem bem ao modelo, pare aqui e consulte a seção "Avaliando o ajuste de um modelo polinomial". Se o modelo ainda não se ajustou bem, dê o quinto passo.

5. **Continue tentando ajustar polinômios de graus mais altos até encontrar um que se ajuste, ou até que o grau do polinômio (o expoente mais alto) fique alto demais para encontrar um padrão confiável.**

 Mas quanto é alto demais? Normalmente se você não conseguir ajustar os dados até um polinômio de terceiro grau, então, talvez seja melhor usar outro tipo de modelo. Ou, ainda, você pode determinar que os dados apresentaram um comportamento dispersivo e desordenado demais para tentar ajustá-los a um modelo.

O Minitab pode realizar todos esses passos por você até o segundo grau (que é o terceiro passo); a partir daí, você vai precisar de um programa mais sofisticado, como o SAS ou o SPSS. No entanto, a maioria dos modelos que você vai precisar ajustar vão até os polinômios de segundo grau.

## ***Usando um polinômio de segundo grau para passar na prova***

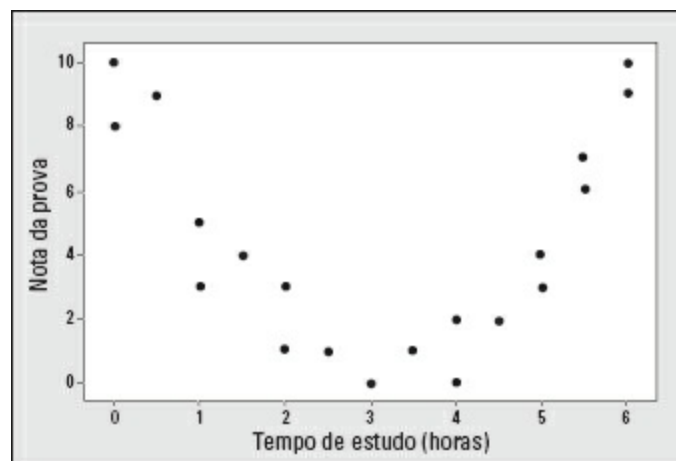
O primeiro passo para o ajuste de um modelo polinomial é colocar os dados em um

diagrama de dispersão e ver se eles se encontram em um padrão específico. Existem muitos tipos diferentes de polinômios para ajustar os dados que se encontram em um padrão do tipo curva. Um dos padrões mais comumente encontrados em dados em curva é o padrão quadrático, ou polinômio de segundo grau, que sobe e desce, ou desce e sobe, à medida que os valores de  $x$  se movimentam da esquerda para a direita (veja a Figura 7-3). O polinômio de segundo grau é o mais simples e o mais utilizado depois da reta, portanto, ele merece uma atenção especial (depois de ter aprendido os conceitos básicos de um polinômio de segundo grau, você pode aplicá-los aos polinômios de graus mais altos).

Suponha que 20 alunos façam uma prova de Estatística. Você, então, registra as notas das provas (cujo valor máximo é dez) e o número de horas que os alunos disseram ter estudado para a prova. Os resultados podem ser vistos na Figura 7-6.

Analisando a Figura 7-6, aparentemente há três tipos de alunos nessa classe: o Tipo 1, na extremidade esquerda do eixo  $x$ , entende do assunto (conforme refletido em suas notas altas), mas não precisou estudar muito (veja que o tempo de estudo no eixo  $x$  é baixo); o Tipo 3 também foi bem na prova (como indicado por suas notas altas), mas teve que estudar muito para consegui-las (como podemos ver na extremidade direita do eixo  $x$ ); os alunos no meio, o Tipo 2, não parecem ter ido bem.

De forma geral, de acordo com o diagrama de dispersão, o tempo de estudo parece explicar as notas obtidas na prova, de modo a indicar um polinômio de segundo grau. Portanto, os dados podem ser ajustados por um modelo de regressão quadrática.

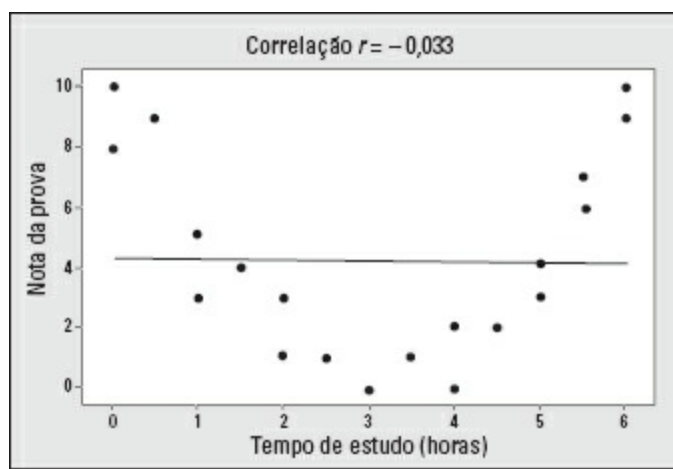


**Figura 7-6:** Diagrama de dispersão mostrando tempo de estudo e nota da prova.

Suponha que um analista de dados (não você!) não conheça a regressão polinomial e simplesmente tente ajustar uma reta aos dados do exemplo em questão. Na Figura 7-7, vemos os dados e a reta que ele tentou ajustar. A correlação mostrada na figura é  $-0,033$ , que é basicamente zero. Essa correlação indica que não existe uma relação linear entre  $x$  e  $y$ . Porém, isso não significa que não exista algum tipo de relação, só indica que essa não é uma relação linear. (Veja o Capítulo 4 para mais informações sobre relações lineares.) Portanto, tentar ajustar uma reta aqui foi, de fato, uma péssima ideia.





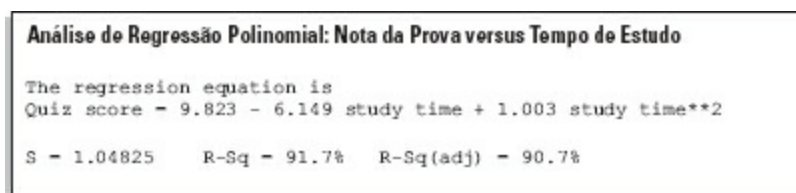


**Figura 7-7:** Tentando ajustar uma reta a dados quadráticos

Depois de saber que um polinômio quadrático parece se ajustar bem aos dados, o próximo desafio é, entre todas as possíveis parábolas, encontrar a equação para a parábola que se ajusta aos dados.

Lembre que em álgebra a equação geral da parábola é  $y = ax^2 + bx + c$ . Agora você deve encontrar os valores de  $a$ ,  $b$  e  $c$  para criar a parábola que melhor se ajusta aos dados (da mesma forma que você encontra  $a$  e  $b$  para criar a reta que melhor se ajusta em um modelo de regressão linear). Esse é o objeto de qualquer análise de regressão.

Suponha que você ajuste um modelo de regressão quadrática aos dados sobre as notas da prova usando o Minitab (veja a saída do Minitab na Figura 7-8 e as instruções de como usar o Minitab para ajustar esse modelo na seção anterior). Na linha superior da saída, você pode ver que a equação da parábola que melhor se ajusta é  $\text{quiz score} = 9,82 - 6,15 * (\text{study time}) + 1,00 * (\text{study time})^2$ . Observe que nesse exemplo  $y$  é a nota da prova e  $x$  é o tempo de estudo, pois estamos utilizando o tempo de estudo para prever a nota da prova.

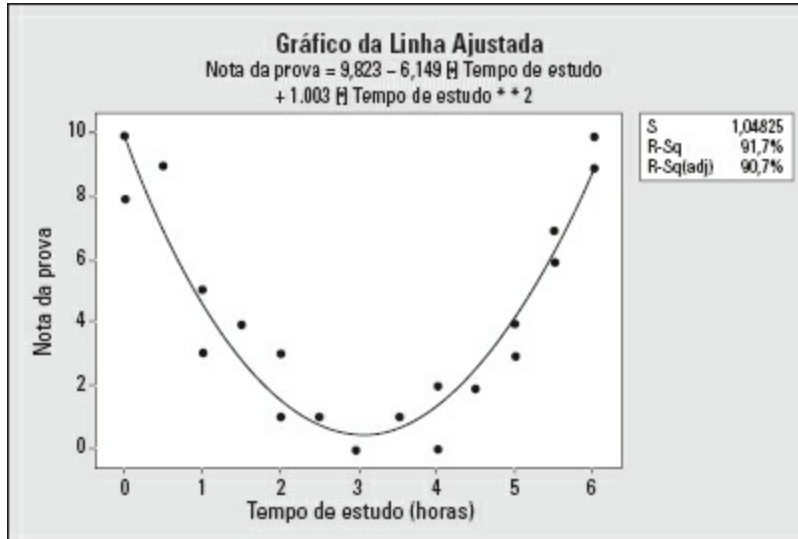


**Figura 7-8:** Saída do Minitab para o ajuste de uma parábola aos dados sobre a nota da prova.

O diagrama de dispersão dos dados e a parábola ajustada por meio do modelo de regressão estão ilustrados na Figura 7-9. Em álgebra você aprendeu que quando o termo quadrático é acompanhado por um coeficiente positivo (neste caso  $a = 1,00$ ), a concavidade da parábola está voltada para cima, o que é o caso aqui.

Analizando a Figura 7-9, realmente parece que o modelo quadrático se ajusta muito bem aos dados, pois estes se encontram bem próximos da curva encontrada pelo Minitab.

Entretanto, nem só de diagramas de dispersão vivem os estatísticos, por isso, a próxima seção vai ajudá-lo a descobrir como avaliar o ajuste de um modelo polinomial de forma mais detalhada.



**Figura 7-9:** A parábola parece se ajustar muito bem aos dados.

## *Avaliando o ajuste de um modelo polinomial*

Você faz um diagrama de dispersão para seus dados e observa um padrão em curva. Portanto, utiliza a regressão polinomial para ajustar um modelo aos dados; o modelo parece se ajustar bem, pois os pontos acompanham de perto a curva encontrada pelo Minitab, mas não pare aqui. Para ter certeza de que seus resultados podem ser generalizados à população de onde os dados foram retirados é preciso fazer mais algumas pequenas verificações, além do diagrama.

Para que o ajuste de um modelo fique acima de qualquer suspeita, além do diagrama de dispersão, observe mais dois itens, normalmente nesta ordem: o valor do  $R^2$  ajustado e os gráficos de resíduos.

Todas essas avaliações devem estar de acordo antes de você concluir que o modelo se ajusta. Se as três avaliações não estiverem de acordo, possivelmente você terá que usar um modelo diferente do polinomial para ajustar os dados, ou terá que mudar as unidades dos dados para fazer com que o modelo polinomial se ajuste melhor. Entretanto, a última opção está fora do escopo da Estatística II e, provavelmente, você não vai se deparar com essa situação.

Nas seções a seguir, vamos nos aprofundar no valor do  $R^2$  ajustado e nos gráficos de resíduos, a fim de descobrir como podemos utilizá-los para avaliar o ajuste de um modelo. (Encontre mais informações sobre diagramas de dispersão na seção "Começando pelos diagramas de dispersão" mais no início deste capítulo.)



## ***Examinando $R^2$ e $R^2$ ajustado***

Encontrar o  $R^2$ , o *coeficiente de determinação* (consulte o Capítulo 5 para mais detalhes), é como chegar ao dia do julgamento de qualquer modelo. Você encontra o  $R^2$  no resultado de sua regressão, listado como "R-Sq" logo abaixo da parte da saída do Minitab onde aparecem os coeficientes e as variáveis. A Figura 7-8 mostra a saída do Minitab para o exemplo da nota da prova; o valor  $R^2$  nesse caso é 91,7%.

O valor de  $R^2$  lhe diz a porcentagem da variação nos valores de  $y$  que o modelo pode explicar. Para interpretar essa porcentagem, observe que  $R^2$  é o quadrado de  $r$ , o coeficiente de correlação (veja o Capítulo 5). Uma vez que os valores de  $r$ , além de  $\pm 0,80$ , são considerados bons, os valores de  $R^2$  acima de 0,64 também são considerados da mesma forma, especialmente para modelos com apenas uma variável  $x$ .

Você pode considerar bons os valores de  $R^2$  acima de 80%, mas os valores abaixo de 60% são ruins. Os valores intermediários são medíocres, pois poderiam ser melhores. (Essa avaliação é uma regra minha; as opiniões podem variar um pouco de um estatístico para outro).

No entanto, há uma coisa que você deve saber: em Estatística, muitas variáveis estragam a sopa. Todas as vezes que você adicionar uma variável  $x$  a um modelo de regressão, o valor de  $R^2$  aumenta automaticamente, quer a variável ajude ou não (isso é apenas um fato matemático). Bem à direita do  $R^2$ , na saída do programa, para qualquer análise de regressão, está o valor de  $R^2$  ajustado, que corrige o valor de  $R^2$  para baixo a cada variável (e a cada potência de cada variável) adicionada ao modelo. Você não pode, simplesmente, sair acrescentando um monte de variáveis no modelo esperando que pequenos incrementos venham, ao serem somados, a chegar a um valor aceitável para  $R^2$ , sem ter que pagar um preço por isso.

Para maior segurança, você deveria sempre usar  $R^2$  ajustado, ao invés de  $R^2$  para avaliar o ajuste de seu modelo, especialmente se você tiver mais de uma variável  $x$  (ou mais de uma potência de uma variável  $x$ ). Os valores  $R^2$  e  $R^2$  ajustado são próximos quando você tem apenas algumas variáveis (ou potências) diferentes no modelo; porém, conforme aumenta o número de variáveis (ou potências) diferentes, a distância entre  $R^2$  e  $R^2$  ajustado também aumenta. Nesse caso, o  $R^2$  ajustado é o coeficiente mais justo e consistente para avaliar o ajuste do modelo.

No exemplo das notas da prova (análise mostrada na Figura 7-8), o valor de  $R^2$  ajustado é 90,7%, ainda assim um valor alto, indicando que o modelo quadrático se ajusta muito bem a esses dados. (Veja no Capítulo 6 mais informações sobre  $R^2$  e  $R^2$  ajustado.)

## ***Verificando os resíduos***



Você já verificou o diagrama de dispersão dos seus dados e viu que o valor de  $R^2$  é alto. O que fazer agora? Agora, você deve examinar quão bem o modelo se ajusta a cada ponto individual no dado para garantir que encontrará quaisquer locais onde o modelo esteja muito discrepante ou onde possa ter deixado passar outro padrão subjacente.

O *resíduo* é uma quantidade de erro, ou sobra, que ocorre quando se ajusta um modelo a um conjunto de dados. Os resíduos representam as distâncias entre os valores previstos no modelo e os valores observados nos dados. Para cada valor de  $y$  observado no conjunto de dados, você tem um valor previsto pelo modelo, normalmente conhecido como  $y$  *chapéu*,  $\hat{y}$ . O resíduo é a diferença entre os valores de  $y$  e  $y$  *chapéu*. Cada valor de  $y$  no conjunto de dados possui um resíduo; você deve examinar todos os resíduos em conjunto, procurando padrões ou valores muito altos (que indicam uma grande diferença entre o  $y$  observado e o  $y$  previsto para aquele ponto; veja o Capítulo 4 para informações mais completas sobre resíduos e seus gráficos).

A fim de que o modelo se ajuste bem, os resíduos devem satisfazer duas condições:

- ✓ **Os resíduos são independentes.** A independência dos resíduos significa que você não vai notar nenhum padrão em seu gráfico. Os resíduos não influenciam uns aos outros e devem ser aleatórios.
- ✓ **Os resíduos possuem uma distribuição normal centralizada em zero e os resíduos padronizados devem seguir o exemplo.** Uma distribuição normal com média zero significa que a maioria dos resíduos deve se centralizar ao redor de zero, havendo pouquíssimos distantes deste valor. Você deve observar aproximadamente a mesma quantia de resíduos acima e abaixo de zero. Se os resíduos estiverem padronizados, ou seja, seu desvio padrão como grupo for 1, você deve esperar que cerca de 95% deles se encontrem entre  $-2$  e  $+2$ , segundo a regra 68-95-99,7 (consulte seu livro de Estatística I).

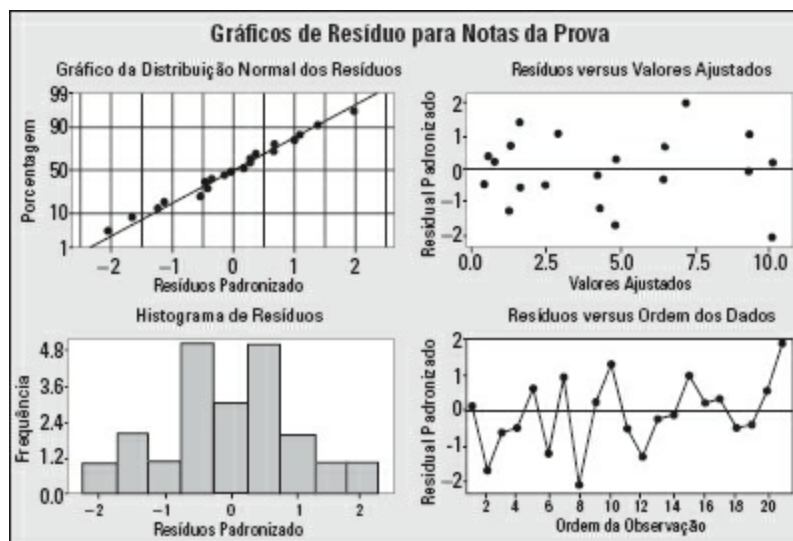
Para determinar se essas duas condições estão sendo atendidas ou não, você deve usar uma série de quatro gráficos chamados *gráficos de resíduos*. A maioria dos estatísticos prefere padronizar os resíduos (o que significa que eles são convertidos em escores- $Z$  ao serem subtraídos de suas médias e divididos por seu desvio padrão) antes de analisá-los, pois, assim, eles podem ser comparados aos valores da distribuição- $Z$ . Caso você também opte por isso, pode pedir ao Minitab que lhe dê uma série de quatro gráficos de resíduos padronizados com os quais vai poder checar as condições. (Veja no Capítulo 4 mais informações sobre os gráficos de resíduos e os gráficos de resíduos padronizados.)

A Figura 7-10 ilustra o gráfico de resíduo para o modelo quadrático, usando o conjunto de dados para as notas da prova das seções anteriores.

- ✓ O gráfico no canto superior esquerdo demonstra que os resíduos padronizados se parecem com uma distribuição normal, pois os pontos de dados e a distribuição normal se encaixam perfeitamente.

- ✓ O gráfico no canto superior direito demonstra que a maioria dos resíduos padronizados se encontra entre  $-2$  e  $+2$ . (Veja o Capítulo 4 para mais informações sobre resíduos padronizados.)
- ✓ O gráfico no canto inferior esquerdo mostra que os resíduos possuem uma certa semelhança com uma distribuição normal.
- ✓ O gráfico no canto inferior direito demonstra a ausência de um padrão entre os resíduos, que parecem ocorrer de forma aleatória.

Quando considerados em conjunto, todos esses gráficos sugerem que as condições impostas aos resíduos foram satisfeitas e o modelo de regressão quadrática selecionado pode ser aplicado.



**Figura 7-10:** Gráficos de resíduos padronizados para os dados em questão, usando o modelo quadrático.

## *Fazendo previsões*

Depois de ter encontrado o modelo que se ajusta bem, utilize-o para estimar o valor de  $y$  a partir de um dado valor de  $x$ . Basta substituir o valor de  $x$  desejado e o modelo lhe dá a previsão para o valor de  $y$ . (Certifique-se de que os valores atribuídos a  $x$  façam parte do conjunto de dados coletados, caso contrário, não é possível garantir o funcionamento do modelo.)

Voltando aos dados referentes às notas da prova, citados nas seções anteriores, é possível usar o tempo de estudo para prever a nota da prova usando um modelo de regressão quadrática. Olhando o diagrama de dispersão e o valor  $R^2$  ajustado (reveja as Figuras 7-8 e 7-9, respectivamente), vemos que o modelo de regressão quadrática parece se ajustar bem aos dados. (Não é ótimo quando você encontra alguma coisa que dá certo?) Os gráficos de resíduos na Figura 7-10 indicam que as condições parecem ter sido atendidas;

os resíduos não apresentam um padrão definido, parecem se centralizar em 1 e a maioria está dentro dos limites normais, de  $-2$  e  $+2$ , dos resíduos padronizados.

Considerando todas essas evidências juntas, nesse caso, o tempo de estudo realmente parece ter uma relação quadrática com a nota da prova. Agora, você pode usar o modelo para fazer estimativas da nota da prova, dado o tempo de estudo. Por exemplo, uma vez que o modelo (mostrado na Figura 7-8) é  $y = 9,82 - 6,15x + 1,00x^2$ , se o tempo de estudo for 5,5 horas (=5h30min), então, a estimativa para a nota da prova é  $9,82 - 6,15 * 5,5 + 1,00 * 5,5^2 = 9,82 - 33,83 + 30,25 = 6,25$  (=6h15min). Esse valor está em conformidade com o que se observa no gráfico da Figura 7-6, se olharmos o local onde  $x = 5,5$ ; os valores de  $y$  estão ao redor de 6 a 7.

Como em qualquer outro modelo de regressão, você não pode estimar o valor de  $y$  a partir de valores de  $x$  que estejam fora do conjunto de dados coletados. Se fizer isso, estará cometendo uma proibição denominada *extrapolação*, que se refere à tentativa de fazer previsões além de onde seus dados lhe permitem. Você não pode garantir que o modelo ajustado a seus dados realmente continue infinitamente para qualquer valor de  $x$ . Nesse exemplo (ver a Figura 7-6), você não pode estimar as notas da prova para um tempo de estudo maior do que seis horas usando esse modelo, pois os dados não mostraram ninguém que tenha estudado mais do que seis horas. Provavelmente, o modelo vai se estabilizar para uma nota dez depois de seis horas de estudo, portanto, estudar mais do que seis horas é suicídio! (No entanto, você nunca me ouviu falar isso!)



# *Subiu? Desceu? Então É Exponencial!*

Os modelos exponenciais funcionam bem em situações onde uma variável  $y$  tanto aumenta quanto diminui exponencialmente ao longo do tempo. Isso significa que a variável  $y$  tanto pode começar lenta e, então, aumentar em uma taxa cada vez mais rápida, ou começar alta e diminuir cada vez mais rápido.

Muitos processos no mundo real se comportam como um modelo exponencial: por exemplo, a mudança no tamanho da população ao longo do tempo, a renda familiar média, o tempo de duração de um produto ou o nível de paciência de uma pessoa à medida que os problemas nas aulas de Estatística aumentam.

Nessa seção, você vai se familiarizar com o modelo de regressão exponencial e aprender a usá-lo para ajustar os dados que aumentam ou diminuem em uma taxa exponencial. Também vai descobrir como construir e avaliar os modelos de regressão exponencial a fim de obter previsões precisas para uma variável de resposta  $y$  usando uma variável explicativa  $x$ .

## *Recordando os modelos exponenciais*

Os modelos exponenciais possuem a forma  $y = \alpha\beta^x$ . Esses modelos envolvem uma constante,  $\beta$ , elevada a potências cada vez mais altas de  $x$  e multiplicadas por uma constante,  $\alpha$ . A constante  $\beta$  representa a curvatura no modelo. Já a constante  $\alpha$  é um multiplicador na frente do modelo que mostra onde este intercepta o eixo de  $y$  (pois, quando  $x = 0$ ,  $y = \alpha * 1$ ).



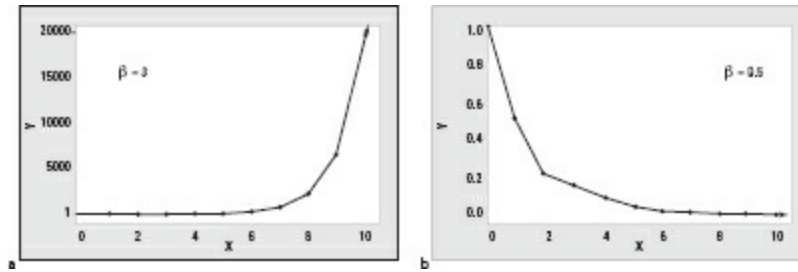
Um modelo exponencial geralmente se parece com a parte superior de uma hipérbole (lembra dela em álgebra avançada?). A *hipérbole* é uma curva que cruza o eixo  $y$  em um ponto, virando-se para baixo em direção ao zero ou se inicia em algum ponto e se curva para cima, rumo ao infinito (veja a Figura 7-11 para exemplos). Se  $\beta$  for maior do que 1, o gráfico é uma curva ascendente em direção ao infinito. Se  $\beta$  for menor do que 1, o gráfico é uma curva descendente em direção a zero. Todos os modelos exponenciais permanecem acima do eixo  $x$ .

Por exemplo, o modelo  $y = 1 * 3^x$  é um modelo exponencial. Vamos supor aqui que  $\alpha = 1$ , indicando que o modelo cruza o eixo  $y$  em 1 (pois, ao substituir  $x$  por 0 na equação, obtemos 1). Vamos estabelecer  $\beta$  como 3, o que indica que queremos que esse modelo tenha um pouco de curvatura. A curva dos valores  $y$  ascende rapidamente a partir do ponto (0,1). Por exemplo, quando  $x = 1$ , temos  $1 * 3^1 = 3$ ; para  $x = 2$ , temos  $1 * 3^2 = 9$ ; para  $x = 3$ , temos  $1 * 3^3 = 27$ ; e assim por diante. A Figura 7-11a mostra um gráfico desse modelo. Observe a grande escala necessária ao eixo  $y$  quando  $x$  é apenas 10.

Agora, suponha que  $\alpha = 1$  e  $\beta = 0,5$ . Esses valores lhe dão o modelo  $y = 1 * 0,5^x$ . Tal modelo eleva 0,5 (uma fração entre 0 e 1) a potências cada vez mais altas, começando em  $1 * 0,5^0 = 1$  e fazendo com que os valores de  $y$  fiquem cada vez menores, mas nunca elevado a zero,



embora cheguem cada vez mais perto. (Por exemplo, 0,5 ao quadrado é 0,25, valor menor do que 0,5, que elevado à décima potência é 0,00098.) A Figura 7-11b mostra um gráfico desse modelo.



**Figura 7-11:** O modelo de regressão exponencial para diferentes valores de  $\beta$ .

## *Em busca do melhor modelo exponencial*

A busca pelo modelo exponencial que melhor se ajusta requer um pouco mais de jogo de cintura se comparada à busca pela reta de melhor ajuste usada na regressão linear simples. Uma vez que ajustar um modelo linear é muito mais fácil do que ajustar um modelo exponencial diretamente a partir dos dados, é preciso transformar os dados em algo a que uma reta se ajuste. Depois, ajuste um modelo linear aos dados transformados. No final, você desfaz a transformação para voltar a um modelo exponencial.

Para a transformação, utilize *logaritmos*, pois são o inverso dos exponenciais. Mas antes de começar a suar, não se preocupe; essa ginástica matemática não é algo para se fazer à mão — o computador faz a maior parte do esforço por você.

O modelo exponencial vai se parecer com isto (se estiver usando a base 10):  $y = 10^{b_0 + b_1 x}$ ; observe que a equação da reta está no expoente. Siga este passo para ajustar um modelo exponencial a seus dados e use-os para fazer previsões:

A mágica matemática usada nestes passos é cortesia da definição do logaritmo, que diz  $\log_b(a) = y \Leftrightarrow b^y = a$ . Suponha que você tem a equação  $\log_{10} y = 2 + 3x$ . Se você elevar 10 em cada um dos lados, tem  $10^{\log_{10}(y)} = 10^{2+3x}$ . Segundo a definição do logaritmo, os dez do lado esquerdo pode ser cancelado, e você tem  $y = 10^{2+3x}$ . Esse modelo é exponencial, pois  $x$  está no expoente. Você pode dar um passo à frente e incluir a fórmula geral do modelo linear  $y = b_0 + b_1 x$ . Usando a definição do logaritmo nessa reta, você tem  $\log_{10}(y) = b_0 + b_1 x \Leftrightarrow 10^{b_0 + b_1 x} = y$ .

- 1. Faça o diagrama de dispersão dos dados e veja se eles aparentam ter um padrão em forma de curva que lembre uma curva exponencial.**

Se os dados seguirem uma curva exponencial, vá para o segundo passo; caso contrário, considere modelos alternativos (tais como a regressão múltipla comentada no Capítulo 5).





O Capítulo 4 lhe ensina a fazer um diagrama de dispersão usando o Minitab. Para mais detalhes sobre o formato buscado, consulte a seção anterior.

## 2. Use o Minitab para ajustar uma reta aos dados $\log(y)$ .



No Minitab, clique em regression model (curve fit). No menu Options, selecione Logten of  $y$ . Depois, selecione Using scale of logten para obter as unidades apropriadas para o gráfico.

O importante aqui é entender o que o Minitab faz durante esse processo; saber calcular isso à mão não é o que nos interessa. Veja o que o Minitab faz:

1. O Minitab aplica o  $\log$  (na base 10) aos valores de  $y$ . Por exemplo, se  $y$  é igual a 100,  $\log_{10} 100$  é igual a 2 (pois 10 ao quadrado é igual a 100). Observe que, se antes os valores de  $y$  ficaram próximos de um modelo exponencial, os valores de  $\log(y)$  ficarão próximos de um modelo linear. Esse fenômeno ocorre pois o logaritmo é o inverso da função exponencial, portanto, basicamente, um cancela o outro e você fica com uma reta.
2. O Minitab ajusta uma reta aos valores de  $\log(y)$  usando a regressão linear simples (veja o Capítulo 4). A equação da reta de regressão linear simples para os dados de  $\log(y)$  é  $\log(y) = b_0 + b_1x$ . O Minitab lhe passa esse modelo em sua saída e você começa desse ponto.
3. **Transforme o modelo de volta em um modelo exponencial, começando pelo modelo linear,  $\log(y) = b_0 + b_1x$ , ajustado aos dados  $\log_{10}(y)$  e, então, elevando a dez os lados esquerdo e direito da equação.**

Segundo a definição do logaritmo, você vai ter  $y$  do lado esquerdo do modelo e dez elevado a  $b_0 + b_1x$  do lado direito. O modelo exponencial resultante para  $y$  é  $y = 10^{b_0 + b_1x}$ .

4. **Use o modelo exponencial do terceiro passo para fazer as previsões para  $y$  (sua variável original), bastando para isso substituir os valores atribuídos a  $x$  no modelo.**

Apenas atribua valores para  $x$ .

5. **Para avaliar o ajuste do modelo, observe o diagrama de dispersão dos dados  $\log(y)$ , o valor de  $R^2$  (ajustado) da reta para  $\log(y)$  e os gráficos de resíduo para os dados  $\log(y)$ .**

As técnicas e os critérios usados para isso são os mesmos que discuto na seção anterior: “Avaliando o ajuste de um modelo polinomial”.

Caso ache esses passos dúbios, cole comigo. O exemplo na próxima seção vai lhe mostrar em primeira mão cada passo, o que vai ajudar muito. No final, você vai ver que fazer previsões usando um modelo exponencial é bem mais fácil do que explicar como fazer.

# ***Espalhando segredos de forma exponencial***

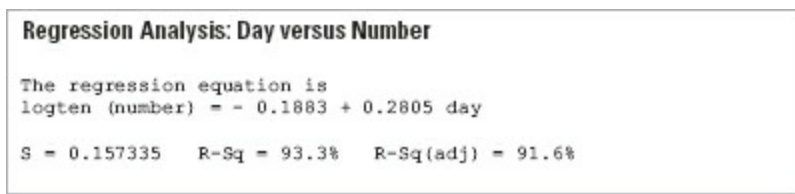
Muitas vezes, a melhor forma de entender algo é vê-lo em ação. Usando o exemplo da difusão de um segredo da Figura 7-2, vamos poder trabalhar os passos da seção anterior a fim de encontrar o modelo exponencial que melhor se ajusta e usá-lo para fazer nossas previsões.

## ***Primeiro passo: Verificar o diagrama de dispersão***

O objetivo no primeiro passo é fazer um diagrama de dispersão dos dados e determinar se sua curva se parece com a de uma função de um modelo exponencial. A Figura 7-2 mostra os dados para um número de pessoas que sabem um segredo em função do número de dias. Podemos ver que o número de pessoas que conhecem o segredo começa pequeno, mas, à medida que cada vez mais pessoas dividem-no com mais e mais pessoas, esse número aumenta rapidamente, até que o segredo passa a não ser mais um segredo. Esse é um bom contexto para um modelo exponencial, graças ao tanto que a curva cresce nesse gráfico.

## ***Segundo passo: Deixe esse por conta do Minitab***

No segundo passo vamos deixar o Minitab encontrar a reta de regressão para os dados  $\log(y)$  (veja a seção "Em busca do melhor modelo exponencial" para saber como fazer isso no Minitab). A saída para a análise dos dados referentes à difusão de um segredo está na Figura 7-12; você poderá ver que a reta com o melhor ajuste é  $\log(y) = -0,19 + 0,28 * x$ , onde  $y$  é o número de pessoas que sabem o segredo e  $x$  é o número de dias.



---

**Figura 7-12:** O Minitab ajusta uma reta ao  $\log(y)$  para os dados referentes à difusão de um segredo.

---

## ***Terceiro passo: Vai, exponencial!***

Com a saída do Minitab em mãos, você está pronto para dar o terceiro passo. Transforme o modelo  $\log(y) = -0,19 + 0,28 * x$  em um modelo para  $y$ , bastando elevar a 10 o lado esquerdo e o lado direito. Transformando a equação  $\log(y)$ , temos  $y = 10^{-0,19 + 0,28x}$ .

## ***Quarto passo: Fazendo previsões***

Com o modelo exponencial do terceiro passo, você pode dar o quarto: fazer previsões para valores adequados de  $x$  (que estejam dentro do conjunto de dados coletados). Usando ainda os dados do exemplo em questão, suponha que você queira estimar o número de

pessoas que conhecem o segredo no quinto dia (veja a Figura 7-2). Basta substituir  $x = 5$  no modelo exponencial para obter  $y = 10^{-0,19 + 0,28x} = 10^{1,21} = 16,22$ . Voltando a observar a Figura 7-2, é possível notar que essa estimativa se alinha aos dados no gráfico.

### ***Quinto passo: Avaliando o ajuste de seu modelo exponencial***

Agora que já encontrou o modelo exponencial de melhor ajuste, o pior já passou. Você chegou ao quinto passo e está pronto para avaliar o ajuste do modelo (além do diagrama de dispersão dos dados originais) para garantir que nenhum problema ocorra.

De forma geral, para avaliar o ajuste de um modelo exponencial, você vai observar a reta de ajuste do  $\log(y)$ . Basta usar esses três itens (em qualquer ordem) da mesma forma descrita na seção "Avaliando o ajuste de um modelo polinomial".

- ✓ **Verifique o diagrama de dispersão dos dados de  $\log(y)$  para ver o quanto eles se parecem com uma reta.** Primeiro, avalie o ajuste de  $\log(y)$  através do diagrama de dispersão mostrado na Figura 7-13. O diagrama demonstra que o modelo parece ajustar bem os dados, pois os pontos estão difundidos bem ao redor da reta.

Durante esse processo, os dados também foram transformados. Começamos com  $x$  e  $y$  e agora temos  $x$  e  $\log(y)$ . Veja  $x$ ,  $y$  e  $\log(y)$  para os dados do exemplo na Tabela 7-2.

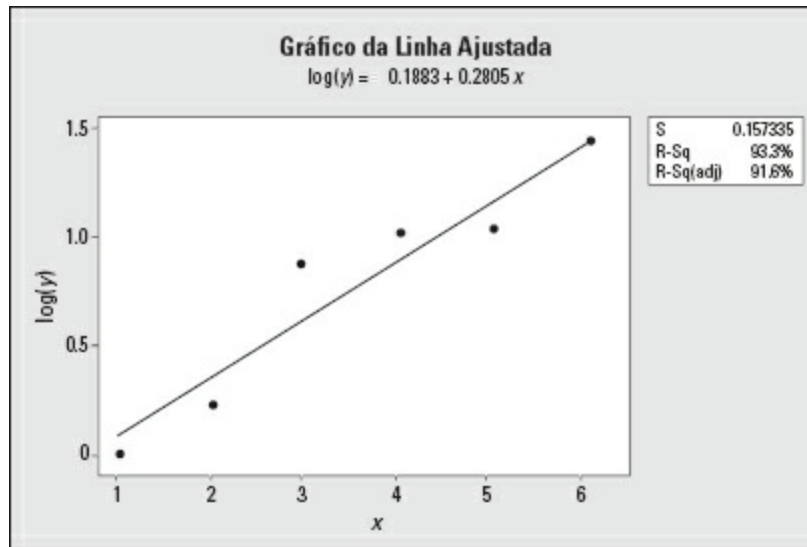
- ✓ **Examine o valor  $R^2$  ajustado para o modelo da reta com o melhor ajuste para  $\log(y)$  feito pelo Minitab.** O valor de  $R^2$  ajustado para esse modelo está na Figura 7-13 como sendo 91,6%. Esse valor também indica um bom ajuste, pois está muito próximo de 100%. Assim, 91,6% da variação no número de pessoas que sabem o segredo é explicado pela quantidade de dias passados desde que o segredo começou a ser espalhado. (Faz sentido).
- ✓ **Observe os gráficos de resíduo para o ajuste da reta aos dados  $\log(y)$ .** Os gráficos de resíduo resultantes dessa análise (veja a Figura 7-14) não mostram grandes desvios da condição que impõe a independência dos resíduos e estes têm uma distribuição normal. Observe que o histograma no canto inferior esquerdo não se parece com um sino, mas também não temos muitos dados nesse exemplo e os demais gráficos parecem estar OK. Portanto, você não tem muito com o que se preocupar.

**Tabela 7-2      Log(y) Valores para Dados Referentes à Difusão de um Segredo**

<i>x (Dia)</i>	<i>y (Número de pessoas)</i>	<i>log(y)</i>
1	1	0,00
2	2	0,30
3	5	0,70
4	7	0,85

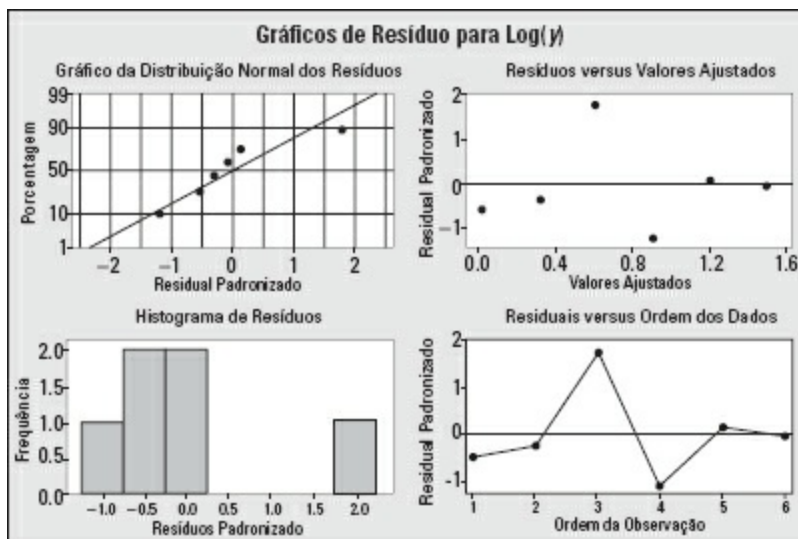


5	17	1,23
6	30	1,48



**Figura 7-13:** Diagrama de dispersão mostrando o ajuste de uma reta aos dados  $\log(y)$ .

No final das contas, agora que você tem um modelo exponencial que explica como tudo acontece, parece que o segredo aqui já foi desvendado.



**Figura 7-14:** Gráficos de resíduos mostrando o ajuste de uma reta aos dados  $\log(y)$ .

# Capítulo 8

## Sim, Não, Talvez: Fazendo Previsões Usando a Regressão Logística

---

### *Neste Capítulo*

- ▶ Aprendendo a hora certa de usar a regressão logística
  - ▶ Construindo modelos de regressão logística para dados de sim ou não
  - ▶ Verificando as condições do modelo e tirando as conclusões certas
- 

**T**odo mundo (inclusive esta que vos fala) tenta fazer previsões sobre a ocorrência ou não de um determinado evento. Por exemplo, qual é a probabilidade de chuva para este fim de semana? Quais são as chances de seu time ganhar o próximo jogo? Qual a probabilidade de eu ter complicações durante uma cirurgia? Essas previsões são muitas vezes baseadas em *probabilidades*, o percentual de vezes a longo prazo que um evento está previsto para acontecer.

No fim das contas, o que você quer é estimar  $p$ , a probabilidade de que um evento ocorra. Neste capítulo, você vai ver como construir e testar modelos para  $p$  com base em um conjunto de variáveis explicativas ( $x$ ). Essa técnica é chamada de *regressão logística* e neste capítulo explico como fazer bom uso dela.

# Entendendo o Modelo de Regressão Logística

Na regressão logística, estima-se a probabilidade de que um evento ocorra para um indivíduo aleatoriamente selecionado versus a probabilidade de que ele não ocorra. Basicamente, você vai observar dados de sim ou não: sim, ocorreu ( $p$  = probabilidade), ou não, não ocorreu (probabilidade =  $1 - p$ ). Os dados de sim ou não que provêm de uma amostra aleatória têm uma distribuição binomial com probabilidade de sucesso (de que o evento ocorra) igual a  $p$ .

Nos problemas envolvendo binômios, vistos em Estatística I, você tinha uma amostra de tamanho  $n$  tentativas, dados de sim ou não, e uma probabilidade de sucesso em cada tentativa, denotada por  $p$ . Em seu curso de Estatística I, para qualquer problema binomial, o valor de  $p$  era de alguma forma dado como um determinado valor, como no caso de uma moeda não viciada que tem probabilidade  $p = 0,50$  para virar do lado cara. Mas, em Estatística II, operamos em um cenário mais realista. De fato, como  $p$  não é conhecido, seu trabalho é estimá-lo através de um modelo.



Para estimar  $p$ , a probabilidade de que um evento ocorra, você precisa de dados que vêm sob a forma de sim ou não, indicando se o fato aconteceu ou não para cada indivíduo no conjunto de dados.

Uma vez que os dados sim ou não não possuem distribuição normal, condição necessária para outros tipos de regressão, você vai precisar de um novo tipo de modelo de regressão para fazer este trabalho, que é o modelo de *regressão logística*.

## Qual é a diferença entre a regressão logística e as outras regressões?

Você usa a regressão logística quando uma variável quantitativa é usada para prever ou adivinhar o resultado de algumas variáveis categóricas com apenas dois resultados (por exemplo, usando a pressão barométrica para prever se irá ou não chover).

Um modelo de regressão logística, em última instância, lhe dá uma estimativa para  $p$ , a probabilidade de que um determinado resultado ocorra em uma situação de sim ou não (por exemplo, a chance de chover versus não). A estimativa se baseia em informações de uma ou mais variáveis explicativas; você pode chamá-las de  $x_1, x_2, x_3, \dots, x_k$ . (por exemplo,  $x_1$  = umidade,  $x_2$  = pressão barométrica,  $x_3$  = presença de nuvens... e  $x_k$  = velocidade do vento).

Já que se trata do uso de uma variável ( $x$ ) para fazer a previsão de outra ( $y$ ), você pode estar pensando em usar regressão — e está corretíssimo. No entanto, há muitos tipos de regressão para escolher, e você precisa determinar qual é o mais adequado a essa situação. Você vai precisar do tipo de regressão que usa uma variável quantitativa ( $x$ ) para prever o resultado de uma variável categórica ( $y$  que tenha apenas dois resultados (sim ou não)).

Assim, como um bom aluno em Estatística II, você busca em sua lista de técnicas estatísticas algo sobre regressão — e imediatamente vê mais do que um tipo.

- ✓ Primeiro, vai ver a regressão linear simples. Não, essa só é usada quando uma variável quantitativa prevê outra variável quantitativa (veja o Capítulo 4).
- ✓ Regressão múltipla? Não, esse método só expande a regressão linear simples para adicionar mais variáveis  $x$  (veja o Capítulo 5).
- ✓ Regressão não linear? Deixe-me ver... não. Essa também trabalha com duas variáveis quantitativas, só que os dados formam uma curva, e não uma reta.

Mas, então, você se depara com a regressão logística e...eureka! Você vê que ela lida com situações onde a variável  $x$  é numérica, e a variável  $y$  é categórica com duas possíveis categorias. Justamente o que estava procurando!

A regressão logística, em essência, estima a probabilidade de que  $y$  esteja em uma ou outra categoria, baseando-se no valor de algumas variáveis quantitativas,  $x$ . Por exemplo, suponha que você queira prever a altura de alguém de acordo com seu sexo. Já que o gênero é uma variável categórica, use a regressão logística para fazer essas previsões. Supondo que 1 indique masculino, parte-se do princípio de que pessoas com probabilidade maior do que 0,5 de serem do sexo masculino (com base em sua altura) sejam do sexo masculino, pessoas com probabilidade menor do que 0,5 de serem do sexo masculino (com base em sua altura) sejam do sexo feminino.

Neste capítulo, apresento o que acontece quando só uma variável explicativa é usada para prever um resultado. Você pode estender essas ideias exatamente da mesma maneira feita no caso do modelo de regressão linear simples para o modelo de regressão múltipla.

## *Utilizando uma curva em $S$ para estimar as probabilidades*

No modelo de regressão linear simples, a equação geral da reta é  $y = \beta_0 + \beta_1 x$ , e  $y$  é uma variável quantitativa. No modelo de regressão logística, a variável  $y$  é categórica, e não quantitativa. O que se está estimando, no entanto, não é a categoria em que o indivíduo se encontra, mas, sim, a probabilidade de que o indivíduo se encontre em uma determinada categoria. Assim, o modelo de regressão logística é baseado na estimativa dessa probabilidade, chamada  $p$ .

Se tivesse que estimar  $p$  usando um modelo de regressão linear simples, você poderia pensar que teria que tentar ajustar uma reta,  $p = \beta_0 + \beta_1 x$ . No entanto, não faz sentido usar uma reta para estimar a probabilidade da ocorrência de um evento com base em outra variável, devido aos seguintes motivos:

- ✓ Os valores estimados para  $p$  nunca podem estar além de  $[0,1]$ , o que vai contra a ideia de uma reta (uma reta é infinita em ambos os sentidos).
- ✓ Não faz sentido forçar o valor de  $p$  para, com base em  $x$ , crescer de forma

**linear.** Por exemplo, um evento pode ocorrer com muita frequência com uma série tanto de valores muito grandes, como com uma série de valores muito pequenos para  $x$ , com pouquíssimas chances de que o evento aconteça no espaço entre elas. Esse tipo de modelo teria um gráfico em forma de U, em vez de formar uma reta.

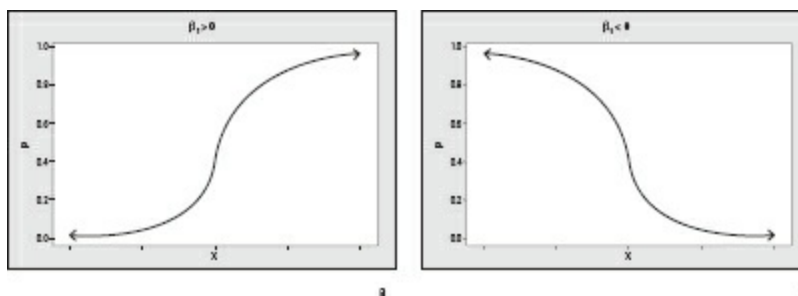
Para chegar a um modelo mais apropriado, os estatísticos criaram uma nova função cujo gráfico é chamado de curva-S. A *curva-S* é uma função que envolve  $p$ , mas também envolve  $e$  (logaritmo natural), bem como uma razão entre as duas funções.

Os valores da curva-S sempre se encaixam entre 0 e 1, o que permite que a probabilidade,  $p$ , mude de baixa para alta ou de alta para baixa, de acordo com uma curva em forma de S.

A equação geral do modelo de regressão logística baseada na curva-S é  $p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$ .

## ***Interpretando os coeficientes do modelo de regressão logística***

O sinal do parâmetro  $\beta_1$  informa a direção da curva S. Se  $\beta_1$  for positivo, a curva-S vai de baixo para cima (veja a Figura 8-1a);  $\beta_1$ , se for negativo, a curva-S vem de cima para baixo (Figura 8-1b).



**Figura 8-1:** Dois tipos básicos de curvas S.

A magnitude de  $\beta_1$  (indicada por seu valor absoluto) informa o quanto de curvatura o modelo apresenta. Valores altos indicam uma curvatura acentuada, e valores baixos indicam curvatura gradual. O parâmetro  $\beta_0$  apenas desloca a curva S para o local adequado a fim de ajustá-la aos dados. Ele demonstra o ponto onde os valores  $x$  mudam de uma probabilidade alta para baixa e viceversa.

## ***O modelo de regressão linear em ação***

Muitas vezes, a melhor forma de entender algo é vê-lo em ação. Nesta seção, vou lhe dar o exemplo de uma situação em que você poderá usar um modelo de regressão logística para estimar uma probabilidade. (Desenvolvo este exemplo mais adiante. Por enquanto, estou apenas estabelecendo um cenário para a regressão logística.)

Suponha que produtores de cinema queiram estimar as chances de que alguém goste de um determinado filme familiar e que você acredite que a idade possa ter algo a ver com isso.



Traduzindo essa pergunta de pesquisa para  $x$ 's e  $y$ 's, a variável de resposta ( $y$ ) representa se uma pessoa vai ou não apreciar o filme, e a variável explicativa ( $x$ ) representa a idade da pessoa. O que você quer estimar é  $p$ , a probabilidade de alguém gostar do filme.

Você coleta dados a partir de uma amostra aleatória com 40 pessoas, mostrada na Tabela 8-1. Com base em seus dados, parece que os jovens gostaram mais do filme do que os mais velhos e que, a partir de certa idade, a tendência muda de gostar do filme para não gostar dele. Com esses dados em mãos, é possível construir um modelo de regressão logística para estimar  $p$ .

**Tabela 8-1**    **Pessoas que Gostaram do Filme (Sim ou Não) versus Idade**

<i>Idade</i>	<i>Gostaram do Filme</i>	<i>Número Total da Amostra</i>
10	3	3
15	4	4
16	3	3
18	2	3
20	2	3
25	2	4
30	2	4
35	1	5
40	1	6
45	0	3
50	0	2

# *Fazendo uma Análise de Regressão Logística*

A ideia básica de qualquer processo de ajuste do modelo é a de observar todos os possíveis modelos no formato da equação geral e encontrar uma que se ajuste melhor a seus dados.

A equação geral do modelo de regressão logística de melhor ajuste é  $\hat{p} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$ , onde  $\hat{p}$  é a estimativa para  $p$ ,  $b_0$  é a estimativa para  $\beta_0$  e  $b_1$  é a estimativa para  $\beta_1$  (da seção anterior “Usando uma curva-S para estimar probabilidades”). Os únicos valores que você tem à sua escolha para construir seu modelo são os valores de  $b_0$  e  $b_1$ . Esses são os valores que você está tentando estimar através da análise de regressão logística.

Para encontrar o modelo de regressão logística que melhor se ajusta a seus dados, siga os passos abaixo:

- 1. Execute uma análise de regressão logística sobre os dados coletados (ver próxima seção).**
- 2. Encontre os coeficientes da constante e de  $x$ , onde  $x$  é o nome de sua variável explicativa.**

Esses coeficientes são  $b_0$  e  $b_1$ , as estimativas para  $\beta_0$  e  $\beta_1$  no modelo de regressão logística.

- 3. Substitua os coeficientes do primeiro passo no modelo de regressão logística:**

$$\hat{p} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

Essa equação será o seu modelo de regressão logística que melhor se ajusta aos dados. Seu gráfico é uma curva em S. (Para mais informações sobre a curva-S, consulte a seção “Usando uma curva-S para estimar probabilidades” neste capítulo.)

Nas seções seguintes, você verá como fazer com que o Minitab realize os passos acima por você. Você também vai ver como interpretar a saída resultante, encontrar a equação do modelo de regressão logística que melhor se ajusta e usar esse modelo para fazer previsões (sempre consciente de que todas as condições devem ser satisfeitas).

## *Fazendo a análise no Minitab*

Veja como realizar uma regressão logística utilizando o Minitab (outros programas estatísticos são semelhantes):

- 1. Insira os dados na planilha sob a forma de uma tabela que relaciona cada valor da variável  $x$  na coluna 1, o número de “sins” para esse valor de  $x$  na coluna 2 e o número total de tentativas para esse valor de  $x$  na coluna 3.**

Essas duas últimas colunas representam o resultado da variável de resposta  $y$ . (Para um exemplo de como inserir seus dados, consulte a Tabela 8-1 referente aos dados da



relação filme versus idade.)

2. Clique em Stat>Regression>Binary Logistic Regression.
3. Ao lado da opção Sucess, selecione o nome da variável da coluna 2 e, ao lado de Trial, selecione o nome da variável para a coluna 3.
4. Embaixo de Model, selecione o nome da variável da coluna um, porque essa é a coluna que contém a variável explicativa ( $x$ ) em seu modelo.
5. Clique em OK e você tem sua saída para a regressão logística.

Quando você ajusta um modelo de regressão logística a seus dados, a saída do software se constitui em duas partes principais:

- ✓ **A parte da construção do modelo:** Nessa parte da saída você poderá encontrar os coeficientes  $b_0$  e  $b_1$ . (Descrevo os coeficientes na próxima seção.)
- ✓ **A parte do ajuste do modelo:** Aqui você poderá ver os resultados de um teste do Qui-quadrado de qualidade de ajuste (ver Capítulo 15), bem como a porcentagem de pares concordantes e discordantes nesta seção da saída. (O *par concordante* indica que o resultado previsto pelo modelo corresponde ao resultado observado a partir dos dados. O *par discordante* é aquele que não corresponde.)

No caso dos dados sobre a relação filme versus idade, a parte da construção do modelo na saída do Minitab é mostrada na Figura 8-2. A parte referente ao ajuste do modelo na saída do Minitab resultante da análise de regressão logística está na Figura 8-4.

Nas seções seguintes você verá como usar essa saída para construir o modelo de regressão logística que melhor se ajusta a seus dados e também verificar esse ajuste.

Logistic Regression Table							
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
Constant	4.86539	1.43434	3.39	0.001			
Age	-0.175745	0.0499620	-3.52	0.000	0.84	0.76	0.93

**Figura 8-2:** Parte referente à construção do modelo para dados sobre a relação filme versus idade na saída do Minitab.

## *Encontrando os coeficientes e construindo o modelo*

Depois de fazer o Minitab executar uma análise de regressão logística em seus dados, você pode encontrar os coeficientes  $b_0$  e  $b_1$  e juntá-los para formar o modelo de regressão logística que melhor se ajusta a seus dados.

Figura 8-2 mostra a parte da saída do Minitab para os dados referentes à relação entre filme versus idade. (Discuto a saída restante na seção “Verificando o ajuste do modelo”.) A primeira coluna de números é chamada Coef, que representa os coeficientes do modelo.

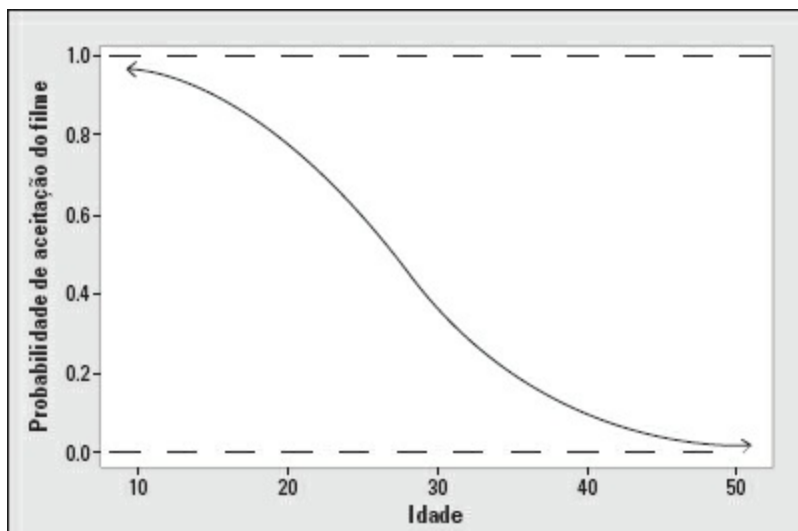
O primeiro coeficiente,  $b_0$ , é nomeado Constante. O segundo coeficiente está na linha nomeada pela variável explicativa,  $x$ . (Nos dados da relação filme versus faixa etária, a variável explicativa é a idade. Esse coeficiente de idade representa o valor de  $b_1$  no modelo.)

De acordo com a saída do Minitab, ilustrada na Figura 8-2, o valor de  $b_0$  é 4,87, e o valor de  $b_1$  é -0,18. Depois de determinar os coeficientes  $b_0$  e  $b_1$  a partir da saída do Minitab para encontrar a curva-S que melhor se ajuste a seus dados, coloque esses dois coeficientes no modelo geral de regressão logística:  $\hat{p} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$ . Para os dados referentes à relação filme versus idade, você obtém  $\hat{p} = \frac{e^{4,87 - 0,18x}}{1 + e^{4,87 - 0,18x}}$ , o modelo de regressão logística com o melhor ajuste para esse conjunto de dados.

O gráfico para o modelo de regressão logística com o melhor ajuste para os dados sobre a relação filme versus idade é ilustrado na Figura 8-3. Observe que o gráfico é uma curva-S com inclinação descendente, pois as maiores probabilidades de alguém gostar do filme se associam às idades inferiores, e as menores probabilidades se associam às idades superiores.

Os produtores agora têm a resposta para sua pergunta. Esse filme tem uma grande chance de ser bem aceito por crianças (e quanto mais jovem, melhor) e uma pequena chance de ser benquisto pelos adultos (e, quanto mais velhos forem, menor será a chance de que gostem do filme).

O ponto onde a probabilidade muda de alta para baixa (ou seja = 0,50) está entre as idades de 25 e 30, ou seja, a probabilidade de alguém gostar do filme parece mudar de mais alta para mais baixa justamente nessa faixa etária. Em cálculo, esse ponto é chamado de *ponto de sela* da curva-S, que é o ponto onde a concavidade do gráfico que era voltada para cima passa a ser voltada para baixo, ou vice-versa.



**Figura 8-3:** Curva-S que melhor se ajusta aos dados para a relação filme versus idade.

## Estimando $p$

Depois de ter determinado o modelo de regressão logística que melhor se ajusta a seus dados, obtido os valores de  $b_0$  e  $b_1$  a partir da análise de regressão logística e encontrado a curva-S que melhor se ajusta aos dados (veja a seção anterior), você está pronto para estimar  $p$  e fazer previsões sobre a probabilidade de que o evento de interesse ocorra, dado o valor da variável explicativa  $x$ .

Para estimar  $p$  usando um determinado valor de  $x$ , substitua esse valor de  $x$  na equação (o modelo de regressão logística com melhor ajuste) e solucione-a usando suas habilidades em álgebra. O número que você obtiver é a estimativa da chance de ocorrência do evento para esse valor de  $x$ , e deve ser um número entre 0 e 1.

Continuando com o exemplo das seções anteriores, suponha que você queira prever se uma pessoa de 15 anos vai gostar do filme. Para estimar  $p$ , substitua  $x$  por 15 no modelo de regressão logística  $\hat{p} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$  para obter  $\hat{p} = \frac{e^{4,87 - 0,18 \times 15}}{1 + e^{4,87 - 0,18 \times 15}} = \frac{e^{2,17}}{1 + e^{2,17}} = \frac{8,76}{9,76} = 0,90$ .

Essa resposta significa que, segundo sua estimativa, a probabilidade de que um jovem de 15 anos goste do filme é de 90%. Você pode ver na Figura 8-3 que, quando  $x$  é 15,  $p$  é aproximadamente 0,90. Por outro lado, se a pessoa tiver 50 anos, a chance de que ela goste desse filme é  $\hat{p} = \frac{e^{4,87 - 0,18 \times 50}}{1 + e^{4,87 - 0,18 \times 50}}$ , ou 0,02 (mostrado na Figura 8-3 para  $x = 50$ ), o que representa uma chance de apenas 2%.

## Verificando o ajuste do modelo

Os resultados obtidos a partir de uma análise de regressão logística, como em qualquer análise de dados, estão sujeitos ao ajustamento do modelo.

Para determinar se seu modelo de regressão logística se ajusta ou não, siga estes passos (sobre os quais falarei com mais detalhes ainda nesta seção):

1. **Localize o valor- $p$  do teste da qualidade de ajuste (encontrado na parte Goodness-of-Fit na saída do programa; veja a Figura 8-4 para um exemplo). Se seu valor- $p$  for maior do que 0,05, concluímos que o modelo se ajusta, mas, se o valor- $p$  for menor do que 0,05, concluímos o oposto.**
2. **Encontre o valor- $p$  para o coeficiente  $b_1$  (que está embaixo de  $P$  na linha para a coluna 1 (explicativas) da variável na parte referente à construção de modelos da saída do Minitab; veja a Figura 8-2 para um exemplo). Se o valor- $p$  for menor do que 0,05, a variável  $x$  é estatisticamente significativa ao modelo e, por isso, deve ser incluída. Se o valor- $p$  for maior ou igual a 0,05, a variável  $x$  não é estatisticamente significativa e não deve ser incluída no modelo.**
3. **Depois, procure a porcentagem de pares concordantes. Essa porcentagem reflete a proporção de tempo em que dados e modelo realmente concordam entre si.**



## Quanto maior a porcentagem, melhor o ajuste do modelo.



A conclusão tirada no primeiro passo, e que se baseia no valor- $p$ , pode parecer estar de trás para frente, mas veja aqui o que acontece: os testes Qui-quadrado de qualidade de ajuste medem a diferença total entre o que você espera ver através do seu modelo e o que você realmente observa em seus dados. (O Capítulo 15 mostra tudo o que você precisa saber sobre os testes Quiquadrado.) A hipótese nula ( $H_0$ ) para esse teste diz que existe uma diferença igual a zero entre o que foi observado e o que você esperava prever a partir do modelo; isto é, o modelo se ajusta. A hipótese alternativa, denotada por  $H_a$ , afirma que o modelo não se ajusta. Se o valor- $p$  obtido for pequeno (menor do que 0,05), rejeite  $H_0$  e conclua que o modelo não se ajusta. Se o valor- $p$  obtido for grande (maior do que 0,05), pode ficar com seu modelo.



Aqui, a impossibilidade de rejeitar  $H_0$  (obter um valor- $p$  grande) apenas significa que você não pode dizer que seu modelo não se ajusta à população de onde a amostra foi retirada. Isso não significa necessariamente que o modelo tem um ajuste perfeito. Os dados é que podem não estar representando a população simplesmente em virtude da casualidade.

## Ajustando o modelo

Você está pronto para verificar o ajuste aos dados a fim de garantir seu emprego depois que o faturamento da bilheteria for contabilizado.

### *Primeiro passo: valor- $p$ para o Qui-quadrado*

Usando a Figura 8-4 para completar a primeira etapa de verificação do ajuste do modelo, você poderá ver diversos testes de qualidade de ajuste. As informações relativas a cada um destes testes estão fora do escopo deste livro, porém, neste caso (como na maioria dos casos), cada teste teve resultados numéricos ligeiramente diferentes e as mesmas conclusões foram tiradas.

Todos os valores- $p$  na coluna quatro da Figura 8-4 estão acima de 0,80, valor muito maior do que o 0,05 que você precisa para rejeitar o modelo. Depois de observar os valores- $p$ , o modelo que usa a idade para prever a aceitação de um filme realmente parece se ajustar a esses dados.

Goodness-of-Fit Test				
Method	Chi-Square	DF	P	
Pearson	2.83474	9	0.970	
Deviance	3.63590	9	0.934	
Hosmer-Lemeshow	2.75232	6	0.839	
Measures of Association:				
(Between the Response Variable and Predicted Probabilities)				
Pairs	Number	Percent	Summary Measures	
Concordant	349	87.3	Somers' D	0.80
Discordant	30	7.5	Goodman-Kruskal Gamma	0.84
Ties	21	5.3	Kendall's Tau-a	0.41
Total	400	100.0		


**Figura 8-4:** Parte referente ao ajuste do modelo para dados sobre a relação filme versus idade na saída do Minitab.

## ***Segundo passo: valor-p para variável $x$***


No segundo passo, vamos observar a significância da variável  $x$ , a idade. Voltando à Figura 8-2, é possível ver a constante para a idade,  $-0,18$ , e, mais adiante, ao longo de sua linha, você pode ver que o valor- $Z$  é  $-3,52$ ; este valor- $Z$  é a estatística de teste para testar  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$ . O valor- $p$  é dado como  $0,000$ , o que significa que ele é menor do que  $0,001$  (um número altamente significativo). Assim, você conclui que o coeficiente na frente de  $x$ , também conhecido como  $\beta_1$ , é estatisticamente significativo (não é igual a zero) e, portanto, você deve incluir  $x$  (idade) no modelo.

## ***Terceiro passo: pares concordantes***

Para completar a terceira etapa do processo de verificação do ajuste, examine o percentual de pares concordantes apresentados na Figura 8-4. Este valor indica a porcentagem de vezes que os dados realmente concordaram com o modelo ( $87,3$ ). Para determinar a concordância, o computador, baseando-se no modelo, faz previsões a fim de saber se o evento deveria ter ocorrido para cada indivíduo e compara esses resultados com o que realmente aconteceu.

 O modelo de regressão logística procura o  $p$ , a probabilidade de que um evento ocorra. Portanto, se o valor estimado para  $p$ , a partir de um dado valor de  $x$ , for  $> 0,50$ , o computador prevê que o evento irá ocorrer (versus não ocorrer). Se o valor estimado para  $p$  for  $< 0,50$ , dado um determinado valor  $x$ , o computador prevê que o evento não irá ocorrer.

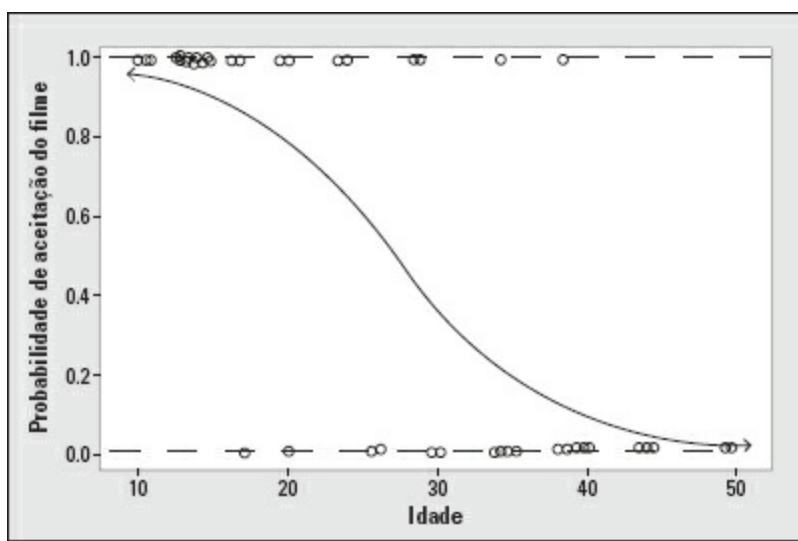
Para os dados referentes à relação filme e idade, a porcentagem de pares concordantes (ou seja, a porcentagem de vezes que o modelo tomou a decisão correta ao prever o que iria acontecer) é de  $87,3\%$ , um valor bastante alto.

 A porcentagem de pares concordantes foi obtida através da divisão do número de pares concordantes pelo número total de pares. Eu começo a me animar quando a porcentagem de pares concordantes é maior do que  $75\%$ ; quanto mais melhor.

A Figura 8-5 mostra o modelo de regressão logística para os dados referentes à relação filme e idade com os valores reais dos dados observados na forma de círculos. A curva-S mostra a probabilidade de aceitação do filme para cada faixa etária, e o computador prevê “1” = vão gostar do filme, quando  $p > 0,50$ . Os círculos indicam se as pessoas dessas faixas etárias realmente gostaram do filme ( $y = 1$ ) ou não ( $y = 0$ ).

Na maior parte do tempo, o modelo tomou decisões corretas; as probabilidades acima de  $0,50$  estão associadas a mais círculos em 1, e as probabilidades abaixo de  $0,50$  estão associadas a mais círculos em zero. Os resultados em que  $p$  está próximo de  $0,50$  são difíceis de prever, pois podem tomar qualquer direção.





**Figura 8-5:** Valores observados (0 e 1) comparados ao modelo

## Qual método usar para comparar? Classificando as situações semelhantes

Os dados vêm sob uma variedade de formas, e cada uma tem a sua própria análise para fazer comparações. Sendo assim, fica difícil decidir qual tipo de análise usar e quando usá-la.

A classificação de algumas situações que parecem semelhantes, mas possuem diferenças sutis que levam a análises muito diferentes, pode ajudar nesse processo. Você pode usar a lista a seguir para comparar essas diferenças sutis, mas importantes:

- ✓ **Caso queira comparar três ou mais grupos de variáveis numéricas**, utilize a ANOVA (veja o Capítulo 10). Para apenas dois grupos, utilize o teste-*t* (veja os Capítulos 3 e 9).
- ✓ **Caso queira estimar uma variável numérica com base em outra**, utilize a regressão linear simples (veja o Capítulo 4).
- ✓ **Caso queira estimar uma variável numérica usando outras variáveis numéricas**, utilize regressão múltipla (veja o Capítulo 5).
- ✓ **Caso queira estimar uma variável categórica usando uma variável numérica**, utilize a regressão logística, o foco deste capítulo, é claro.
- ✓ **Caso queira comparar duas variáveis categóricas**, utilize um teste de Quiquadrado (veja o Capítulo 14).

Todas essas evidências ajudam a confirmar que seu modelo se ajusta bem aos seus dados. Pode seguir em frente e fazer previsões com base neste modelo para o próximo indivíduo, cujo resultado você desconhece (veja a seção “Estimando *p*”, já mencionada neste capítulo).



# **Parte III:** **Analisando a Variância com ANOVA**

**A 5ª Onda**

Por Rich Tennant



**"Estes são o meu antigo professor de Estatística II, sua esposa, Doris, e seus dois filhos, Wilcox e Kruskal."**



## *Nesta parte...*

**V**ocê vai aprender tudo o que precisa para entender a análise de variância com um e com dois fatores (também conhecida como ANOVA), a qual compara as médias de várias populações de uma só vez com base em uma ou duas características diferentes. Também vai ver como interpretar as tabelas ANOVA e a saída do software, além de começar a visitar os bastidores a fim de entender as grandes ideias que estão por trás das fórmulas usadas na análise de variância. E, para terminar, vai poder ver diversos procedimentos de comparações múltiplas que focam as médias diferentes. (Não precisa perder o fôlego. Apenas apresento as fórmulas por uma questão de conhecimento.)

## Capítulo 9

# Precisando Testar Várias Médias? Venha para a ANOVA!

### *Neste Capítulo*

- ▶ Estendendo o teste- $t$  para realizar a comparação entre duas médias usando a ANOVA
- ▶ Utilizando o processo ANOVA
- ▶ Realizando um teste- $F$
- ▶ Navegando pela tabela ANOVA

**U**ma das técnicas mais utilizadas em Estatística II é a *análise de variância* (carinhosamente conhecida como ANOVA). Já que o nome tem a palavra *variância*, você poderia pensar que esta técnica tem algo a ver com variância — e tem mesmo. A análise de variância examina a quantidade de variabilidade em  $y$  (variável de resposta) e tenta entender de onde essa variabilidade vem.

Uma forma de usar a ANOVA é comparando várias populações em relação a uma variável quantitativa,  $y$ . As populações que você deseja comparar constituem diferentes grupos (representados por uma variável  $x$ ), tais como partido político, faixa etária ou diferentes marcas de um produto. A ANOVA também é empregada em situações que envolvam um experimento em que você aplica determinados tratamentos ( $x$ ) a alguns indivíduos a fim de medir a resposta ( $y$ ) obtida.

Neste capítulo, começamos com o teste- $t$  para duas médias populacionais, o precursor da ANOVA. Em seguida, passaremos para os conceitos básicos da análise de variância para a comparação de mais de duas médias: a soma de quadrados, o teste- $F$  e a tabela ANOVA. Esses princípios podem ser aplicados à análise de variância com *um fator* ou *one-way*, em que as respostas são comparadas com base apenas em uma variável de tratamento. (No Capítulo 11, você poderá ver esses princípios básicos aplicados a uma ANOVA com dois fatores, ou seja, que possui duas variáveis de tratamento.)

# Comparando Duas Médias com um Teste-*t*

O teste-*t* de duas amostras foi concebido com a finalidade de verificar se duas médias populacionais são diferentes. As condições para o teste-*t* de duas amostras são as seguintes:

- ✓ As duas populações devem ser independentes. Em outras palavras, seus resultados não se influenciam.
- ✓ A variável de resposta (*y*) é uma variável quantitativa, ou seja, seus valores têm um significado numérico e representam algum tipo de quantidade.
- ✓ Os valores de *y* para cada população têm distribuição normal. No entanto, suas médias podem ser diferentes; isso é o que o teste-*t* vai determinar.
- ✓ As variâncias das duas distribuições normais devem ser iguais.



Para tamanhos amostrais grandes com variância conhecida, use o teste-*Z* para as duas médias populacionais. No entanto, o teste-*t* lhe permite testar duas médias populacionais quando as variâncias forem desconhecidas e/ou os tamanhos amostrais forem pequenos. Isso ocorre muitas vezes em situações em que um experimento é realizado com um número limitado de indivíduos. (Consulte um livro de Estatística I ou o *Estatística Para Leigos* da Editora Alta Books para obter informações sobre o teste-*Z*.)

Embora você já tenha estudado teste-*t* nas aulas de Estatística I, pode ser bom rever os princípios básicos. O teste-*t* testa as hipóteses  $H_0: \mu_1 = \mu_2$  versus  $H_a: \mu_1 \leq, \geq, \text{ ou } \neq \mu_2$ , sendo que é a situação que determina qual destas hipóteses será usada. (**Observação:** com a ANOVA, é possível estender essa ideia a *k* diferentes médias populacionais e a única versão da  $H_a$  que interessa é  $\neq$ ).



Para realizar o teste-*t* de duas amostras, colete dois conjuntos de dados das duas populações usando duas amostras independentes. Para obter a estatística de teste (a estatística-*t*), subtraia as duas médias amostrais e divida a diferença pelo erro padrão (a combinação de dois desvios padrão das duas amostras e de seus respectivos tamanhos). Depois, compare a estatística-*t* à distribuição-*t* com  $n_1 + n_2 - 2$  graus de liberdade e encontre o valor-*p*. Se o valor-*p* for inferior ao nível  $\alpha$  predeterminado, digamos 0,05, você tem provas suficientes para dizer que as médias populacionais são diferentes. (Para mais informações sobre testes de hipótese, veja o Capítulo 3.)

Por exemplo, suponha que você esteja participando de um campeonato de cuspe de sementes de melancia, onde cada concorrente deve colocar as sementes na boca e cuspi-las o mais longe possível. Os resultados são medidos em metros e tratados com a mesma reverência dos resultados obtidos na prova de tiro realizada nos Jogos Olímpicos. Suponha que você queira comparar a distância alcançada pelas sementes cuspidas por adultos dos sexos feminino (females) e masculino (males) e, para isso, seu conjunto de dados inclui dez pessoas de cada grupo.

Veja os resultados do teste- $t$  na Figura 9-1. A distância média do cuspe para as mulheres foi de 1,21 metros e a média para os homens foi de 1,44 metros já a diferença (mulheres – homens) é 0,22 metros, ou seja, as mulheres da amostra cuspiram as sementes a distâncias mais curtas, em média, do que os homens. A estatística- $t$  para a diferença entre as duas médias (mulheres – homens) é  $t = -2,23$ , com valor- $p$  de 0,039 (veja a última linha da saída na Figura 9-1). No nível  $\alpha = 0,05$ , essa diferença é significativa (pois  $0,039 < 0,05$ ). Assim, você conclui que homens e mulheres se diferem em relação à distância média alcançada pelas sementes de melancia. Mas também poderá dizer que os homens cospem mais longe, pois a média amostral deles foi mais alta.

Two-sample T for females vs males				
	N	Mean	StDev	SE Mean
females	10	1,21	0,23	0,073
males	10	1,44	0,21	0,066
Difference = mu (females) - mu (males)				
Estimate for difference: -0.22000				
95% CI for difference: (-0.49086, -0.04914)				
T-Test of difference = 0 (vs not =): T-Value = -2.23 P-Value = 0.039 DF = 18				

---

**Figura 9-1:** Teste- $t$  comparando as distâncias médias alcançadas por mulheres versus homens na prova de cuspe de semente de melancia.

---

# Avaliando Mais Médias com ANOVA

Quando você compara muito duas populações independentes, em algum momento essas duas populações não serão mais suficientes. Suponha que você queira comparar mais de duas populações em relação a uma variável de resposta ( $y$ ). Essa ideia aperfeiçoa o teste- $t$  no território da ANOVA. O procedimento ANOVA se constrói em torno de um teste de hipótese chamado teste- $F$ , que compara o quanto os grupos diferem entre si em relação à quantidade de variabilidade dentro de cada grupo. Nesta seção, ofereço um exemplo de quando usar a ANOVA e mostro os passos envolvidos no processo de análise de variância. Você pode, então, aplicar esses passos ao exemplo a seguir ao longo de todo este capítulo.

## Cuspe de sementes: Uma situação perfeita para a ANOVA

Antes de mergulhar de cabeça na ANOVA, você deve descobrir qual a pergunta que deseja responder e coletar os dados necessários.

Suponha que você queira comparar as distâncias atingidas pelas sementes cuspidas por quatro diferentes faixas etárias: 6–8 anos, 9–11, 12–14 e 15–17. As hipóteses para esse exemplo são  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$  versus  $H_a$ : pelo menos duas dessas médias são diferentes, onde a média populacional  $\mu$  representa as médias das faixas etárias, respectivamente. Ao longo dos anos desta competição, você coletou dados sobre 200 crianças de cada faixa etária, por isso, tem uma ideia de quais serão as distâncias alcançadas. Este ano são 20 participantes, cinco em cada faixa etária. Veja os dados deste ano, em metros, na Tabela 9-1.

Tabela 9-1 Distâncias na prova de cuspe de sementes de melancia para quatro faixas etárias de crianças (medidas em metros)				
6–8 anos	9–11 anos	12–14 anos	15–17 anos	
0,97	0,97	1,12	1,12	
0,99	0,99	1,09	1,19	
1,07	1,02	1,02	1,14	
1,02	1,12	1,12	1,14	
1,04	1,09	1,14	1,17	

Você vê alguma diferença entre as distâncias alcançadas por essas faixas etárias com base apenas nesses dados? Se você combinasse todos os dados, veria uma pequena diferença (o intervalo da combinação dos dados vai de 0,97 metros até 1,19 metros) e poderia suspeitar que as crianças mais velhas conseguem cuspir mais longe.

Talvez a faixa etária a que cada competidor pertence explique pelo menos parte do que acontece. Mas você não deve parar aqui. A próxima seção mostra os passos que precisam

ser dados a fim de encontrar uma resposta à sua pergunta.

## *Seguindo os passos da ANOVA*

Depois de ter decidido a variável de resposta quantitativa ( $y$ ) que deseja comparar para  $k$  diversas médias populacionais (ou tratamento) e coletado uma amostra aleatória de dados de cada população (consulte a seção anterior), você está pronto para realizar a análise de variância nos dados e ver se as médias da população são diferentes para a variável de resposta,  $y$ .

A característica que distingue essas populações é chamada *variável de tratamento*,  $x$ . Os estatísticos usam a palavra *tratamento* neste contexto, pois um dos maiores usos da ANOVA é feito em ensaios clínicos, onde os indivíduos são selecionados aleatoriamente para participar de tratamentos e as reações são comparadas entre os grupos de diferentes tratamentos. Assim, os estatísticos usam a palavra tratamento mesmo quando o estudo não é um ensaio clínico e populações comuns estão sendo comparadas. Ei, não me culpe! Estou apenas seguindo a terminologia estatística apropriada.

Veja aqui os passos para uma ANOVA com um fator:

- 1. Verifique as condições de análise de variância utilizando os dados coletados de cada uma das  $k$  populações.**

Consulte a próxima seção, "Verificando as condições", para saber que condições são essas.

- 2. Estabeleça as hipóteses  $H_0: \mu_1 = \mu_2 \dots = \mu_k$  versus  $H_a$ : pelo menos duas das médias populacionais são diferentes.**

Outra maneira de expressar a sua hipótese alternativa é dizendo  $H_a$ : pelo menos duas de  $\mu_1, \mu_2, \dots, \mu_k$  são diferentes.

- 3. Colete dados de  $k$  amostras aleatórias, uma de cada população.**
- 4. Realize um teste-F para os dados do terceiro passo, utilizando as hipóteses do segundo passo, e encontre o valor- $p$ .**

Consulte a seção "Realizando um Teste- $F$ ", mais adiante neste capítulo, para obter essas instruções.

- 5. Tire suas conclusões: se rejeitar  $H_0$  (quando o valor- $p$  for inferior a 0,05 ou o nível  $\alpha$  predeterminado), conclui-se que pelo menos duas das médias populacionais são diferentes, caso contrário, conclui-se que não temos evidências suficientes para rejeitar  $H_0$  (não se pode dizer que as médias são diferentes).**

Se parecer que estou falando grego, não se preocupe — vou descrever cada um desses passos em detalhes nas seções seguintes.



# *Verificando as Condições*

A primeira etapa da ANOVA é verificar se todas as condições necessárias foram atendidas antes de mergulhar de cabeça na análise dos dados. As condições de utilização ANOVA são apenas uma extensão das condições para um teste-t. (Veja a seção "Comparando duas médias com um teste-t".) Todas as condições a seguir precisam ser satisfeitas a fim de que a análise de variância possa ser realizada.

- ✓ As  $k$  populações devem ser independentes. Em outras palavras, seus resultados não se influenciam.
- ✓ As  $k$  populações devem ter distribuição normal.
- ✓ As variâncias das  $k$  distribuições normais devem ser iguais.

## *Verificando a independência*

Para verificar a primeira condição, examine a forma como os dados foram coletados de cada uma das populações. A fim de manter a independência, os resultados de uma população não podem afetar os resultados das outras populações. Se os dados foram coletados através de uma amostra aleatória independente de cada população (aleatório aqui significa que cada indivíduo da população teve uma oportunidade igual de ser selecionado), esse fator garante essa independência em seu nível mais forte.

Nos casos dos dados da prova de cuspe de sementes (ver Tabela 9-1), eles não foram amostrados de forma aleatória, pois representam todos os que participaram da competição. Mas você pode argumentar que, na maioria dos casos, as distâncias alcançadas por uma faixa etária não afetam as distâncias atingidas por outras faixas etárias, de modo que a suposição da independência é relativamente boa.

## *Procurando o que é normal*

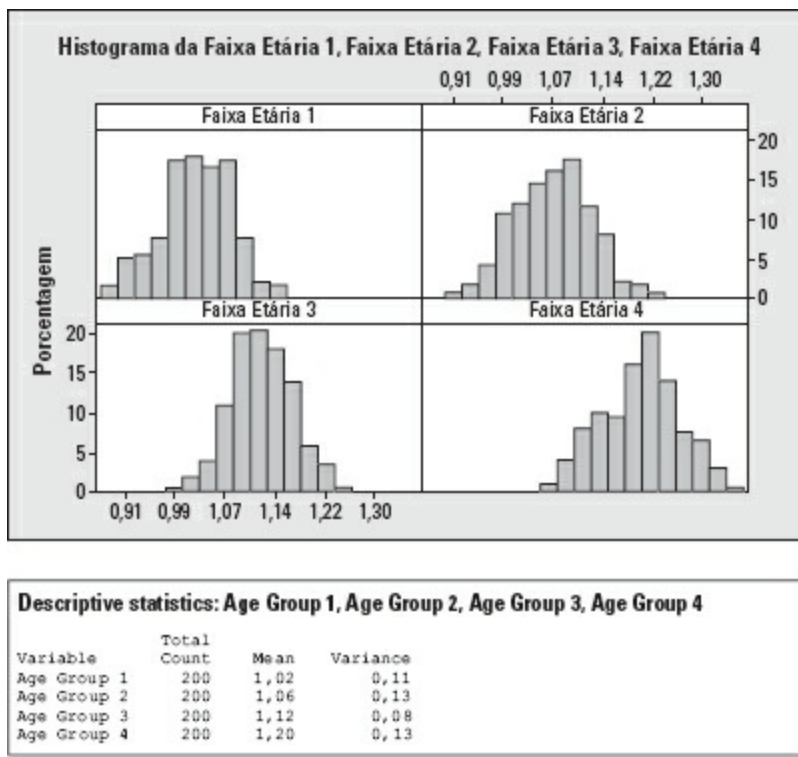
A segunda condição para a realização da ANOVA é que cada uma das  $k$  populações tenha uma distribuição normal. Para verificar esta condição, faça um histograma individual dos dados de cada grupo e veja se ele se assemelha a uma distribuição normal. Os dados de uma distribuição normal devem ser simétricos (isto é, se você dividir o histograma ao meio, cada um dos lados será idêntico) e ter a forma de um sino. Não espere que os dados em cada histograma sigam uma distribuição normal exata (lembre-se de que isso é apenas uma amostra), mas também não devem ser muito diferentes de uma distribuição normal, ou seja, em forma de sino.

Como os dados do exemplo em questão contêm apenas cinco crianças por faixa etária, a verificação das condições pode ser duvidosa. Mas, neste caso, você tem dados de anos anteriores para 200 crianças em cada faixa etária e, sendo assim, poderá usá-los para verificar as condições. Os histogramas e as estatísticas descritivas dos dados dessa

competição para as quatro faixas etárias estão na Figura 9-2, tudo em um só painel, para que você possa compará-los uns aos outros na mesma escala.

Olhando os quatro histogramas na Figura 9-2, você pode ver que cada gráfico se assemelha à forma de um sino, portanto, a condição de normalidade não está sendo gravemente violada aqui. (O botão vermelho deve ser acionado se você notar dois picos nos dados, uma forma assimétrica onde o pico esteja longe de um lado ou um histograma plano, por exemplo.)

Você pode usar o Minitab para construir histogramas para cada uma das amostras e fazer com que todos eles apareçam em um grande painel, todos usando a mesma escala. Para fazer isso, clique em Graph> Histogram e clique em OK. Escolha as variáveis que representam os dados de cada amostra, selecionando-as no campo à esquerda e clicando em Select. Em seguida, clique em Multiple Graphs e uma nova janela será aberta. Abaixo da opção Show Graph Variables, marque o seguinte campo: em painéis separados do mesmo gráfico. Na opção Same Scales for Graphs, marque os campos para x e y. Essa opção lhe dá a mesma escala tanto para o eixo de  $x$  quanto para o eixo de  $y$  em todos os histogramas. Depois, clique em OK.



**Figura 9-2:** Verificando as condições necessárias para a realização da ANOVA através de histogramas e de estatísticas descritivas.

## Notando a dispersão

A terceira condição para a realização da ANOVA é que a variância em cada uma das  $k$  populações seja igual; os estatísticos a chamam de *condição de igualdade de variâncias*. Há duas maneiras de verificar essa condição em seus dados:

- ✓ Calcule cada uma das variâncias de cada amostra e compare-as.
- ✓ Construa um gráfico que mostre lado a lado todos os boxplots de cada amostra. Este tipo de gráfico é chamado de *boxplot paralelo*. (Consulte um livro de Estatística I ou o *Estatística Para Leigos* da Alta Books para obter informações sobre boxplots.)

Se uma ou mais das variâncias calculadas forem significativamente diferentes das demais, a condição de igualdade de variância provavelmente não vai ser atendida. O que quer dizer *significativamente diferente*? Um teste de hipótese para igualdade de variâncias é a ferramenta estatística utilizada para resolver essa questão; no entanto, está fora do escopo da maioria dos cursos de Estatística II, por isso, por ora, você pode tirar uma conclusão subjetiva. Sempre digo que, se as diferenças entre as variâncias calculadas forem suficientes para você (digamos que se diferenciem em 10% ou mais), é provável que a condição de igualdade de variância não seja respeitada.

Da mesma forma, se o comprimento de um ou mais boxplots paralelos parecerem diferentes o suficiente para você, é provável que a condição de igualdade de variância não seja cumprida (mas, escute, se você realmente se preocupa com qualquer questão estatística, talvez deva apimentar um pouco mais sua vida).

O comprimento da caixa de um boxplot é chamado de *intervalo interquartil*. Para calculá-lo, subtraia o terceiro quartil (percentil 75) do primeiro quartil (percentil 25). (Consulte um livro de Estatística I ou o *Estatística Para Leigos* para mais informações.)

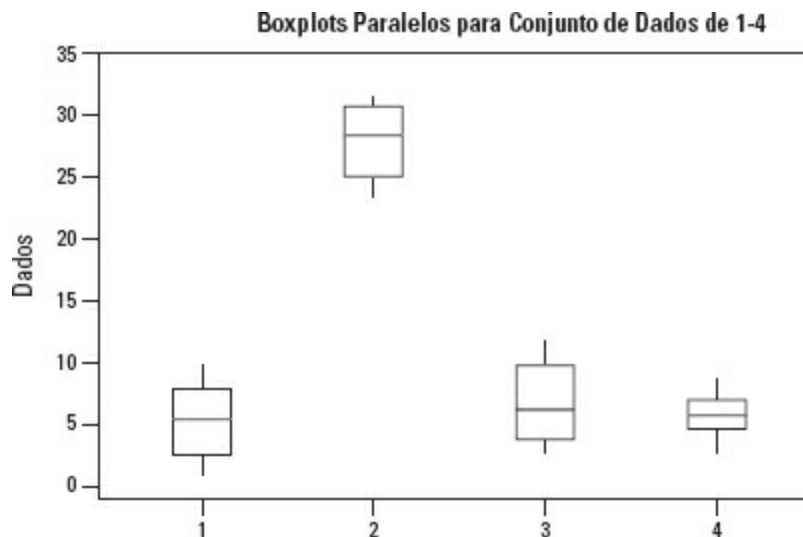
A Tabela 9-2 mostra um exemplo de quatro pequenos conjuntos de dados com suas respectivas variâncias calculadas e demonstradas na última linha. Note que a variância do conjunto de dados 4 é significativamente menor do que a dos demais. Neste caso, é seguro dizer que a condição de igualdade de variância não foi atendida.

**Tabela 9-2 Comparando variâncias de quatro conjuntos de dados para verificar a condição de igualdade de variância**

<i>Conjunto de dados 1</i>	<i>Conjunto de dados 2</i>	<i>Conjunto de dados 3</i>	<i>Conjunto de dados 4</i>
1	32	4	3
2	24	3	4
3	27	5	5
4	32	10	5
5	31	7	6
6	28	4	6
7	30	8	7
8	26	12	7
9	31	9	8

10	24	10	9
Variância = 9,167	Variância = 9,833	Variância = 9,511	Variância = 3,333

A Figura 9-3 mostra os boxplots paralelos para esses mesmos quatro conjuntos de dados. É possível notar que o boxplot para o conjunto de dados 4 tem um intervalo interquartil (comprimento da caixa) significativamente menor do que os demais. Calculei os intervalos interquartis reais para estes quatro conjuntos de dados, são eles: 5,50; 5,75; 6,00 e 2,50, respectivamente. Esses achados confirmam a conclusão de que a condição de igualdade de variância não foi atendida, pois a variabilidade do grupo 4 é muito menor.



**Figura 9-3:** Analisando boxplots paralelos para verificar a condição de igualdade de variância.

Para encontrar a estatística descritiva (incluindo a variância e o intervalo interquartil) para cada amostra, clique em Stat > Basic Statistics > Display Descriptive Statistics. Clique em cada variável no campo à esquerda para as quais deseja a estatística descritiva e clique em Select. Ao clicar na opção Statistics, você vai abrir uma janela com vários tipos diferentes de estatísticas. Marque as que desejar e desmarque as que não quiser. Clique em OK. Depois, clique em OK de novo. Pronto! Suas estatísticas descritivas estão calculadas.

Para encontrar boxplots paralelos no Minitab, clique na opção Graph > Boxplot. Uma janela vai aparecer. Clique sobre a imagem para Multiple Y's, Simple e, em seguida, clique em OK. Selecione, à esquerda, as variáveis que deseja comparar e clique em Select. Depois, clique em OK.

Observe que não é preciso que os tamanhos amostrais de cada grupo sejam iguais para realizar uma análise de variância. Porém, em Estatística II, costuma-se ver o que os estatísticos chamam de um delineamento balanceado, onde cada amostra de cada população tem o mesmo tamanho. (Como explico no Capítulo 3, para obter maior precisão em seus dados, quanto maior o tamanho da amostra, melhor.)

Para os dados do exemplo da competição de cuspe de sementes, as variâncias para cada faixa etária estão listadas na Figura 9-2. Essas variâncias estão próximas o bastante para que possamos dizer que a condição de igualdade de variância foi atendida.

# ***Estabelecendo as Hipóteses***

O segundo passo da ANOVA é a formulação das hipóteses a serem testadas. O objetivo de seu teste é ver se todas as médias populacionais podem ser consideradas iguais. A hipótese nula para a ANOVA é que todas as médias populacionais são iguais. Ou seja,  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ , onde  $\mu_1$  é a média da primeira população,  $\mu_2$  é a média da segunda população e assim por diante, até chegar a  $\mu_k$  (a média da  $k$ -ésima população).



O que aparece na hipótese alternativa ( $H_a$ ) deve ser o oposto do que aparece na hipótese nula ( $H_0$ ). Qual é o oposto de ter todas as médias das  $k$  populações iguais umas às outras? Talvez você possa pensar que o contrário é que elas sejam todas diferentes, mas esse não é o caso. Para mandar a  $H_0$  para o espaço, tudo o que você precisa é que pelo menos duas dessas médias não sejam iguais. Assim, a hipótese alternativa,  $H_a$ , é que pelo menos duas das médias das populações são diferentes. Ou seja,  $H_a$ : pelo menos duas das  $\mu_1, \mu_2, \dots, \mu_k$  são diferentes.

Note que  $H_0$  e  $H_a$  para a ANOVA são uma extensão das hipóteses para um teste- $t$  de duas amostras (que só compara duas populações independentes). E, mesmo que a hipótese alternativa em um teste- $t$  seja a de que uma média é maior do que, menor do que ou diferente das outras na ANOVA, não considere outra alternativa senão  $\neq$ . (Os estatísticos usam modelos mais complexos para as demais hipóteses. Você não fica feliz em saber que é outra pessoa que faz isso?)



Nesta fase do jogo, o importante é saber se as médias são iguais ou não. Depois de chegar à conclusão de que  $H_0$  é rejeitada, pode proceder para descobrir como as médias se diferem, quais são as maiores e assim por diante, usando as comparações múltiplas. Esses detalhes estão no Capítulo 10.

# Realizando o Teste-F

O terceiro passo na realização da ANOVA é coletar os dados, o que inclui a coleta de  $k$  amostras aleatórias, uma de cada população. O quarto passo é fazer o teste-F para esses dados, que é o coração do processo de análise de variância. Este é o real teste de hipótese de  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$  versus  $H_a$ : pelo menos, duas de  $\mu_1, \mu_2, \dots, \mu_k$  são diferentes.

Três são as principais etapas a serem realizadas a fim de completar o teste-F. Observação: não confunda estas etapas com as principais etapas para a realização da ANOVA; considere o teste-F como os passos de uma etapa:

1. **Desmembre a variância de  $y$  em somas de quadrados.**
2. **Encontre as médias para essas somas.**
3. **Junte as médias das somas dos quadrados para obter a estatística-F.**

Nas seções a seguir, descrevo cada etapa do teste-F em pormenor e aplico-as ao exemplo que compara as distâncias atingidas na competição de cuspe de sementes de melancia (ver Tabela 9-1).



Os analistas de dados contam com os softwares computacionais para realizar cada etapa do teste-F e você pode fazer o mesmo. Todos os pacotes de software organizam e resumem as informações importantes obtidas a partir do teste-F em uma tabela.

Este quadro de resultados para a ANOVA se chama... (Adivinha?) Tabela ANOVA. Como a tabela ANOVA é uma parte crítica de todo o processo da ANOVA, começo as seções seguintes descrevendo como executar a ANOVA no Minitab para obter sua tabela e continuo me referindo a esta seção à medida que descrever cada etapa do processo de análise de variância.

## ANOVA no Minitab



Para executar uma ANOVA no Minitab, primeiro insira os dados das  $k$  amostras. Você pode inserir os dados de duas maneiras:

- ✓ **Stacked data:** Insira todos os dados em duas colunas, já as respostas ( $y$ ) ficam na coluna 2. Para analisar esses dados, clique na opção Stat>ANOVA>One-Way Stacked. Selecione a variável de resposta ( $y$ ) e clique em Select. Em seguida, selecione a variável fator (população) e clique em Select. Clique em OK.
- ✓ **Unstacked data:** Insira os dados de cada amostra em colunas individuais: para analisar esses dados, clique na opção Stat>ANOVA>One-Way Unstacked. Selecione os nomes das colunas onde os dados estão localizados e clique em OK.

Normalmente, uso a versão unstacked para inserir os dados só porque acho que ajuda a visualizá-los. No entanto, a escolha é sua, e o resultado será o mesmo, independentemente

do método escolhido, contanto que você seja consistente.

## ***Desmembrando a variância em somas de quadrados***

O primeiro passo para realizar um teste- $F$  é dividir a variabilidade de  $y$  em porções e definir de onde ela está vindo. Cada porção da variabilidade recebe o nome de soma de quadrados. O termo análise de variância é a descrição exata de como realizar um teste para  $k$  médias populacionais. Com o objetivo principal de verificar se  $k$  médias populacionais (ou tratamentos) são iguais, colete uma amostra aleatória de cada uma das  $k$  populações.

Primeiro, coloque todos os dados juntos em um grande grupo e meça a quantidade de variabilidade total presente; esta variabilidade é a chamada soma total de quadrados, ou SQTO. Se os dados realmente forem diversificados, a SQTO será grande. Se os dados forem semelhantes, a SQTO será pequena.

É possível decompor a variabilidade total de um conjunto de dados combinados (SQTO) em duas partes:

- ✓ **SQT:** A variabilidade entre os grupos, conhecida como *soma de quadrados entre os grupos* (ou variabilidade entre tratamentos).
- ✓ **SQE:** A variabilidade dentro dos grupos, conhecida como *soma de quadrados dentro dos grupos* (ou variabilidade dentro dos tratamentos, ou seja, representa a variação devido ao erro experimental).

Assim, você está decompondo a variabilidade dos resultados de seus dados em uma das igualdades mais importantes na ANOVA:

$$\text{SQTO} = \text{SQT} + \text{SQE}$$

A fórmula para SQTO é o numerador da fórmula para  $s^2$ , a variância de um único conjunto de dados, então  $\text{SQTO} = \sum_i n_i (\bar{x}_i - \bar{x})^2$ , onde  $i$  e  $j$  representam o  $j$ -ésimo valor na amostra retirada da  $i$ -ésima população e é a *média amostral global* (média de todo o conjunto de dados). Assim, tratandose de análise de variância, SQTO é a distância total ao quadrado entre os valores dos dados e sua média global.

A fórmula para a SQT é  $\sum_i n_i (\bar{x}_i - \bar{x})^2$ , onde  $n_i$  é o tamanho da amostra proveniente da  $i$ -ésima população e é a média amostral global. SQT representa a distância total ao quadrado entre as médias de cada amostra e a média global.

A fórmula para SQE é  $\sum_i \sum_j (x_{ij} - \bar{x}_i)^2$ , onde  $x_{ij}$  é o  $j$ -ésimo valor da amostra retirada da  $i$ -ésima população e é a média amostral proveniente da  $i$ -ésima população. Essa fórmula representa a distância total ao quadrado entre os valores de cada amostra e suas médias amostrais correspondentes. Através da álgebra é possível confirmar (depois de ter os cotovelos seriamente esfolados) que  $\text{SQTO} = \text{SQT} + \text{SQE}$ .





A saída produzida pelo Minitab para os dados do concurso de cuspe de sementes de melancia envolvendo quatro faixas etárias é mostrada na Figura 9-4. Embaixo da coluna Source da tabela ANOVA, vemos Factor na linha um. A variável fator (como descrito pelo Minitab) representa o tratamento ou a variável população. Na coluna três da linha Factor você vê o SQT, equivalente a 89,75. Na linha Error (linha dois), vemos o SQD na coluna três, equivalente a 56,80. Na coluna três da linha Total (linha três), vemos STQO igual a 146,55. Utilizando os valores de SQT, SQE e SQTO, obtidos na saída do Minitab, podemos verificar que  $SQT + SQE = SQTO$ .

One-Way ANOVA: Age Group 1, Age Group 2, Age Group 3, Age Group 4					
Source	DF	SS	MS	F	P
Factor	3	89.75	29.92	8.43	0.001
Error	16	56.80	3.55		
Total	19	146.55			

S = 1.884 R-Sq = 61.24% R-Sq(adj) = 53.97%

**Figura 9-4:** Saída do Minitab para ANOVA do exemplo referente à competição de cuspe de semente de melancia.

Agora você está pronto para usar essas somas de quadrado para completar o próximo passo na realização do teste- $F$ .

## Localizando as médias das somas de quadrados

Depois de ter a soma de quadrados entre os grupos, SQT, e as somas de quadrados dentro dos grupos, SQE (consulte a seção anterior para saber mais sobre esse assunto), você deseja compará-las para ver se a variabilidade dos valores de  $y$  causada pelo modelo (SQT) é grande quando comparada à quantidade de resíduo deixado nos dados depois que os grupos foram contabilizados (SQE). Ou seja, você quer uma razão que, de algum modo, compare SQT a SQE.

Para fazer com que essa razão gere uma estatística com a qual eles saibam lidar (neste caso, uma estatística- $F$ ), os estatísticos decidiram encontrar as médias de SQT e SQE e trabalhar com elas. Encontrar as médias das somas de quadrados é o segundo passo para a realização do teste- $F$ , e as médias das somas são as seguintes:

- ✓ **MQT** é a *média das somas de quadrados para os tratamentos*, que mede a variação média ocorrida entre os diferentes tratamentos (as diferentes amostras que compõem o conjunto de dados). O que estamos procurando é a quantidade de variação presente nos dados à medida que passamos de uma amostra para outra. Uma grande variabilidade entre as amostras (tratamentos) pode indicar que as populações também são diferentes. Para encontrar a MQT, divida a SQT por  $k - 1$  (onde  $k$  é o número de tratamentos).
- ✓ **MQE** é a *média das somas de quadrados dentro dos grupos*, que mede a variação média dentro do tratamento. A *variabilidade dentro do tratamento* é a quantidade

de variabilidade observada dentro de cada amostra em si, devido ao acaso e/ou outros fatores não incluídos no modelo.

Para encontrar a MQE, divida a SQE por  $n - k$  (onde  $n$  é o tamanho total da amostra e  $k$  é o número de tratamentos). Os valores de  $k - 1$  e  $n - k$  são chamados de *graus de liberdade* (ou gl) para SQT e SQE, respectivamente.

O Minitab calcula e mostra os graus de liberdade para SQE, SQT, MQT e MQE na tabela ANOVA nas colunas dois e quatro, respectivamente.

A partir da tabela ANOVA para os dados da competição de cuspe de sementes ilustrada na Figura 9-4, é possível notar que a coluna dois se chama DF, sigla para *degrees of freedom*, ou, em português, graus de liberdade. Os graus de liberdade para SQT estão na linha Factor (linha dois); seu valor é igual a  $k - 1 = 4 - 1 = 3$ . O valor dos graus de liberdade para SQE é  $n - k = 20 - 4 = 16$ . (Lembre-se: são quatro faixas etárias com cinco crianças em cada, formando um total de  $n = 20$  valores de dados). Os graus de liberdade para SQTO são  $n - 1 = 20 - 1 = 19$  (encontrado na linha Total embaixo da coluna DF). Você pode ver que os graus de liberdade para SQTO = os graus de liberdade de SQT + graus de liberdade de SQE.

Os valores do MQE e MQT são mostrados na coluna quatro da Figura 9-4, cujo título é MS. O valor para MQE está na linha Factor, 29,92. Esse valor foi obtido pela divisão de SQT = 89,75 e por seus graus de liberdade, 3. O valor de MQE está na linha Error e é igual a 3,55. MQE foi encontrado através da divisão de SQE = 56,80 por seus graus de liberdade, 16.

Ao encontrar as médias das somas de quadrados, você conclui o segundo passo do teste- $F$ , mas não pare por aqui! Para concluir o processo é preciso chegar até a próxima seção.

## ***Chegando à estatística- $F$***

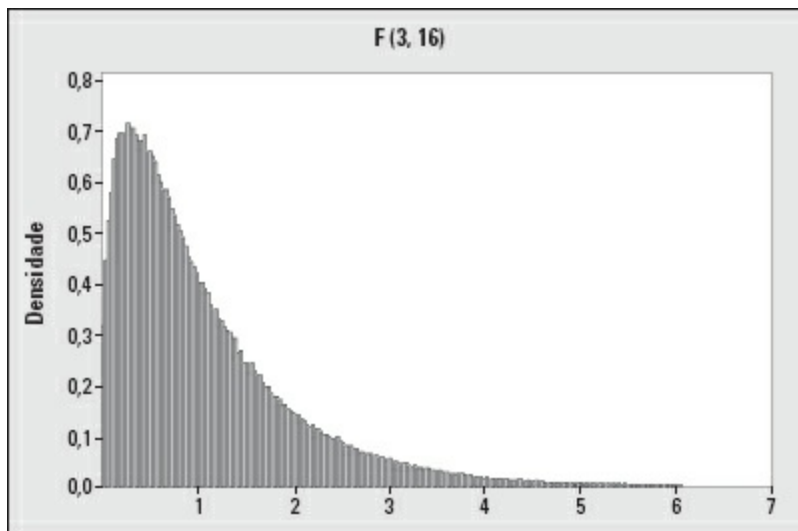
A estatística para o teste de igualdade das  $k$  médias populacionais é  $F = \frac{MQT}{MQE}$ . O resultado desta fórmula recebe o nome de estatística- $F$ . A estatística- $F$  tem uma distribuição- $F$ , equivalente ao quadrado de um teste- $t$  (quando o grau de liberdade do numerador é 1; veja mais sobre essa interessante relação entre as distribuições- $t$  e  $F$  no Capítulo 12). Todas as distribuições  $F$  começam em zero e se deslocam para a direita. O grau de curvatura e a altura de cada distribuição- $F$  se refletem em dois graus de liberdade, representados por  $k - 1$  e  $n - k$ . (Estes vêm dos denominadores de MQT e MQE, respectivamente, onde  $n$  é o tamanho total da amostra e  $k$  é o número total de tratamentos ou populações.) Uma forma abreviada de denotar a distribuição- $F$  nesse teste é  $F_{(k - 1, n - k)}$ .

No exemplo da competição de cuspe de sementes de melancia, estamos comparando quatro médias e temos um tamanho amostral igual a cinco para cada população. A Figura 9-5 mostra a distribuição correspondente, que possui graus de liberdade  $4 - 1 = 3$  e  $20 - 4 =$



16, isto é,  $F_{(3, 16)}$ .

A estatística- $F$  aparece na saída ANOVA do Minitab (ver Figura 9-4) na linha Factor, sob a coluna indicada por  $F$ . Para o exemplo da prova de cuspe de sementes, o valor da estatística- $F$  é 8,43, obtido pela divisão de  $MQT = 29,92$  por  $MQE = 3,55$ . Em seguida, localize 8,43 na distribuição- $F$  na Figura 9-5 para ver sua posição, em termos de valor- $p$ . (E parece que ele está beeeem longe dali. Veja mais sobre esse assunto na próxima seção.)



**Figura 9-5:** Distribuição- $F$  com (3,16) graus de liberdade.

Cuidado para não trocar a ordem dos graus de liberdade para a distribuição- $F$ . A diferença entre  $F_{(3, 16)}$  e  $F_{(16, 3)}$  é bem grande.

## ***Tirando conclusões a partir da ANOVA***

Se tiver concluído o teste- $F$  e encontrado a estatística- $F$  (o quarto passo do processo de análise de variância), você está pronto para o quinto passo: tirar conclusões para seu teste de hipótese para as  $k$  médias populacionais. Porém, se ainda não tiver feito isso, você pode comparar a estatística- $F$  à distribuição- $F$  correspondente a  $(k - 1, n - k)$  graus de liberdade para ver onde ela se encontra e tirar suas próprias conclusões. Você pode tirar suas conclusões de duas maneiras: seguindo a abordagem do valor- $p$  ou a abordagem do valor crítico. A abordagem a ser usada depende principalmente da possibilidade de se ter acesso a um computador, especialmente no caso de uma prova. Descrevo essas duas abordagens nas seções seguintes.

### ***Usando a abordagem do valor- $p$***

Na saída ANOVA do Minitab (ver Figura 9-4), o valor da estatística- $F$  está localizado na linha Factor, sob a coluna indicada por  $F$ . O valor- $p$  para o teste- $F$  está na linha Factor embaixo da coluna indicada por  $P$ , e é ele que indica se você pode ou não rejeitar  $H_0$ .

- ✓ Se o valor- $p$  for menor do que o nível predeterminado para  $\alpha$ , normalmente 0,05, rejeite  $H_0$ . Conclua que as  $k$  médias populacionais não são todas iguais e que pelo menos duas delas são diferentes.
- ✓ Se o valor- $p$  for maior do que  $\alpha$ , então você não pode rejeitar  $H_0$ . Não há provas suficientes nos dados para dizer que as  $k$  médias populacionais têm alguma diferença.

A estatística- $F$  para a comparação entre as distâncias médias alcançadas na prova de cuspe de sementes de melancia pelas quatro faixas etárias é 8,43. O valor- $p$  indicado na Figura 9-4 é 0,001, ou seja, os resultados são estatisticamente muito significativos. Sendo assim, rejeite  $H_0$  e conclua que pelo menos um par de faixas etárias se difere com relação às distâncias médias alcançadas. (Você esperaria que um jovem de 17 anos pudesse se dar muito melhor na competição do que uma criança de 6 anos, mas talvez a criança tenha muito mais prática do que o jovem.)

Na Figura 9-5, você vê como a estatística- $F$  8,43 se posiciona na distribuição- $F$  com  $(4 - 1, 20 - 4) = (3, 16)$  graus de liberdade. Podemos perceber que ela está muito para a direita, fora de nosso campo de visão. Portanto, faz sentido que o valor- $p$ , que mede a probabilidade de estar além dessa estatística- $F$ , seja 0,001.

### *Usando os valores críticos*

Se você estiver em uma situação em que não tem acesso a um computador (como é o caso em muitos cursos de Estatística quando se fala em provas), não é possível encontrar o valor- $p$  exato para a estatística- $F$  utilizando uma tabela. Você simplesmente deve escolher o valor- $p$  da estatística- $F$  que estiver mais próximo do seu. No entanto, se você tiver acesso a um computador quando estiver fazendo uma lição de casa ou uma prova, o programa calculará automaticamente todos os valores- $p$  de forma exata e, portanto, você vai encontrá-los em qualquer saída computacional.

Para aproximar o valor- $p$  à sua estatística- $F$ , caso você não tenha um computador ou um resultado obtido por computador à sua disposição, encontre um valor de corte na distribuição- $F$  com  $(k - 1, n - k)$  graus de liberdade que delimite a rejeição e a não rejeição da  $H_0$ . Esse valor de corte, também conhecido como *valor crítico*, é determinado pelo  $\alpha$  predeterminado (normalmente, 0,05). O valor crítico deve ser escolhido de modo a fazer com que a área à sua direita sobre a distribuição- $F$  seja igual a  $\alpha$ .

Vários livros de Estatística e sites na web disponibilizam tabelas de distribuição- $F$  para outros valores de  $\alpha$ ; porém,  $\alpha = 0,05$  é de longe o nível  $\alpha$  mais usado para a distribuição- $F$ , além de ser suficiente para seus propósitos.

A tabela de valores para a distribuição- $F$  é chamada de *tabela- $F$*  e os alunos geralmente recebem-na com a prova. Para o exemplo da competição de cuspe de sementes, a estatística- $F$  tem uma distribuição- $F$  com graus de liberdade  $(3, 16)$ , onde  $3 = k - 1$  e  $16 = n - k$ . Para encontrar o valor crítico, consulte uma tabela- $F$  (Tabela A-5 no apêndice).



Procure os graus de liberdade (3,16) e verá que o valor crítico é 3,2389 (ou 3,24). Sua estatística- $F$  para o exemplo em questão é igual a 8,43, que está bem além deste valor crítico (se observar a Figura 9-5, poderá ver como 8,43 se compara a 3,24), sua conclusão é a de rejeitar  $H_0$  ao nível  $\alpha$ . Pelo menos, duas das faixas etárias se diferem em relação às distâncias médias alcançadas.



Com a abordagem do valor crítico, qualquer estatística- $F$  situada além do valor crítico resulta na rejeição da  $H_0$ , independentemente da distância em que se encontra. Se sua estatística- $F$  estiver além do valor encontrado na tabela- $F$ , então você deve rejeitar  $H_0$  e dizer que pelo menos dois dos tratamentos (ou populações) têm médias diferentes.

## ***O que fazer agora?***

Depois de ter rejeitado  $H_0$  no teste- $F$  e concluído que nem todas as médias populacionais são iguais, a sua próxima pergunta pode ser: “Quais são diferentes?” A resposta a essa pergunta pode ser encontrada com uma técnica estatística chamada *comparações múltiplas*. Os estatísticos usam diversos procedimentos de comparações múltiplas para investigar a fundo as médias quando o teste- $F$  é rejeitado. No Capítulo 10, discuto e aplico algumas das técnicas mais comuns de comparação múltipla.

# Verificando o Ajuste do Modelo ANOVA

Como acontece com qualquer outro modelo, você deve determinar a qualidade do ajuste do modelo ANOVA antes de utilizar seus resultados com confiança. No caso da análise de variância, o modelo basicamente se resume a uma variável tratamento (também conhecida como população de interesse), mais um termo de erro. Para avaliar o quanto esse modelo se ajusta aos dados, consulte os valores de  $R^2$  e o  $R^2$  ajustado na última linha da saída ANOVA abaixo da tabela ANOVA. Para os dados da prova de cuspe de sementes esses valores estão na parte inferior da Figura 9-4.

- ✓ **O valor de  $R^2$  mede a porcentagem de variabilidade na variável de resposta (y) explicada pela variável explicativa (x).** No caso da ANOVA, a variável x é o fator em virtude do tratamento (onde o tratamento pode representar a população que está sendo comparada). Um valor alto para  $R^2$  (digamos, acima de 80%) indica um bom ajuste do modelo.
- ✓ **O  $R^2$  ajustado, a medida preferida, ajusta o valor de  $R^2$  de acordo com o número de variáveis presentes no modelo.** No caso da ANOVA com um fator, tem-se apenas uma variável, o fator em virtude do tratamento, portanto,  $R^2$  e  $R^2$  ajustado não serão muito diferentes. (Veja no Capítulo 6 mais informações sobre  $R^2$  e  $R^2$  ajustado.)

Para os dados do exemplo em questão, o valor de  $R^2$  ajustado (encontrado na última linha da Figura 9-4) é apenas de 53,97%. Isso significa que a faixa etária (que se mostrou estatisticamente significativa no teste- $F$ ; consulte a seção "Tirando conclusões a partir da ANOVA") explica apenas um pouco mais do que a metade da variabilidade nas distâncias alcançadas na competição de cuspe de melancia. Por causa dessa conexão, você pode encontrar outras variáveis que possam ser examinadas, além de faixa etária, a fim de estabelecer um modelo ainda melhor para a previsão das distâncias atingidas pelas sementes.

Como você pode ver na Figura 9-1, os resultados do teste-t realizado para comparar as distâncias atingidas por homens e mulheres na seção "Comparando duas médias com um teste-t" mostram que homens e mulheres obtiveram médias significativamente diferentes ( $p = 0,039 < 0,05$ ). Então, eu arriscaria supor que se incluíssemos sexo, além da faixa etária, criando o que os estatísticos chamam de ANOVA com dois fatores (ou *two-way ANOVA*), o modelo resultante se ajustaria ainda melhor aos dados, resultando em valores mais altos para  $R^2$  e  $R^2$  ajustado. (O Capítulo 11 o guia pela ANOVA com dois fatores.)

# de recusa

Muitos estudos médicos e psicológicos utilizam experimentos planejados para comparar as respostas de vários tratamentos diferentes, sempre procurando as diferenças. O *experimento planejado* é um estudo em que se selecionam aleatoriamente indivíduos para diferentes tratamentos (condições experimentais) e se registram suas respostas. Os resultados são, então, usados para comparar os tratamentos a fim de saber quais são os melhores, quais funcionam igualmente bem e assim por diante.

Os pesquisadores da Ohio State University conduziram um experimento usando a ANOVA para determinar a maneira mais eficaz de escrever uma carta de recusa. (E, por acaso, existe realmente uma maneira melhor de dizer “não” a alguém? Acontece que a resposta é “sim”.) O experimento testou três formatos tradicionais para se escrever uma carta de recusa:

- ✓ Usando uma evasiva, uma frase neutra ou positiva que retarda a informação negativa.
- ✓ Colocar a razão antes da recusa.
- ✓ Finalizar a carta com uma nota positiva, como forma de minimizar a situação.

Os participantes foram distribuídos aleatoriamente aos tratamentos e suas respostas às cartas de rejeição foram comparadas (provavelmente, com algum tipo de escala, como de 1 = muito negativo a 7 = muito positivo, com 4 sendo uma resposta neutra).

Você pode analisar essa situação utilizando ANOVA, pois está comparando três tratamentos (formatos de cartas de rejeição) em relação a uma variável quantitativa (resposta à carta). Você pode dizer que a resposta a uma carta não é uma variável contínua, no entanto, ela possui valores suficientes para provar que a ANOVA é razoável. Os dados também demonstraram ter a forma de um sino.

A hipótese nula seria  $H_0$ : as respostas médias aos três tipos de cartas de recusa são iguais versus  $H_a$ : pelo menos dois formatos de carta de recusa resultaram em diferentes respostas médias.

No final, os pesquisadores realmente encontraram alguns resultados significativos. Os diferentes formatos em que a carta de recusa foi escrita contagiaram os participantes de forma diferente (por isso, o teste-F foi rejeitado). Usando os procedimentos de comparações múltiplas (veja o Capítulo 10), é possível prosseguir e determinar quais os formatos de cartas provocaram respostas diferentes e como elas se diferenciaram.

Caso, algum dia, você tenha que escrever uma carta de recusa, os pesquisadores recomendam o seguinte:

- ✓ Não utilize evasivas para iniciar.
- ✓ Dê uma razão para a recusa, quando isso fizer com que o chefe do remetente não se zangue tanto.
- ✓ Apresente a mensagem negativa de forma positiva, mas clara; ofereça uma alternativa ou se coloque à disposição, se possível.
- ✓ Um final positivo não é necessário.



# Capítulo 10

## Organizando as Médias Através das Comparações Múltiplas

### *Neste Capítulo*

- ▶ Descobrimos quando e como acompanhar a ANOVA através das comparações múltiplas
- ▶ Comparando dois métodos muito conhecidos de comparação múltipla
- ▶ Levando em consideração os procedimentos adicionais

**I**magine: você está comparando as médias não de duas, mas de  $k$  populações independentes e descobre (usando a ANOVA; ver Capítulo 9) que deve rejeitar a  $H_0$ : todas as médias populacionais são iguais, e fica com a  $H_a$ : pelo menos duas das médias populacionais são diferentes. Agora você tem que saber — quais dessas populações são diferentes? A resposta a essa pergunta exige um procedimento de acompanhamento da análise de variância chamado *comparações múltiplas*, o que faz sentido, uma vez que vamos comparar as várias médias para ver quais são diferentes.

Neste capítulo, você vai descobrir quando é necessário usar um procedimento de comparação múltipla. Dois dos procedimentos de comparação múltipla mais conhecidos são o LSD de Fisher (least significant difference, que em português quer dizer diferença mínima significativa) e o teste de Tukey. Eles poderão ajudá-lo a responder esta difícil questão: OK, algumas das médias são diferentes, mas quais? Neste capítulo, também falo sobre outros procedimentos de comparação com os quais você pode se deparar ou querer experimentar.

**Observação:** os novos procedimentos de comparações múltiplas geralmente recebem o nome das pessoas que os criaram (da mesma forma que uma estrela é nomeada por seu descobridor, mas muito menos romântico e muito mais trabalhoso).



# Acompanhado a ANOVA

A principal razão para o uso da ANOVA na análise de dados é saber se há alguma diferença em um grupo de médias populacionais. Sua hipótese nula é a de que não há diferenças e a hipótese alternativa é a de que há pelo menos uma diferença entre duas das médias (observe que isso não quer dizer que todas as médias devem ser diferentes).

Se ficar provado que pelo menos duas das médias populacionais são diferentes, naturalmente, a próxima pergunta seria: "Mas quais são diferentes?" Embora esta pareça uma questão muito simples, a resposta para ela não é. O conceito de que as médias são diferentes pode ser interpretado de várias formas. Uma é maior do que todas as demais? Três pares se diferenciam uns dos outros e o resto é tudo igual? Os estatísticos têm trabalhado muito duro para chegar a um amplo conjunto de opções de procedimentos que investiguem e descubram as diferenças de todos os tipos entre duas ou mais médias populacionais. Essa família de procedimentos é chamada de comparações múltiplas.

Esta seção começa com um exemplo em que o procedimento ANOVA foi utilizado e a  $H_0$ , rejeitada, levando-o para o próximo passo: as comparações múltiplas. Assim, você vai ter uma visão geral de como e por que os procedimentos de comparação múltipla funcionam.

## Comparando o uso de minutos no celular: Um exemplo

Suponha que você queira comparar o número médio de minutos de celular usado mensalmente por várias faixas etárias definidas da seguinte forma:

- ✓ Grupo 1: 19 anos ou menos
- ✓ Grupo 2: 20–39 anos
- ✓ Grupo 3: Homens adultos de 40–59 anos
- ✓ Grupo 4: Mulheres adultas de 60 anos ou mais

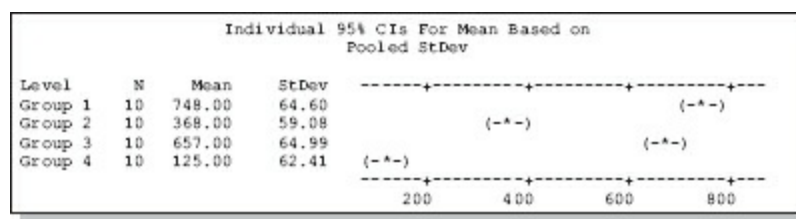
Você, então, coleta dados sobre uma amostra aleatória de dez pessoas pertencentes a cada grupo (onde ninguém conheça ninguém para manter a independência) e grava o número de minutos que cada pessoa usou em um mês. As dez primeiras linhas de um conjunto de dados hipotético são mostradas na Tabela 10-1.

Tabela 10-1		Minutos de Celular Usados em um Mês	
19 ou Menos (Grupo 1)	20–39 (Grupo 2)	40–59 (Grupo 3)	60 ou Mais (Grupo 4)
800	250	700	200
850	350	700	120
800	375	750	150
650	320	650	90
750	430	550	20

680	380	580	150
800	325	700	200
750	410	700	130
690	450	590	160
710	390	650	30

As médias e os desvios padrão dos dados da amostra são mostrados na Figura 10-1, bem como os intervalos de confiança para cada uma das médias populacionais separadamente. (Veja o Capítulo 3 para obter mais informações sobre os intervalos de confiança.) Observando a Figura 10-1, parece que todas as quatro médias são diferentes, sendo que o grupo 1 lidera o ranking, seguido pelos grupos 3 (que não fica muito atrás), 2 e 4 (na lanterna).

Sabendo que não se pode viver apenas dos resultados amostrais, você decide que a realização da ANOVA é necessária para ver se as diferenças que aparecem nas amostras podem ser estendidas à população (ver Capítulo 9). Usando o procedimento ANOVA, você testa se a média dos minutos utilizados é a mesma para todos os grupos. Os resultados da ANOVA, usando os dados da Tabela 10-1, estão ilustrados na Figura 10-2.



**Figura 10-1:** Estatística básica e intervalos de confiança para dados referentes a minutos de celular.

Observando a Figura 10-2, vemos que o teste- $F$  para a igualdade de todas as médias tem um valor- $p$  de 0,000, ou seja, ele é muito menor do que 0,001. Isso nos diz que pelo menos duas dessas faixas etárias tem uma diferença significativa com relação ao uso do celular. (Veja o Capítulo 9 para informações a respeito do teste- $F$  e seus resultados.)

ANOVA com um fator: Grupo 1, Grupo 2, Grupo 3, Grupo 4					
Source	DF	SS	MS	F	P
Factor	3	2416010	805337	204.13	0.000
Error	36	142030	3945		
Total	39	2558040			

S = 62.81 R-Sq = 94.5% R-Sq(adj) = 93.99%

**Figura 10-2:** Resultados da ANOVA para a comparação do uso do celular por quatro faixas etárias.

Então, qual é a sua próxima pergunta? Você apenas descobriu que o número médio de minutos de celular usado por mês não é o mesmo entre esses quatro grupos. Isso não

significa que todos os quatro grupos sejam diferentes (ver Capítulo 9), mas, necessariamente, significa que pelo menos dois grupos se diferenciam significativamente em relação ao uso do celular. Assim, suas perguntas passam a ser:

- ✓ Que grupos são diferentes?
- ✓ Como eles se diferenciam?

## ***Preparando o terreno para os procedimentos de comparação múltipla***

Determinar quais populações têm médias diferentes, depois de o teste- $F$  da ANOVA ter sido rejeitado, envolve a utilização de uma nova técnica de análise de dados chamada comparações múltiplas. A ideia básica dos procedimentos de comparações múltiplas é comparar as diversas médias e indicar onde estão e quais são as diferenças. Por exemplo, você pode concluir, a partir de um procedimento de comparação múltipla, que a primeira população tinha uma média estatisticamente inferior à da segunda, mas estatisticamente superior à média da terceira.

Existe uma infinidade de diferentes procedimentos de comparação múltipla lá fora; como saber qual deles usar e quando? Dois elementos básicos distinguem os diferentes procedimentos de comparações múltiplas: eu os chamo de finalidade e preço.

- ✓ **Finalidade:** Quando se sabe que em um grupo de médias nem todas são iguais, aumenta-se o foco para investigar as relações entre elas, dependendo da finalidade de sua pesquisa. Talvez você só queira descobrir quais médias se equivalem e quais não. Talvez seja melhor classificá-las em grupos estatisticamente equivalentes da menor para a maior. Ou também pode ser importante comparar a média de um grupo de médias à média de outro grupo. Os diferentes procedimentos de comparação múltipla foram construídos para diferentes fins. Em sua grande maioria, se você usá-los para os fins que foram projetados, terá mais chances de encontrar as diferenças específicas que procura, desde que essas diferenças realmente estejam lá.
- ✓ **Preço:** Qualquer procedimento estatístico tem um preço: a probabilidade de cometer o erro Tipo I em suas conclusões em algum momento do procedimento, devido ao acaso. (O *erro Tipo I* é cometido quando a  $H_0$  é rejeitada quando não deveria; ou seja, você acha que duas médias são diferentes, mas elas não são. Consulte um livro de Estatística I ou meu livro *Estatística Para Leigos*, da Alta Books, para mais informações.) A probabilidade de cometer pelo menos um erro Tipo I durante um procedimento de comparações múltiplas é chamada de *taxa de erro global* (também conhecida como taxa de erro por experimento (EER) ou taxa de erro familywise). Obviamente, desejamos as pequenas taxas de erro global. Cada procedimento de comparação múltipla tem sua própria taxa de erro global. Geralmente, quanto mais específicas são as relações que se está tentando encontrar,

menor a taxa de erro global, desde que você esteja usando um procedimento projetado para sua finalidade.

Na próxima seção, descrevo dois procedimentos de comparações múltiplas para mil e uma finalidades: o LSD de Fisher e o teste de Tukey.



Não tente investigar os dados com um procedimento de comparação múltipla se o teste para a igualdade das populações não foi rejeitado. Neste caso, você deve concluir que não tem provas suficientes para dizer que as médias da população não são todas iguais e, portanto, deve parar por aí. Sempre examine o valor- $p$  do teste- $F$  na saída da ANOVA antes de começar qualquer procedimento de comparações múltiplas.

# Identificando as Médias Diferentes com Fisher e Tukey

Depois de ter realizado uma ANOVA para ver se um grupo de  $k$  populações tem a mesma média e rejeitado  $H_0$ , você conclui que, pelo menos duas dessas populações têm médias diferentes. Mas não precisa parar por aqui, você pode seguir em frente a fim de descobrir quantas e quais são as médias diferentes por meio dos testes de comparações múltiplas.

Nesta seção, você vai conhecer dois dos mais conhecidos métodos de comparação múltipla: o *LSD de Fisher* (também conhecido como *LSD protegido de Fisher* ou *teste de Fisher*) e o *teste de Tukey* (também conhecido como *intervalos de confiança simultâneos de Tukey*).

Embora eu só discuta dois procedimentos em detalhes neste capítulo, existe uma infinidade de outros procedimentos de comparação múltipla. (Consulte a seção "Tantos outros procedimentos, tão pouco tempo!" no final deste capítulo.) Embora os métodos dos outros procedimentos sejam muito diferentes uns dos outros, seu objetivo global é o mesmo: descobrir quais médias populacionais se diferem através da comparação das médias amostrais.

## Pescando diferenças com o LSD de Fisher

Nesta seção, descrevo o procedimento original da diferença mínima significativa (LSD) e a melhoria feita por R.A. Fisher (devidamente chamado de procedimento da diferença mínima significativa de Fisher, ou o LSD de Fisher). Tanto o LSD quanto os procedimentos LSD de Fisher comparam os dois pares de médias utilizando uma forma de testes- $t$ , mas o fazem de formas diferentes (veja o Capítulo 3, ou consulte seu livro de Estatística I para saber mais informações sobre o teste- $t$ ). Aqui, você também vai ver o LSD de Fisher aplicado ao exemplo do celular citado no início deste capítulo (consulte a seção "Depois da ANOVA").

### O procedimento LSD original

Para usar o LSD (abreviação para diferença mínima significativa — *least significant difference*) original (pré-Fisher), basta escolher alguns pares de médias e realizar um teste- $t$  para cada par com nível  $\alpha = 0,05$  para procurar diferenças. O LSD não requer que um teste de ANOVA seja realizado antes (problema que mais tarde foi percebido por R.A. Fisher). Se as  $k$  médias populacionais fossem comparadas em pares pelo LSD, o número de testes- $t$  realizados seria representado por  $\frac{k(k-1)}{2}$ .

Veja como contar o número de testes- $t$  quando todas as médias são comparadas. Para começar, compare a primeira e a segunda médias, a primeira e a terceira, e assim por diante, até comparar a primeira e a  $k$ -ésima médias. Em seguida, compare a segunda e a terceira, a segunda e a quarta, e assim por diante, até comparar a  $(k-1)$ -ésima e  $k$ -ésima médias. O número total de pares de médias a serem comparados é igual a  $k * (k-1)$ . Uma

vez que a ordem de comparação das duas médias (média um e média dois versus média dois e média um) gera o mesmo resultado, independentemente de qual seja a maior, divida o total por 2 para evitar a duplicidade na contagem. Por exemplo, se você tiver quatro populações identificadas como A, B, C e D, deverá realizar  $\frac{4(4-1)}{2} = 6$  testes- $t$ : A versus B; A versus C; A versus D, B versus C, B versus D e C versus D.

O procedimento LSD original é muito simples, fácil de ser conduzido e fácil de ser entendido, mas tem alguns problemas. Uma vez que cada teste- $t$  é realizado com nível  $\alpha$  igual a 0,05, existe uma chance de 5% de que cada teste realizado cometa um erro Tipo I (rejeitar  $H_0$  quando ela não deveria ser rejeitada, conforme explico no Capítulo 3).

Embora uma taxa de erro de 5% para cada teste não pareça ser tão ruim, os erros possuem um efeito multiplicador à medida que o número de testes aumenta. Por exemplo, a chance de cometer pelo menos um erro do Tipo I com seis testes- $t$ , cada um com nível  $\alpha = 0,05$ , é 26,50%, que é a sua taxa de erro global para o procedimento.

Se você quer ou precisa saber como cheguei a 26,50% para a taxa de erro geral do exemplo anterior, veja aqui: a probabilidade de cometer um erro Tipo I para cada teste é de 0,05. A chance de cometer pelo menos um erro em seis testes é igual a  $1 -$  a probabilidade de não cometer erros em seis testes. A probabilidade de não cometer um erro em um teste é  $1 - \alpha = 0,95$ . A chance de não haver nenhum erro em seis testes é essa mesma quantidade vezes ela mesma, seis vezes, ou  $(0,95)^6$ , que equivale a 0,735. Agora, faça  $1 -$  essa quantidade e obtenha  $1 - 0,735 = 0,2650$  ou 26,50%.

## *Usando o novo e aperfeiçoado LSD de Fisher*

R.A. Fisher sugeriu uma melhoria em relação ao procedimento LSD original, que acabou gerando o procedimento conhecido como LSD de Fisher, ou LSD protegido de Fisher. Ele acrescenta a exigência de que um teste- $F$  ANOVA seja realizado e rejeitado antes que qualquer par de médias possa ser comparado individualmente ou coletivamente. Ao exigir que o teste- $F$  seja rejeitado, conclui-se que existe pelo menos uma diferença entre as médias. Ao adicionar este requisito, a taxa de erro global do LSD de Fisher passa a se encontrar em algum lugar na área de  $\alpha$ , valor muito inferior ao que você consegue com o procedimento LSD original.

A desvantagem do LSD de Fisher é que, uma vez que cada teste- $t$  é conduzido ao nível  $\alpha$  e a taxa de erro global também é próxima de  $\alpha$ , ele é bom para encontrar diferenças que realmente existam, mas também produz alguns alarmes falsos no processo (principalmente, dizendo que há uma diferença, quando na verdade ela não existe).

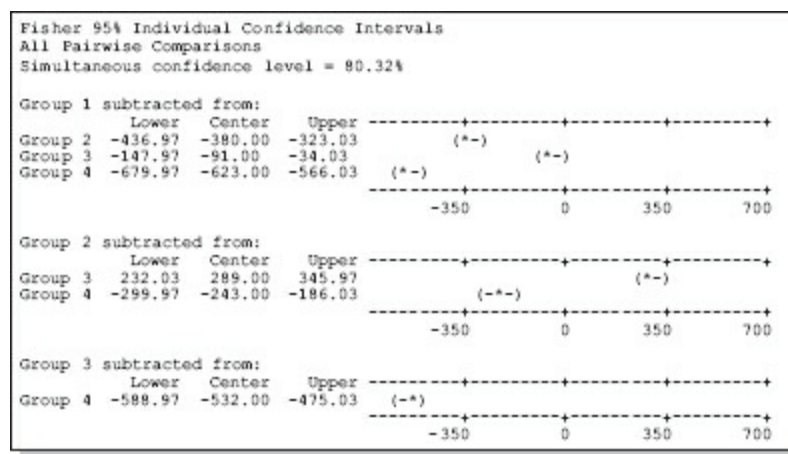
Para realizar o LSD de Fisher no Minitab, clique no menu Stat>ANOVA>One-way or One-way unstacked. (Se os seus dados aparecem em duas colunas, com a coluna um representando o número da população, e a coluna 2, a resposta, apenas clique em One-way, pois seus dados estão empilhados. Se os dados são mostrados em  $k$  colunas, uma para cada

uma das  $k$  populações, clique em One-way unstacked.) Selecione os dados para os grupos que estão sendo comparados e clique em Select. Em seguida, clique em Comparisons e depois em Fisher's. A taxa de erro individual aparece como 5 (por cento), que é o padrão. Se você quiser mudá-la, digite a taxa de erro desejada (entre 0,5 e 0,001) e clique em OK. Você pode digitar a taxa de erro em forma de decimal, 0,05, ou como um número maior do que 1, como o 5. Os números maiores do 1 são interpretados como porcentagens.

Um procedimento ANOVA foi realizado para os dados referentes ao uso de celulares apresentados na Tabela 10-1 a fim de comparar o número médio de minutos utilizados por quatro faixas etárias. Analisando a saída ilustrada na Figura 10-2, é possível ver que a  $H_0$  (todas as médias populacionais são iguais) foi rejeitada. O próximo passo é a realização de comparações múltiplas, usando o LSD de Fisher para saber quais médias populacionais diferem. A Figura 10-3 mostra o resultado da análise do Minitab para esses testes.

O primeiro bloco de resultados mostra "Grupo 1 subtraído de" onde o Grupo 1 = 19 anos ou menos. Cada linha depois dessa representa as outras faixas etárias (Grupo 2 = 20 a 39 anos, Grupo 3 = 40 a 59 anos e Grupo 4 = 60 anos ou mais). Cada linha mostra os resultados da comparação da média de algum grupo menos a média do Grupo 1.

Por exemplo, a primeira linha mostra o Grupo 2 sendo comparado ao Grupo 1. Movendo-se para a direita na mesma linha, você verá o intervalo de confiança para a diferença entre essas duas médias, que acaba sendo  $-436,97$  a  $-323,03$ . Já que o zero não está contido neste intervalo, você conclui que essas duas médias também são diferentes nas populações. Além disso, uma vez que essa diferença ( $\mu_2 - \mu_1$ ) é negativa, você também pode dizer que  $\mu_2$  é menor do que  $\mu_1$ . Uma maneira melhor de pensar isso poderia ser que  $\mu_1$  é maior do que  $\mu_2$ . Ou seja, a média do Grupo 1 é superior à média do Grupo 2.



**Figura 10-3:** Saída mostrando a aplicação do LSD de Fisher aos dados referentes ao uso de celular.

Se duas médias são iguais, a diferença entre elas é igual a zero e seu intervalo de confiança deve incluí-lo. Se o zero não for incluído, pode-se dizer que as médias são diferentes.



Neste caso, cada linha subsequente à linha "Grupo 1 subtraído de", mostrada na Figura 10-3, demonstra resultados semelhantes. Nenhum dos intervalos de confiança contém zero, assim sendo, você conclui que a média de uso do celular para o Grupo 1 é diferente da média de uso de outro grupo qualquer.

Além disso, uma vez que todos os intervalos de confiança estão em território negativo, pode-se concluir que a média de uso do celular para os usuários com 19 anos é maior do que a de todos os outros. (Lembre-se: a média para o Grupo 1 é subtraída da média dos outros grupos, assim, uma diferença negativa significa que a média do Grupo 1 é maior.)

Este processo continua à medida que você se move por toda a saída, até que todos os seis pares de médias sejam comparados. E, então, você pode chegar a uma conclusão. Por exemplo, na segunda parte da saída, o Grupo 2 é subtraído dos Grupos 3 e 4. Você vê que o intervalo de confiança para o "Grupo 3" é (232,03, 345,97), intervalo que fornece possíveis valores para a média do Grupo 3 menos a média do Grupo 2. O intervalo é inteiramente positivo, então, conclui-se que a média do Grupo 3 é maior do que a média do Grupo 2 (de acordo com estes dados).

Na próxima linha, o intervalo para o Grupo 4 menos o Grupo 2 é  $-299,97$  a  $-186,03$ . Todos esses números são negativos, então, conclui-se que a média do Grupo 4 é inferior a do Grupo 2. Combine as conclusões para dizer que a média do Grupo 3 é maior do que a do Grupo 2, que é maior do que a do Grupo 4.

No exemplo do celular, nenhuma das médias é igual, e, com base nas indicações dos intervalos de confiança e nos resultados de todas as comparações pareadas, prevalece a seguinte ordem para a média do uso de celular:  $\mu_1 > \mu_3 > \mu_2 > \mu_4$ . (Dados hipotéticos à parte, pode ser que o grupo de homens de 40 a 59 anos usem seus celulares por mais tempo por causa do trabalho.) Comparando esses resultados com as médias amostrais da Figura 10-1, essa ordem faz sentido, e as médias estão separadas o suficiente para serem declaradas como estatisticamente significativas.

Observe que próximo à parte superior da Figura 10-3 você vê "Simultaneous confidence level = 80.32%". Isso significa que a taxa de erro global para este procedimento é  $1 - 0,8032 = 0,1968$ , ou seja, próximo a 20% — um pouquinho alta.

## ***O teste de Tukey***

A ideia básica por trás do teste de Tukey é fornecer uma série de testes simultâneos para as diferenças nas médias. Ele ainda analisa todos os possíveis pares de médias, além de manter a taxa de erro global e a taxa individual de erro Tipo I para cada par de médias em  $\alpha$ . O que o distingue é o fato de realizar todos os testes ao mesmo tempo.

Embora os detalhes das fórmulas usadas para o teste de Tukey estejam fora do escopo deste livro, eles não se baseiam no teste-*t*, mas, sim, em uma estatística chamada *amplitude estudentizada*, que, por sua vez, baseia-se na média mais alta e na mais baixa do grupo e

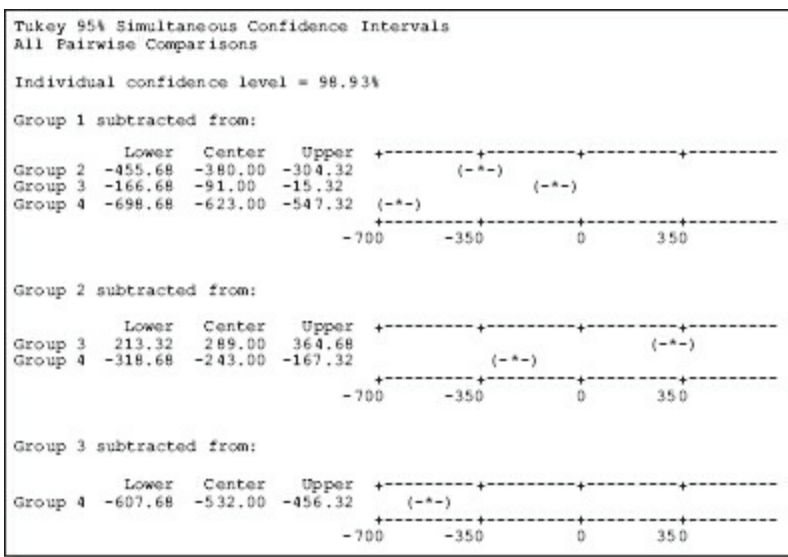


na diferença entre elas. As taxas de erro individual são mantidas em 0,05, pois Tukey desenvolveu um valor de corte para sua estatística de teste que se baseia em todas as comparações pareadas (independentemente do número de médias em cada grupo).

Para realizar o teste de Tukey no Minitab, clique no menu Stat>ANOVA>Oneway ou One-way unstacked. (Se os seus dados aparecem em duas colunas com a coluna 1 representando o número da população e a coluna 2, a resposta, apenas clique em One-way, pois seus dados estão empilhados. Se os dados são mostrados em  $k$  colunas, uma para cada uma das  $k$  populações, clique em One-way unstacked.) Selecione os dados para os grupos que estão sendo comparados e clique em Select. Em seguida, clique em Comparisons e, depois, em Tukey's. A taxa de erro familywise (global) aparece como 5 (por cento), que é o padrão. Se você quiser mudá-la, digite a taxa de erro desejada (entre 0,5 e 0,001) e clique em OK. Você pode digitar a taxa de erro em forma decimal, 0,05, ou como um número maior do que 1, como o 5. Os números maiores do 1 são interpretados como porcentagens.

A saída do Minitab para a comparação dos grupos do exemplo em questão conduzida por meio do teste de Tukey aparece na Figura 10-4. Estes resultados podem ser interpretados da mesma forma usada na interpretação dos resultados da Figura 10-3. Alguns dos números nos intervalos de confiança são diferentes, mas, neste caso, as principais conclusões continuam sendo as mesmas: o grupo das pessoas com 19 anos, ou menos, é o que mais usa o celular, seguido pelo grupo dos homens de 40 a 59 anos, das pessoas de 20 a 39 anos e, por último, o grupo das mulheres com 60 anos ou mais.

Os resultados de Fisher e de Tukey nem sempre são iguais, geralmente porque a taxa de erro global do processo de Fisher é maior do que a do teste de Tukey (exceto quando apenas duas médias estão envolvidas). A maioria dos estatísticos que conheço prefere o procedimento de Tukey ao de Fisher. Isso não significa que não tenham outros procedimentos dos quais gostem mais ainda, no entanto, o teste de Tukey é um procedimento comumente usado e muito aceito.



**Figura 10-4:** Saída para o teste de Tukey usado na comparação do uso de celulares.

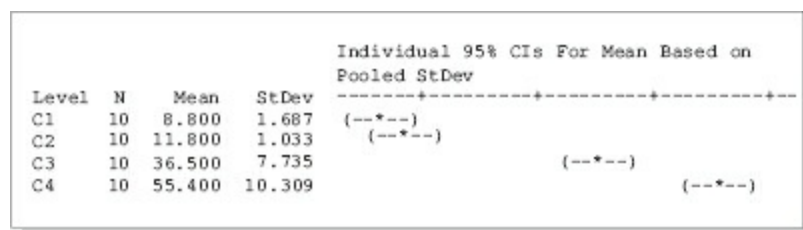
### *Examinando a Saída para Determinar a Análise*

Às vezes, o processo de responder perguntas é invertido nos cursos de Estatística. Em vez de lhe fazer uma pergunta e pedir que você use a informação do computador para respondê-la, o professor pode lhe dar a saída computacional e lhe pedir para determinar a pergunta respondida pela análise (algo como um programa de perguntas e respostas). Para fazer o caminho inverso, rumo à pergunta, você deve procurar pistas que indiquem o tipo de análise feita e depois montar o quebra-cabeça usando o que já sabe sobre aquele determinado tipo de análise.

Por exemplo, o professor lhe dá uma saída comparando as idades de dez consumidores de cada uma de quatro marcas de cereais, designados de C1–C4 (veja a Figura 10-5). Na análise, você pode ver a idade média dos consumidores dos quatro cereais sendo comparadas umas às outras, mas ela também mostra e compara os intervalos de confiança para essas médias. Sendo assim, a comparação dos intervalos de confiança lhe mostra que se trata de um procedimento de comparação múltipla.

Lembre-se: você está analisando se os intervalos de confiança de cada grupo de cereais se sobrepõem; se não o fizerem, isso significa que os consumidores dos quatro cereais têm diferentes idades médias. Mas, se eles se sobrepuserem, as médias de idade dos consumidores não poderão ser declaradas diferentes.

Com base nos dados da Figura 10-5, você pode ver que os cereais um (C1) e dois (C2) não são significativamente diferentes, mas, no caso do cereal três (C3), percebe-se que os consumidores têm uma idade média mais elevada do que a idade dos consumidores do C1 e C2. Já os consumidores do cereal quatro (C4) têm uma idade significativamente mais alta do que a idade dos outros três grupos. Depois do procedimento de comparações múltiplas, você sabe quais são os cereais diferentes e como eles se comportam uns em relação aos outros.



**Figura 10-5:** Resultados da comparação múltipla para o exemplo dos cereais.

Alguns procedimentos de comparações múltiplas lhe dão os grupos de médias equivalentes, os diferentes ou os sobrepostos. Neste caso, o resultado final é  $\mu_{C1} = \mu_{C2} < \mu_{C3} < \mu_{C4}$ .

# *Tantos Outros Procedimentos, Tão Pouco Tempo!*

Existem muitos procedimentos de comparação além das imaginações de Fisher e de Tukey. Os que discuto nesta seção são um pouco mais especializados na finalidade a que foram projetados em relação ao teste de Tukey e de Fisher. Por exemplo, você pode querer saber se uma determinada combinação de médias é maior do que outra combinação; ou talvez você possa querer comparar apenas as médias específicas, e não todos os pares de médias.



Uma coisa que deve se observar, no entanto, é que, em muitos casos, você não sabe exatamente o que está procurando quando compara médias — está apenas procurando diferenças. Se esse for o caso, o caminho a percorrer é o de um procedimento mais geral, como o de Fisher ou o de Tukey, construídos com a finalidade de realizar uma investigação geral e o fazem melhor do que os procedimentos mais especializados.

Esta seção fornece uma visão geral dos outros procedimentos de comparações múltiplas existentes e fala um pouco sobre cada um, incluindo as pessoas que os desenvolveram. Dadas as datas de quando esses procedimentos foram desenvolvidos, acho que você vai concordar comigo que a década de 1950 foi a época dourada para os procedimentos de comparação múltipla.

## *Cortando a conversa fiada com o ajustamento de Bonferroni*

O *ajustamento de Bonferroni* (ou *correção de Bonferroni*) é uma técnica usada em uma série de situações, e não apenas para comparações múltiplas. Ela foi criada basicamente para impedir que as pessoas superanalisassem os dados. Há um limite para o que pode ser feito em se tratando de análise de dados; existe uma linha que, quando ultrapassada, resulta em uma coisa que os estatísticos chamam de *data snooping*. E o ajustamento de Bonferroni veio justamente para limitar isso.

O data snooping ocorre quando alguém analisa seus dados várias e várias vezes, até conseguir um resultado que possa dizer ser estatisticamente significativo (ou seja, há poucas chances de que o resultado tenha acontecido por acaso; veja o Capítulo 3). Como o número de testes realizados pelo data snooper é alto demais, é provável que ele encontre algo significativo apenas por acaso. E, muito provavelmente, esse resultado será falso.


Por exemplo, suponha que um pesquisador deseje descobrir as variáveis relacionadas à venda de chinelos de quarto. Ele coleta dados sobre tudo o que consegue imaginar, incluindo o tamanho dos pés das pessoas, a frequência com que elas saem para pegar o jornal usando os chinelos e suas cores favoritas. Porém, não encontrando nada de significativo, ele passa a examinar o estado civil, a idade e a renda das pessoas.

Ainda sem fazer muito progresso, ele, então, extrapola e começa a relacionar a cor do cabelo, se as pessoas já foram ou não ao circo e seus lugares preferidos no avião (janela ou corredor, senhor?). Então, como era de se esperar, ele consegue. Segundo seus dados, as pessoas que se sentam nos corredores dos aviões são mais propensas a comprar chinelos

de quarto do que aquelas que se sentam à janela ou no meio.


O que há de errado nisso? Testes demais! Cada vez que o pesquisador analisa uma variável e realiza um teste para ela, ele deve escolher um nível  $\alpha$  em que o teste será realizado. (Lembre-se de que o nível  $\alpha$  é a probabilidade a que você se dispõe em rejeitar a hipótese nula e acionar um falso alarme.) Conforme o número de testes aumenta, o nível  $\alpha$  também cresce.

Suponha que o  $\alpha$  escolhido seja 0,05. O pesquisador, então, tem uma chance de 5% de se enganar ao encontrar uma conclusão significativa, apenas em virtude do acaso. Então, se ele conduzir 100 testes, cada um com 5% de chance de erro, em média, 5 dos 100 testes irão gerar um resultado estatisticamente significativo apenas por acaso. No entanto, os pesquisadores que não sabem disso (ou que sabem, mas o fazem mesmo assim) encontram resultados ditos significativos, embora, na verdade, sejam falsos.



Em 1950, um matemático italiano chamado Carlo Emilio Bonferroni (1892– 1960) disse "basta!" e criou o que os estatísticos chamam de ajustamento de Bonferroni para controlar esses disparates. O *ajustamento de Bonferroni* simplesmente diz que se você fizer  $k$  testes para seus dados, não poderá realizar cada um ao nível  $\alpha = 0,05$ , mas deverá ter um nível  $\alpha$  para cada teste igual a  $0,05 \div k$ .

Por exemplo, alguém que realiza 20 testes para um conjunto de dados tem que fazer cada um a um nível  $\alpha = 0,05 \div 20 = 0,0025$ . Esse ajustamento faz com que seja mais difícil encontrar uma conclusão significativa, pois o valor- $p$  de qualquer teste deve ser inferior a 0,0025. O ajustamento de Bonferroni freia as chances de ocorrência de data snooping até você encontrar algo falso.



A desvantagem do ajustamento de Bonferroni é que ele é muito conservador. Embora reduza a possibilidade de concluir que duas médias são diferentes, quando, na verdade, elas não o são, essa técnica não consegue localizar algumas diferenças que realmente existem. Estatisticamente falando, o ajustamento de Bonferroni tem problemas com o poder estatístico. (Consulte um livro de Estatística I ou o *Estatística Para Leigos* para uma discussão sobre poder estatístico.)

## ***Comparando combinações usando o método de Scheffe***

O *método de Scheffe* foi desenvolvido em 1953 por Henry Scheffe (1907–1977). Esse método não apenas compara duas médias ao mesmo tempo, como os testes de Tukey e Fisher, mas também todas as diferentes combinações (chamadas contrastes) das médias. Por exemplo, se você tem as médias de quatro populações, pode querer testar para ver se a soma delas é igual a um determinado valor, ou se a média de duas delas é igual à média das outras duas.

## ***O teste de Dunnett***

O teste de Dunnett foi desenvolvido em 1955 por Charles Dunnett (1921–1977). O *teste de Dunnett* é um procedimento de comparação múltipla especial utilizado em um experimento planejado que contenha um grupo de controle. Tal teste compara cada grupo de tratamento ao grupo de controle e determina quais são os melhores tratamentos.

Comparado a outros procedimentos de comparações múltiplas, o teste de Dunnett tem maior capacidade de encontrar verdadeiras diferenças nesse tipo de situação, pois se concentra apenas sobre as diferenças entre cada grupo de tratamento e de controle — e não sobre as diferenças entre todos os pares de tratamentos do estudo.

## ***O teste de Student Newman-Keuls***

O teste de Student Newman-Keuls é uma abordagem diferente da abordagem de Tukey e Fisher para a comparação de pares de médias em um procedimento de comparação múltipla. Este teste é o resultado do trabalho de três pessoas: "Student", Newman e Keuls.

O *procedimento Student Newman-Keuls* baseia-se em uma abordagem por etapas ou camadas. Primeiro, você deve ordenar as médias amostrais da menor para a maior e, em seguida, examinar as diferenças entre elas.

Teste primeiro a maior diferença menos a menor e, caso ela seja estatisticamente significativa, conclua que as duas respectivas populações se diferenciam em relação às médias. Das médias restantes, busque uma diferença significativa entre as que estiverem mais afastadas na ordem, e assim por diante. Pare quando não encontrar mais diferenças.

## ***O teste de Duncan***

David B. Duncan desenvolveu o *teste de Duncan* em 1955. O teste se baseia no teste de Student Newman-Keuls, mas aumentou o poder de sua capacidade em detectar quando a hipótese nula não é verdadeira (veja o Capítulo 3), pois aumenta o valor de  $\alpha$  a cada etapa do teste de Student Newman-Keuls. O teste de Duncan é usado especialmente em agronomia (cultivo e manejo da terra) e outros tipos de pesquisas agrícolas. Uma das coisas mais interessantes em ser um estatístico é que você nunca sabe com que tipos de problemas vai trabalhar ou quem vai usar seus métodos e resultados.

Embora Duncan tenha ganhado o apoio de muitos pesquisadores que utilizaram o teste (e ainda ganha), ele sofreu várias críticas. John Tukey (criador do teste de Tukey) e Henry Scheffe (criador do teste de Scheffe) acusaram o teste de Duncan de ser muito liberal, uma vez que não controla a taxa de erro global (conhecida entre os grandes como *taxa de erro familywise*). Porém, Duncan manteve sua posição. Ele disse que, de qualquer forma, as médias geralmente nunca são iguais e, portanto, preferia cometer o erro de acionar um alarme falso (erro Tipo I) a perder uma oportunidade (erro Tipo II) em descobrir quando as médias são diferentes.

Todo procedimento em Estatística tem alguma chance de nos levar à conclusão errada, não



por causa de um erro no processo, mas pelo fato de que os resultados variam de conjunto de dados para conjunto de dados. Você precisa apenas conhecer sua situação e escolher o procedimento que melhor funciona para ela. Em caso de dúvida, consulte um estatístico para ajudá-lo a resolver tudo isso.

## A vida secreta dos estatísticos

Às vezes, é difícil imaginar que pessoas famosas possam ter uma vida normal, e pode ser ainda mais difícil imaginar estatísticos fazendo outra coisa que não seja ficar em uma sala fazendo cálculos. Mas a verdade é que os estatísticos famosos são pessoas interessantes que têm vidas interessantes, assim como você e eu. Considere estas estrelas do mundo da estatística:

- ✓ **Scheffe Henry:** Scheffe foi um estatístico excepcional na University of California em Berkeley. Um de seus cinco livros, *A Análise de Variância*, escrito em 1959, é um clássico sobre o assunto, e ainda hoje é usado (usei na faculdade e ainda tenho uma cópia em meu escritório). Scheffe gostava de trilhas, natação, ciclismo, literatura e música, sendo que aprendeu a tocar flauta doce quando adulto. Infelizmente, morreu em um acidente de bicicleta no caminho para a universidade em 1977.
- ✓ **Dunnett Charles:** Apelidado de “Charlie”, (você imaginava que estatísticos famosos pudessem ter apelidos?), Dunnett foi um excepcional e premiado professor nos Departamentos de Matemática, Estatística, Epidemiologia Clínica e Bioestatística da McMaster University, em Ontário, no Canadá. Escreveu muitos artigos, dos quais dois foram tão importantes, que foram incluídos na lista dos 25 artigos estatísticos mais citados de todos os tempos.
- ✓ **William Sealy Gosset, ou “Student”:** O primeiro nome incluído no teste de Student Newman-Keuls é uma história por si só. “Student” é o pseudônimo do estatístico inglês William Sealy Gosset (1876–1937). Gosset trabalhava como estatístico para a cervejaria Guinness, em Dublin, na Irlanda, quando se tornou famoso por desenvolver o teste- $t$ , também conhecido como a distribuição- $t$  de Student (veja o Capítulo 3), um dos testes de hipóteses mais utilizados em estatística no mundo. Gosset planejou o teste- $t$  como uma forma mais barata de monitorar a qualidade da cerveja. Ele, então, publicou seu trabalho na melhor das revistas científicas de Estatística, mas seu empregador, considerando que o uso de suas estatísticas no controle da qualidade seria um segredo industrial, não o deixou usar seu verdadeiro nome em suas publicações (embora todos os seus companheiros soubessem exatamente quem era “Student”). Então, se não fosse a cerveja Guinness, o teste- $t$  de Student teria sido chamado de teste- $t$  de Gossett (ou talvez você estaria bebendo a cerveja “Gosset”).

## *Ficando não paramétrico com o teste de Kruskal-Wallis*

O teste de Kruskal-Wallis foi desenvolvido em 1952 pelos estatísticos americanos, William Kruskal (1919–2005) e W. Allen Wallis (1912–1998). O *teste de Kruskal-Wallis* é a versão não paramétrica para um procedimento de comparação múltipla. Os procedimentos não paramétricos não têm tantas condições a serem satisfeitas como seus

correspondentes tradicionais. Todos os outros procedimentos descritos neste capítulo requerem que as populações estejam em distribuições normais e, muitas vezes, também requerem a igualdade de variação.

O teste de Kruskal-Wallis não utiliza os valores reais dos dados, mas suas classificações (ordenação do menor para o maior). O teste classifica todos os dados juntos e, em seguida, analisa a forma como suas posições estão distribuídas entre as amostras que representam populações individuais. Se uma amostra obtém todas as posições menores, conclui-se que a população a que ela corresponde tem uma média menor do que a das outras, e assim por diante. (Vá ao Capítulo 16 para a história completa sobre as estatísticas não paramétricas e ao Capítulo 19 para todos os detalhes sobre o teste de Kruskal-Wallis.)



# Capítulo 11

## Percorrendo os Caminhos da ANOVA com Dois Fatores

---

### *Neste Capítulo*

- ▶ Construindo e conduzindo uma ANOVA com dois fatores
  - ▶ Familiarizando-se com os (e procurando pelos) efeitos de interação e efeitos principais
  - ▶ Colocando os termos à prova
  - ▶ Desmistificando a tabela ANOVA com dois fatores
- 

**A** *análise de variância* (ANOVA) é muito utilizada em experimentos para verificar se diferentes níveis de uma variável explicativa ( $x$ ) obtêm diferentes resultados sobre uma variável quantitativa  $y$ . A variável  $x$ , neste caso, é chamada de *fator* e possui determinados níveis, dependendo de como o experimento está configurado.

Por exemplo, digamos que você queira comparar a variação média da pressão arterial em relação a determinadas doses de um medicamento. Sendo assim, o fator em questão aqui é a dosagem do medicamento. Suponha que ele tenha três níveis: 10mg por dia, 20mg por dia ou 30mg por dia. Suponha, também, que outra pessoa esteja estudando a resposta a esse mesmo medicamento e decida examinar se a frequência diária da administração do medicamento (uma ou duas vezes) tem algum efeito sobre a pressão arterial. Neste caso, o fator é o número de vezes que o remédio é ingerido ao dia e possui dois níveis: uma e duas vezes.

Suponha que você queira estudar os efeitos da dosagem *e* o número de vezes juntos, pois acredita que ambos possam ter efeito sobre a resposta. Então, o que você tem em mãos é a chamada *ANOVA de dois fatores*, que utiliza dois fatores em conjunto a fim de comparar a resposta média. É a extensão da ANOVA com um fator (veja o Capítulo 9), mas com um detalhe: os dois fatores usados, quando juntos, podem operar sobre a resposta de forma diferente do que quando separados.

Neste capítulo, primeiro vou lhe dar um exemplo de quando é preciso usar uma ANOVA com dois fatores. Depois, mostro-lhe como configurar o modelo, usar a tabela ANOVA, conduzir os testes- $F$  e tirar as devidas conclusões.



# Configurando o Modelo ANOVA com Dois Fatores

O modelo ANOVA com dois fatores expande as ideias do modelo ANOVA com um fator e adiciona um termo de interação para examinar como as várias combinações dos dois fatores influenciam a resposta. Nesta seção, você vai ver as bases para a construção de uma ANOVA com dois fatores: os tratamentos, os efeitos principais, o termo de interação e a equação das somas de quadrados, o responsável por juntar tudo.

## Determinando os tratamentos

O modelo ANOVA em questão contém dois fatores, A e B, e cada fator tem um certo número de níveis — dizemos  $i$  níveis do Fator A e  $j$  níveis do Fator B.

No exemplo do estudo sobre o medicamento, apresentado na introdução do capítulo, tem-se A = dosagem do medicamento com  $i = 1, 2$  ou  $3$  e B = número de vezes por dia com  $j = 1$  ou  $2$ . Cada pessoa envolvida no estudo está sujeita a uma das três diferentes doses do medicamento e irá tomá-lo segundo um dos dois métodos apresentados. Isso significa que você tem  $3 * 2 = 6$  diferentes combinações entre os fatores A e B que podem ser aplicadas aos sujeitos e, portanto, poderá estudar essas combinações e seus efeitos sobre a variação da pressão arterial através do modelo ANOVA com dois fatores.

No modelo, cada diferente combinação de níveis dos fatores A e B é chamada de *tratamento*. A Tabela 11-1 mostra os seis tratamentos para o exemplo em questão. Por exemplo, o Tratamento 4 é a combinação de 20mg do medicamento administrada em duas doses de 10mg por dia.

Tabela 11-1 Seis Combinações de Tratamentos para um Ensaio Clínico

Dosagem	Uma Dose por Dia	Duas Doses por Dia
10mg	Tratamento 1	Tratamento 2
20mg	Tratamento 3	Tratamento 4
30mg	Tratamento 5	Tratamento 6

Se o fator A tem  $i$  níveis e o Fator B tem  $j$  níveis, você tem  $i * j$  diferentes combinações de tratamentos em seu modelo ANOVA com dois fatores.

## Em busca das somas de quadrados

O modelo ANOVA com dois fatores contém os três termos a seguir:

- ✓ **O efeito principal A:** Termo que se refere ao efeito do Fator A sobre a resposta.
- ✓ **O efeito principal B:** Termo que se refere ao efeito do Fator B sobre a resposta.

- ✓ **A interação entre A e B:** O efeito da combinação dos Fatores A e B (designado como AB).

A equação da soma de quadrados para a ANOVA com um fator (abordada no Capítulo 9) é  $SQTO = SQT + SQE$ , onde  $SQTO$  é a variabilidade total da variável resposta,  $y$ ;  $SQE$  é a variabilidade explicada pela variável tratamento (que vamos chamar de fator A); e  $SQE$  é a variabilidade dentro dos tratamentos, representando a variação em virtude do erro experimental.

A finalidade de um modelo ANOVA com um fator é verificar se os diferentes níveis do Fator A produzem diferentes respostas na variável  $y$ . A maneira de fazer isso é usando  $H_0: \mu_1 = \mu_2 = \dots = \mu_i$ , onde  $i$  é o número de níveis do Fator A (variável de tratamento). Se você rejeitar  $H_0$ , o Fator A (que separa os dados nos grupos que estão sendo comparados) é significativo. Se você não pode rejeitar  $H_0$ , não poderá concluir que o Fator A é significativo.

Na ANOVA com dois fatores, adiciona-se outro fator à mistura (B) mais um termo de interação (AB). A equação da soma de quadrados para o modelo ANOVA com dois fatores é  $SQTO = SQA + SQB + SQAB + SQE$ . Aqui,  $SQTO$  é a variabilidade total nos valores de  $y$ ;  $SQA$  é a soma de quadrados devido ao Fator A (que representa a variabilidade dos valores de  $y$  explicada pelo Fator A) e o mesmo para  $SQB$  e o Fator B.  $SQAB$  é a soma de quadrados devido à interação dos Fatores A e B, e  $SQE$  é a variabilidade deixada sem explicação e considerada erro.

Embora os detalhes matemáticos de todas as fórmulas desses termos sejam pesados e estejam além do escopo deste livro, eles apenas expandem as fórmulas para a ANOVA com um fator abordada no Capítulo 9. A ANOVA faz os cálculos por você, portanto, não precisa se preocupar com essa parte.

Para realizar uma ANOVA de dois fatores no Minitab, insira seus dados em três colunas.

- ✓ A coluna 1 contém as respostas (os dados reais).
- ✓ A coluna 2 representa o nível do Fator A (o Minitab o chama de *row factor*).
- ✓ A coluna 3 representa o nível do Fator B (o Minitab o chama de *column factor*).

Clique em Stat>Anova>Two-way. Clique em Coluna 1, no campo à esquerda, e ela aparece no campo Response, ao lado direito. Clique em Coluna 2, e ela aparece no campo row factor, clique em Coluna 3, e ela aparece no campo factor. Clique em OK.

Por exemplo, suponha que você tenha seis valores de dados em Coluna 1: 11, 21, 38, 14, 15 e 62. Suponha que a Coluna 2 contenha 1, 1, 2, 2, 2 e a Coluna 3 contenha 1, 2, 3, 1, 2, 3. Isso significa que o Fator A tem dois níveis (1, 2), e o Fator B, três (1, 2, 3). A Tabela 11-2 mostra o desdobramento dos valores de dados e quais combinações de níveis e

fatores se associam a eles.

**Tabela 11-2**                      **Dados e Seus Respectivos Níveis a Partir de Dois Fatores**

<i>Valor dos Dados</i>	<i>Nível do Fator A</i>	<i>Nível do Fator B</i>
11	1	1
21	1	2
38	1	3
14	2	1
15	2	2
62	2	3

Suponha que o Fator A tenha  $i$  níveis e o Fator B,  $j$  níveis, com uma amostra de tamanho  $m$  coletada para cada combinação de A e B. Os graus de liberdade para o Fator A, Fator B e para a interação AB são  $(i - 1)$ ,  $(j - 1)$  e  $(i - 1) * (j - 1)$ , respectivamente. Essa fórmula é apenas uma extensão dos graus de liberdade para o modelo com um fator para os Fatores A e B. A equação dos graus de liberdade para SQTO é  $(i * j * m) - 1$  e dos graus de liberdade para SQE é  $i * j * (m - 1)$ . (Consulte o Capítulo 9 para detalhes sobre os graus de liberdade.)

# *Entendendo os Efeitos da Interação*

O efeito de interação é o eixo central do modelo ANOVA com dois fatores. É importante saber e considerar que os dois fatores podem, quando juntos, agir de maneira diferente da qual agiriam se estivessem separados. Nesta seção, vamos ver as muitas maneiras em que a interação AB e os efeitos principais dos Fatores A e B afetam a variável de resposta em um modelo ANOVA com dois fatores.

## *Mas, afinal, o que é interação?*

A *Interação* é quando dois fatores se encontram, ou interagem um com o outro, influenciando a resposta de forma diferente da qual cada fator a afetaria se estivessem separados.

Por exemplo, antes de verificar se a dosagem do medicamento (o Fator A) ou o número de vezes administradas (o Fator B) são importantes para explicar as variações na pressão arterial, você deve observar de que forma eles, juntos, influenciam a pressão sanguínea. Ou seja, você deve examinar o termo interação.

Suponha que esteja tomando um tipo de medicamento para o colesterol e outro para um problema cardíaco. Imagine que os pesquisadores tenham observado somente os efeitos de cada medicamento individualmente, dizendo que cada um deles foi aprovado para o controle do problema para o qual foram produzidos com pouco ou nenhum efeito colateral. Agora, você mistura os dois medicamentos em seu organismo. Com relação aos resultados do estudo individual, não há problema algum. No entanto, se esses estudos não continuarem, ninguém vai saber como os medicamentos interagem um com o outro, e você poderá se encenar rapidinho se tomá-los juntos.

Felizmente, as indústrias farmacêuticas e os pesquisadores trabalham muito para estudar as interações entre medicamentos e seu farmacêutico também conhece essas interações. Mas pode apostar que um dia um estatístico esteve envolvido nesse trabalho!

Outro bom exemplo de como a interação funciona está na cozinha. Ingira um ovo cru de uma só vez, beba um copo de leite e coma uma xícara de açúcar, uma de farinha e uma colher de margarina. Depois, tome uma xícara de gotas de chocolate. Cada um desses itens tem um determinado sabor, efeito e uma determinada textura sobre o seu paladar, que, na maioria dos casos, não é dos melhores. Mas misture-os em uma tigela e voilá! Você tem um monte de biscoitos de chocolate graças aos efeitos mágicos da interação.

Na ANOVA com dois fatores, primeiro você deve verificar o termo de interação. Se A e B interagem entre si e essa interação é estatisticamente significativa, você não deve examinar os efeitos de cada fator isoladamente. Seus efeitos estão interligados e não podem ser separados.



## ***Interagindo com os gráficos de interação***

No modelo ANOVA com dois fatores, você está lidando com dois fatores e a interação entre eles. A partir desse modelo, poderíamos conseguir uma série de resultados referentes à significância dos termos individuais, como você pode ver na lista a seguir:

- ✓ Fatores A e B são significativos.
- ✓ O Fator A é significativo, mas o Fator B não.
- ✓ O Fator B é significativo, mas o Fator A não.
- ✓ Nem o Fator A nem o Fator B são significativos.
- ✓ A interação AB é significativa, portanto, você não deve examinar A ou B isoladamente.

A Figura 11-1 mostra cada uma dessas cinco situações em um diagrama usando o exemplo do estudo de um medicamento. Os gráficos que mostram como os Fatores A e B reagem separadamente e em conjunto sobre a variável de resposta  $y$  são chamados de *gráficos de interação*. Nas próximas seções, descrevo cada uma dessas cinco situações em detalhes, explicando o que os gráficos lhe dizem e o significado dos resultados no contexto do exemplo dado.

### ***Fatores A e B são significativos.***

A Figura 11-1a mostra a situação em que A e B são significativos ao modelo e não há interação. As retas representam os níveis do Fator B (vezes ao dia); o eixo  $x$  representa os níveis do Fator A (dosagem); e o eixo  $y$  representa o valor médio da variável de resposta  $y$ , que é a variação da pressão arterial em cada combinação de tratamentos.

Para interpretar esses gráficos de interação, primeiro observe as tendências gerais de cada reta. A reta superior na Figura 11-1a está em ascendência, da esquerda para a direita, o que significa que, quando o medicamento é tomado duas vezes ao dia, a variação da pressão sanguínea aumenta à medida que o nível de dosagem também aumenta. A reta inferior demonstra um resultado semelhante quando o medicamento é tomado uma vez por dia; a variação da pressão arterial aumenta conforme o nível de dosagem também se eleva. Supondo que essas diferenças sejam grandes o suficiente, conclui-se que o nível de dosagem (Fator A) é significativo.

Agora, observe como as retas se comportam uma em relação à outra. Observe que as retas, embora paralelas, estão muito distantes. De modo geral, a variação total da pressão arterial é maior quando o medicamento é tomado duas vezes ao dia (reta superior) do que quando tomado uma vez ao dia (reta inferior). Novamente, supondo que essas diferenças sejam grandes o suficiente, conclui-se que a quantidade de vezes que o remédio é tomado ao dia (Fator B) é significativa.

Neste caso, as diferentes combinações entre os Fatores A e B não afetam de forma oposta

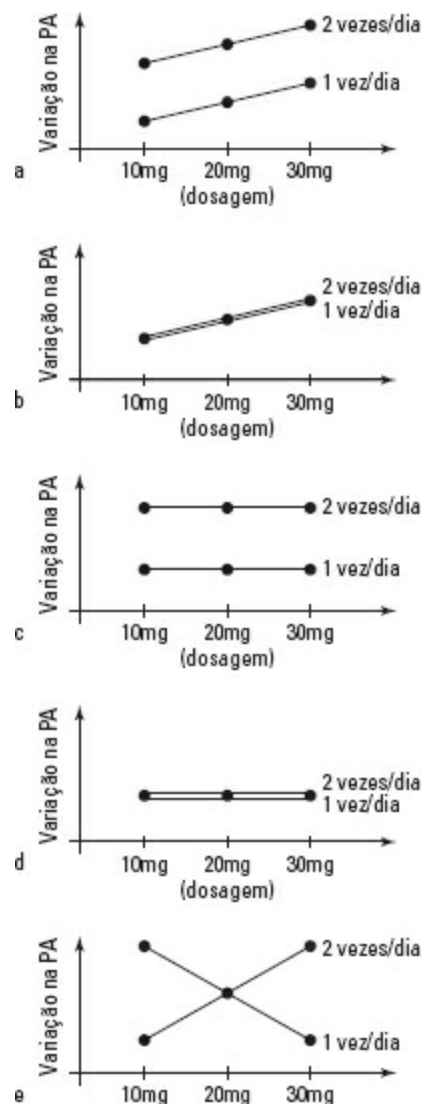
as tendências gerais da variação da pressão arterial (ou seja, as retas não se interceptam), assim, não há efeito da interação entre o nível de dosagem e a quantidade diária de vezes que o remédio é tomado.



A presença de duas retas paralelas em um gráfico de interação indica ausência de um efeito de interação. Em outras palavras, o efeito do Fator A sobre a resposta não se altera à medida que você se movimenta pelos diferentes níveis do Fator B. No exemplo em questão, os níveis de A não alteram a pressão arterial de forma distinta da forma que os diferentes níveis de B a alteram.

### ***O Fator A é significativo, mas o Fator B não***

A Figura 11-1b demonstra que a variação da pressão arterial (PA) aumenta ao longo do nível de dosagem para as pessoas que tomam o medicamento uma ou duas vezes ao dia. No entanto, as duas retas estão tão próximas, que não faz diferença se você tomar o remédio uma ou duas vezes por dia. Assim, o Fator A (dosagem) é significativo, mas o Fator B (vezes ao dia) não é. As retas paralelas indicam a ausência do efeito de interação.



**Figura 11-1:** Cinco exemplos de resultados para ANOVA de dois fatores com interação.

## ***O Fator B é significativo, mas o Fator A não***

A Figura 11-1c mostra onde o Fator B (vezes ao dia) é significativo, mas o Fator A (dosagem) não é. As retas ao longo dos níveis de dosagem são contínuas, indicando que a dosagem não tem efeito sobre a pressão arterial. No entanto, as duas retas para o fator de vezes ao dia se afastam e, assim, seu efeito sobre a pressão arterial é significativo. As retas paralelas indicam não haver efeito de interação.

## ***Nenhum dos fatores é significativo***

A Figura 11-1d mostra duas retas contínuas muito próximas uma da outra. Pelas discussões anteriores, a respeito das Figuras 11-1b e 11-1c, você pode imaginar que essa figura representa o caso em que, nem o Fator A nem o Fator B são significativos, e que não há um efeito de interação, pois as retas são paralelas.

## ***A interação AB é significativa***

Finalmente, chegamos à Figura 11-1e, o mais interessante de todos os gráficos de interação. O interessante aqui é que, uma vez que as duas retas se interceptam, os Fatores A e B interagem entre si de forma a influenciar a resposta. Se eles não interagissem, as retas deveriam ser paralelas.

Comece com a reta que ascende da esquerda para a direita (a que representa 2 vezes/dia). Essa reta mostra que, quando o medicamento é tomado duas vezes ao dia a uma baixa dosagem, tem-se uma pequena alteração na pressão arterial; à medida que a dosagem aumenta, a alteração da pressão arterial também aumenta. Mas, quando o remédio é tomado uma vez ao dia, acontece o oposto, como mostra a outra reta na Figura 11-1e, que desce da esquerda para a direita.

Se você não procurou um possível efeito de interação antes de examinar os efeitos principais, é porque pode ter pensado que, independentemente de quantas vezes tomar este medicamento ao dia, os efeitos serão os mesmos. Mas não é bem assim! Sempre verifique primeiro o termo de interação em qualquer ANOVA com dois fatores. Se o termo de interação for significativo, você não tem como afirmar que os efeitos foram causados apenas pelo Fator A ou pelo B.

A verificação dos efeitos principais do Fator A ou do B, sem a prévia verificação do termo de interação AB, é considerada uma proibição no mundo da ANOVA com dois fatores. Outro tabu é o exame individual dos fatores (também conhecido como a *análise dos efeitos principais*) quando o termo de interação é significativo.



# Testando os Termos na ANOVA com Dois Fatores

Na ANOVA com um fator, temos apenas um teste de hipótese geral; utilizamos o teste- $F$  para determinar se as médias dos valores de  $y$  são iguais ou diferentes ao longo dos níveis de um fator. Na ANOVA com dois fatores, temos mais itens para testar, além do modelo global. Para analisar, em primeiro lugar, temos o termo de interação AB e, possivelmente, os efeitos principais de A e B. Cada teste em uma ANOVA com dois fatores é um teste- $F$  baseado nas ideias da ANOVA com um fator (veja o Capítulo 9 para mais informações sobre o assunto).

Para conduzir os testes- $F$  para esses termos, você basicamente deve observar o quanto mais da variabilidade total nos valores de  $y$  pode ser explicado pelo termo que está testando em comparação com o que sobrou no termo de erro. Um valor grande para  $F$  significa que o termo testado é significativo.

Primeiro, teste se o termo de interação AB é significativo. Para isso, use a estatística de teste- $\frac{MQ_{AB}}{MQE}$ , que tem uma distribuição- $F$  com  $(i - 1) * (j - 1)$  graus de liberdade a partir de  $MQ_{AB}$  (média da soma de quadrados para o termo de interação entre A e B) e  $i * j * (m - 1)$  graus de liberdade a partir de  $MQE$  (média da soma de quadrados do erro), respectivamente. (Lembre-se de que  $i$  e  $j$  representam o número de níveis de A e B, e  $m$  é o tamanho da amostra para cada combinação de A e B.)

Se o termo de interação não for significativo, tire o termo AB do modelo e explore os efeitos isolados dos Fatores A e B em relação à variável de resposta  $y$ . O teste para o

Fator A utiliza a estatística de teste- $\frac{MQ_A}{MQE}$ , que tem uma distribuição- $F$  com  $(i - 1)$  graus de liberdade a partir de  $MQ_A$  (média da soma de quadrados para o Fator A) e  $i * j * (m - 1)$  graus de liberdade a partir de  $MQE$  (média da soma de quadrados do erro),

respectivamente. O teste para o Fator B utiliza a estatística de teste- $\frac{MQ_B}{MQE}$ , que tem uma distribuição- $F$  com  $(j - 1)$  e  $i * j * (m - 1)$  graus de liberdade. (Consulte o Capítulo 9 para todos os detalhes sobre testes- $F$ ,  $MQE$  e graus de liberdade).

Os resultados que você pode obter a partir dos testes feitos aos termos do modelo ANOVA são iguais aos representados na Figura 11-1. Todos aparecem na saída do Minitab descrita na próxima seção, incluindo a soma de quadrados, os graus de liberdade, a média da soma dos quadrados e os valores- $p$  para seus devidos testes- $F$ .





# *Executando uma Tabela ANOVA*

A tabela ANOVA para a ANOVA com dois fatores inclui os mesmos elementos da tabela ANOVA para a ANOVA com um fator (veja o Capítulo 9). Mas, onde na ANOVA com um fator você tem uma linha para as contribuições do Fator A, agora você adiciona linhas para os efeitos do Fator B e do termo de interação AB. O Minitab constrói a tabela ANOVA para você como parte da saída resultante da execução de uma ANOVA com dois fatores.

Nesta seção, você vai descobrir como interpretar os resultados de uma ANOVA com dois fatores, avaliar o ajuste do modelo e utilizar um procedimento de comparações múltiplas. Tudo isso utilizando os dados referentes ao estudo sobre o medicamento.

## *Interpretando os resultados: Números e gráficos*

O exemplo do estudo em questão envolve quatro pessoas em cada combinação de tratamento com três possíveis níveis de dosagem (10mg, 20mg e 30mg ao dia) e duas possíveis frequências para a administração do medicamento (uma vez ao dia e duas vezes ao dia). O tamanho total da amostra é  $4 * 3 * 2 = 24$ . Fiz cinco diferentes conjuntos de dados em que as análises representam cada um dos cinco cenários ilustrados na Figura 11-1. As tabelas ANOVA criadas pelo Minitab estão na Figura 11-2.

Observe que cada tabela ANOVA na Figura 11-2 mostra que os graus de liberdade para a dosagem são  $3 - 1 = 2$ ; os graus de liberdade para o número de vezes que o medicamento é tomado ao dia são  $2 - 1 = 1$ ; os graus de liberdade para o termo de interação são  $(3 - 1) * (2 - 1) = 2$ ; os graus de liberdade para o total são  $3 * 2 * 4 - 1 = 23$ ; e os graus de liberdade para o resíduo são  $3 * 2 * (4 - 1) = 18$ .

A ordem dos gráficos da Figura 11-1 e a das tabelas ANOVA na Figura 11-2 não é a mesma. Você consegue adivinhar qual corresponde a qual? (Prometo lhe dar a resposta, basta continuar lendo.)

Veja aqui como os gráficos da Figura 11-1 se correspondem com a saída da Figura 11-2:

- ✓ Na tabela ANOVA da Figura 11-2, o termo de interação não é significativo (valor- $p = 0,526$ ), sendo assim, os efeitos principais podem ser estudados. Os valores- $p$  para o Fator A (dosagem) e para o Fator B (frequência de administração) são 0,000 e 0,001, indicando que, tanto o Fator A quanto o B, são significativos; o que corresponde ao gráfico da Figura 11-1a.
- ✓ Na figura 11-2b, o valor- $p$  para a interação é significativo (valor- $p = 0,000$ ), então, você não deve analisar os efeitos principais dos Fatores A e B (ou seja, não observe seus valores- $p$ ). Essa tabela, então, representa a situação ilustrada pela Figura 11-1e.
- ✓ A Figura 11-2c mostra que nada é significativo. O valor- $p$  para o termo de interação é 0,513; valores- $p$  para os efeitos principais dos fatores A (dosagem) e B

(frequência da administração) são 0,926 e 0,416, respectivamente. Esses resultados coincidem com a Figura 11-1d.

- ✓ A Figura 11-2d corresponde à Figura 11-1b. Ela demonstra a ausência do efeito de interação (valor- $p = 0,899$ ); a dosagem (o Fator A) é significativa (valor- $p = 0,000$ ), e a frequência de administração (o Fator B) não é (valor- $p = 0,207$ ).
- ✓ A Figura 11-2e corresponde à Figura 11-1c. O termo de interação, dosagem \* vezes ao dia não é significativo (valor- $p = 0,855$ ); o Fator B (vezes ao dia) é significativo, tendo um valor- $p$  de 0,000, mas o Fator A (nível da dosagem) não é (seu valor- $p = 0,855$ ).

Two-way ANOVA: BP versus Dosage, Times						
Source	DF	SS	MS	F	P	
Dosage	2	56.3333	28.1667	112.67	0.000	
Times	1	4.1667	4.1667	16.67	0.001	
Interaction	2	0.3333	0.1667	0.67	0.526	
Error	18	4.5000	0.2500			
Total	23	65.3333				
S = 0.5      R-Sq = 93.11%      R-Sq(adj) = 91.20%						

Two-way ANOVA: BP versus Dosage, Times						
Source	DF	SS	MS	F	P	
Dosage	2	0.0833	0.04167	0.16	0.855	
Times	1	0.3750	0.37500	1.42	0.249	
Interaction	2	16.7500	8.37500	31.74	0.000	
Error	18	4.7500	0.26389			
Total	23	21.9583				
S = 0.5137      R-Sq = 78.37%      R-Sq(adj) = 72.36%						

Two-way ANOVA: BP versus Dosage, Times						
Source	DF	SS	MS	F	P	
Dosage	2	0.0833	0.041667	0.08	0.926	
Times	1	0.3750	0.375000	0.69	0.416	
Interaction	2	0.7500	0.375000	0.69	0.513	
Error	18	9.7500	0.541667			
Total	23	10.9583				
S = 0.7360      R-Sq = 11.03%      R-Sq(adj) = 0.00%						

Two-way ANOVA: BP versus Dosage, Times						
Source	DF	SS	MS	F	P	
Dosage	2	36.7500	18.3750	47.25	0.000	
Times	1	0.6667	0.6667	1.71	0.207	
Interaction	2	0.0833	0.0417	0.11	0.899	
Error	18	7.0000	0.3889			
Total	23	44.5000				
S = 7.6236      R-Sq = 84.27%      R-Sq(adj) = 79.90%						

Two-way ANOVA: BP versus Dosage, Times						
Source	DF	SS	MS	F	P	
Dosage	2	0.0833	0.0417	0.16	0.855	
Times	1	12.0417	12.0417	45.63	0.000	
Interaction	2	0.0833	0.0417	0.16	0.855	
Error	18	4.7500	0.2639			
Total	23	16.9583				
S = 0.5137      R-Sq = 71.99%      R-Sq(adj) = 64.21%						

Figura 11-2 Tabelas ANOVA para gráficos de interação da Figura 11-1.

Para avaliar o ajuste dos modelos ANOVA com dois fatores, você pode usar o  $R^2$  ajustado (veja o Capítulo 6). Quanto maior o  $R^2$  ajustado, melhor (o máximo é 100% ou 1,00). Observe que todas as tabelas ANOVA na Figura 11-2 mostram um  $R^2$  ajustado bastante elevado, exceto a Figura 11-2c, onde nenhum dos termos foi significativo.

### ***Comparações múltiplas***

Caso você encontre um efeito de interação estatisticamente significativo, realize comparações múltiplas para ver quais combinações entre os Fatores A e B geram diferentes resultados na resposta. As ideias aqui são as mesmas para as comparações múltiplas abordadas no Capítulo 10, exceto pelo fato de que os testes podem ser realizados para todas as interações  $i * j$ .

Para realizar um procedimento de comparações múltiplas para uma ANOVA com dois fatores usando o Minitab, insira suas respostas (dados) em Column 1 (C1), os níveis do Fator A em Column 2 (C2) e os seus níveis do Fator B em Column 3 (C3). Clique em Stat>ANOVA>General Linear Model. No campo Responses, insira a variável da Column 1. No Modelo, digite C1 <espaço> C2 <espaço> C1 \* C2 (para os efeitos principais e de interação, respectivamente; aqui, <espaço> significa deixar um espaço). Clique em Comparisons. Em Terms, insira as Columns 2 e 3. Assinale o Method que deseja usar para suas comparações múltiplas (consulte o Capítulo 10) e clique em OK.



# *O Branco Fica Mais Branco na Água Quente? Mais um Caso para a ANOVA com Dois Fatores*

A ANOVA com dois fatores é usada quando se quer comparar as médias de  $n$  populações que são classificadas de acordo com duas variáveis categóricas (fatores). Por exemplo, suponha que você queira ver como quatro marcas de sabão em pó (Marcas A, B, C, D) e a temperatura da água (1 = frio, 2 = morna, 3 = quente) trabalham juntas para influenciar a brancura das camisetas que estão sendo lavadas (os institutos para teste de produtos podem usar essa informação, bem como os fabricantes de sabão em pó, para investigar ou anunciar como um produto está em relação a seus concorrentes).

Já que essa questão envolve dois fatores diferentes e seus efeitos sobre uma variável numérica (quantitativa), você sabe que precisa usar uma ANOVA com dois fatores. Não se pode presumir que a temperatura da água afeta a brancura da roupa da mesma forma para as quatro marcas, por isso, é preciso incluir um efeito de interação entre marca e temperatura no modelo ANOVA com dois fatores. Como existem quatro possíveis tipos (ou níveis) para a marca de sabão em pó e a temperatura da água possui três possíveis valores (ou níveis), tem-se  $4 * 3 = 12$  combinações diferentes para serem analisadas em relação a como a marca e a temperatura da água interagem. Tais combinações são: Marca A em água fria, Marca A em água morna, Marca A em água quente; Marca B em água fria, Marca B em água morna, Marca B em água quente, e assim por diante.

O resultado do modelo ANOVA com dois fatores se parece com este:  $y = b_i + w_j + bw_{ij} + e$ , onde  $b$  representa a marca de sabão em pó,  $w$  representa a temperatura da água,  $y$  representa a brancura das roupas após a lavagem, e  $bw_{ij}$  representa a interação da marca  $i$  ( $i = A, B, C, D$ ) e a temperatura  $j$  da água ( $j = 1, 2, 3$ ). (Note que  $e$  representa a quantidade de variação nos valores de  $y$  (brancura) que não é explicada nem pela marca nem pela temperatura).

Suponha que você decida executar a experiência cinco vezes para cada uma das 12 combinações, o que significa 60 observações (ou seja, lavar 60 camisetas — isso é o que chamo de um trabalho sujo, mas alguém tem que fazê-lo). Os resultados da ANOVA com dois fatores são mostrados na Figura 11-3.

ANOVA Table: Clothing Example					
Source	DF	SS	MS	F	P
Brand	3	22.983	7.6611	20.89	0.000
Water	2	1.433	0.7167	1.95	0.153
Interaction	6	308.167	51.3611	140.08	0.000
Error	48	17.600	0.3667		
Total	59	350.183			
S = 0.6055      R-Sq = 94.97%      R-Sq(adj) = 93.82%					

**Figura 11-3:** Tabela ANOVA para o exemplo das roupas.

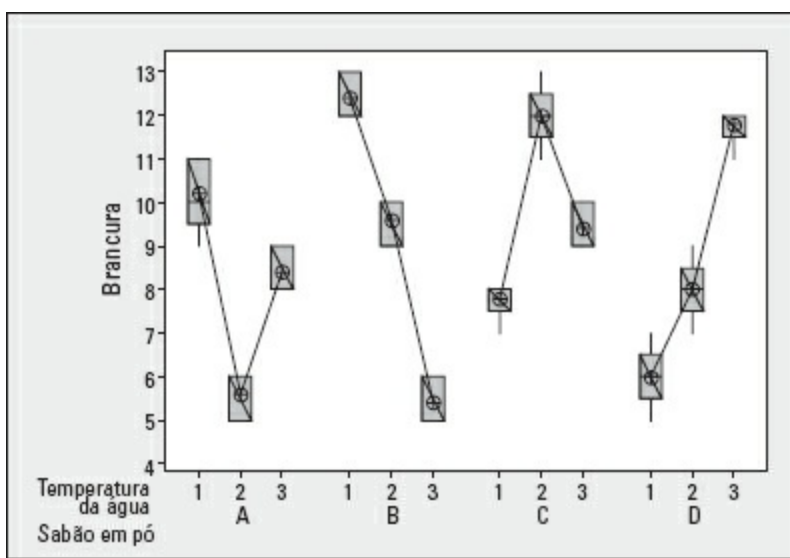
Observe que os graus de liberdade (gl) para marca, água, interação, erro (resíduos) e total foram obtidos a partir dos seguintes cálculos:

- ✓ gl para a marca:  $4 - 1 = 3$
- ✓ gl para temperatura da água:  $3 - 1 = 2$
- ✓ gl para o termo de interação:  $(4 - 1) * (3 - 1) = 6$
- ✓ gl para erro (resíduo):  $60 - (4 * 3) = 48$
- ✓ gl para o total:  $n - 1 = 60 - 1 = 59$

Analisando a tabela ANOVA na Figura 11-3, você pode ver que o modelo se ajusta muito bem aos dados, com  $R^2$  ajustado igual a 93,82%. O termo de interação (marca do sabão em pó interagindo com a temperatura da água) é significativo, com valor- $p$  de 0,000. Isso significa que você não deve analisar o efeito da marca de sabão em pó ou da temperatura da água de forma isolada. Uma marca de sabão em pó nem sempre é a melhor, como também uma temperatura da água nem sempre é a melhor; é a combinação entre elas que resulta em diferentes efeitos.

Sua próxima pergunta poderia ser: qual é a melhor combinação entre marca de sabão em pó e temperatura da água? Para responder a essa pergunta, realizei comparações múltiplas para as médias de todas as 12 combinações (para fazer isso, segui as instruções do Minitab na seção anterior). Felizmente, o teste de Tukey me dá uma taxa de erro global de apenas 5%, assim, a realização de tantos testes não me leva a tirar diversas conclusões erradas.

Devido ao elevado número de combinações a serem comparadas, faz sentido que todos os resultados na saída de Tukey sejam um pouco complicados. Em vez disso, optei primeiro por fazer boxplots dos dados de cada combinação entre marca e temperatura da água para me ajudar a visualizar o que estava acontecendo. Os resultados de minha boxplots estão na Figura 11-4.



**Figura 11-4:** Boxplots mostrando como a marca do sabão em pó e a temperatura da água interagem para influenciar a brancura das roupas.

Para criar um conjunto de boxplots para os dados de cada uma das combinações em uma ANOVA com dois fatores, em primeiro lugar, faça com que o Minitab realize uma ANOVA com dois fatores (as instruções estão na seção anterior, “Em busca das somas de quadrados”). Na mesma janela do Minitab para a ANOVA com dois fatores, clique em Graphs, e uma nova janela se abrirá. Clique em Boxplots of Data e, depois, em OK. Por último, clique em OK para executar a análise e, com ela, obter os boxplots.

A Figura 11-4 mostra quatro grupos de três caixas conectadas; cada grupo de três representa os dados de uma marca de sabão em pó testada em cada uma das três temperaturas de água (1 = fria, 2 = morna e 3 = quente). Por exemplo, o primeiro grupo de três mostra os dados da Marca A testada em cada uma das três temperaturas da água, 1, 2 e 3, respectivamente. Cada boxplot mostra os níveis de brancura para as cinco camisas lavadas com uma combinação de temperatura da água e sabão em pó.

Analisando esses gráficos, você pode ver que cada sabão em pó reage de forma diferente conforme as diferentes temperaturas da água. Por exemplo, a Marca A se sai melhor em água fria (nível 1), mas, em água morna (nível 2), se sai muito mal, enquanto que a Marca C faz justamente o oposto, alcançando as pontuações mais elevadas quando em água morna e as mais baixas quando em água fria. Cada sabão em pó se sai melhor/pior em uma combinação diferente de temperatura da água. Sendo assim, você pode realmente perceber por que o termo de interação nesse modelo é significativo!

Mas, afinal, qual combinação entre temperatura da água e sabão em pó se sai melhor? Se você olhar os gráficos, a Marca B em água fria realmente parece ser uma boa combinação, assim como a Marca C em água morna, seguida bem de perto pela Marca D em água quente. E é aqui que as comparações múltiplas de Tukey entram em cena.

Ao realizar comparações múltiplas para todas as 12 combinações entre temperatura da água e sabão em pó, é possível confirmar que as três primeiras combinações são todas

significativamente superiores do que todas as outras (pois suas médias amostrais foram maiores e suas diferenças em relação a todas as outras médias tiveram valores- $p$  inferiores a 0,05). Porém, as três primeiras não podem ser distinguidas umas das outras (pois todos os valores- $p$  para as diferenças entre elas ultrapassaram 0,05). O teste de Tukey também nos informa que as três piores combinações são: a Marca A em água morna, a Marca B em água quente e Marca D em água fria. E todas estão juntas no fundo do poço (suas médias são significativamente menores do que as médias de todas as demais, porém, não podem ser distinguidas umas das outras). Assim, nenhuma combinação pode ser considerada melhor do que a outra.

Você pode pensar nas muitas outras comparações que poderia fazer a partir daqui a fim de classificar as outras combinações em algum tipo de ordem, mas acho que, para esse caso, a classificação em melhor e pior é a mais interessante. Como os comentários sobre o que as celebridades usam nas noites de premiação (seja qual for a roupa que elas usaram, esperamos que seus estatísticos lhes digam qual a marca de sabão em pó e temperatura da água usar na hora de lavá-las).

# Capítulo 12

## Regressão e ANOVA: Uma Relação Inesperada!

---

### *Neste Capítulo*

- ▶ Reescrevendo uma reta de regressão como um modelo ANOVA
  - ▶ Ligando as equações de regressão à tabela ANOVA
- 

**E**ntão, você estava tranquilamente passeando por seu curso de Estatística II, percorrendo os caminhos da regressão (onde estimava  $y$  usando uma ou mais variáveis  $x$ , veja o Capítulo 4), até que apareceu um novo tópico, ANOVA, sigla para *análise de variância*, técnica que se refere à comparação das médias de várias populações (consulte o Capítulo 9), que, no final, você também tirou de letra. Mas, espere um pouco, agora seu professor está começando a falar sobre como a ANOVA se relaciona com a regressão e, de repente, tudo começou a ficar fora de controle. Como conciliar duas técnicas que parecem ser tão diferentes? Este capítulo vai tratar justamente disso.

Pense neste capítulo como uma ponte de ligação entre a regressão linear simples e a ANOVA, ponte esta que vai permitir que você continue seu passeio tranquilamente e responder a todas as perguntas que o professor lhe fizer. Tenha em mente que você não vai aplicar essas duas técnicas neste capítulo (para encontrar essa informação, consulte os Capítulos 4 e 9). O objetivo deste capítulo é determinar e descrever a relação entre regressão e ANOVA.



# *Vendo a Regressão Através dos Olhos da Variação*

Todos os modelos estatísticos básicos tentam explicar porque os diferentes resultados ( $y$ ) são o que são. Eles tentam descobrir quais fatores ou variáveis explicativas ( $x$ ) podem ajudar a explicar a variabilidade nesses  $y$ 's. Nesta seção, você começa apenas com os valores de  $y$  e vê como sua variabilidade desempenha um papel importante no modelo de regressão. Este é o primeiro passo para a aplicação da ANOVA ao modelo de regressão.



Independentemente da variável  $y$  em que está interessado, sempre haverá variabilidade em seus valores. Se quiser prever o comprimento de um peixe, por exemplo, deverá saber que existem peixes com muitos comprimentos diferentes (indicando uma grande variabilidade). Mesmo que você coloque todos os peixes da mesma idade e da mesma espécie juntos, ainda terá alguma variação em seus comprimentos (menos do que antes, mas ainda terá). O primeiro passo para entender os princípios básicos da regressão e da ANOVA é entender que a variabilidade dos valores de  $y$  é esperada e que seu trabalho é tentar descobrir o que pode explicar a maior parte dela.

## *Localizando a variabilidade e encontrando uma “x-plicação”*

Tanto a regressão quanto a ANOVA trabalham para conseguir explicar a variabilidade na variável  $y$  usando uma variável  $x$ . Depois de coletar os dados, você pode encontrar o desvio padrão da variável  $y$  para ter uma noção de quanto os dados variam dentro da amostra. A partir daí, colete dados para uma variável  $x$  e determine o quanto ela contribui para explicar essa variabilidade.

Suponha que você tenha notado que as pessoas passam diferentes quantidades de horas na Internet e decide explorar o motivo para que isso ocorra. Para tal, você começa com a coleta de uma pequena amostra de 20 pessoas e registra quantas horas por mês elas passam na web. Os resultados (em horas) são 20, 20, 22, 39, 40, 19, 20, 32, 33, 29, 24, 26, 30, 46, 37, 26, 45, 15, 24 e 31. A primeira coisa a notar sobre esses dados é sua grande variabilidade. O desvio padrão (distância média entre os valores dos dados e suas médias) desse conjunto de dados é 8,93 horas, considerado bastante grande, dado o tamanho dos números no conjunto de dados.

Assim, você descobre que os valores de  $y$  — a quantidade de horas que alguém usa a Internet — possui uma grande variabilidade. Mas, afinal, o que pode ajudar a explicar isso? Sabemos que parte da variabilidade se deve ao acaso. Porém, você suspeita que uma variável lá fora (vamos chamá-la de  $x$ ) tem alguma ligação com a variável  $y$ , e que a variável  $x$  pode ajudá-lo a entender melhor esse intervalo, aparentemente muito amplo, para os valores de  $y$ .

Suponha que você imagine que o número de anos de estudo poderia estar relacionado ao uso da Internet. Neste caso, a variável explicativa (variável de entrada,  $x$ ) representa os anos de estudo (escolaridade) e será usada para tentar estimar  $y$ , o número de horas que

uma pessoa passa na Internet durante um mês. Você, então, pergunta a escolaridade de uma amostra aleatória de 250 usuários da Internet (assim,  $n = 250$ ). Confira as dez primeiras observações de seu conjunto de dados contendo os pares  $(x, y)$  na Tabela 12-1. Se houver uma ligação significativa entre os valores de  $x$  e de  $y$ , então, pode-se dizer que  $x$  ajuda a explicar uma parte da variabilidade em  $y$ . Se  $x$  explicar uma parte suficiente dessa variabilidade, você pode colocá-lo em um modelo de regressão simples e usá-lo para estimar  $y$ .

**Tabela 12-1**                      **Primeiras Dez Observações do Exemplo Escolaridade e Uso da Internet**

<i>Anos de Estudo</i>	<i>Horas Gastas na Internet (em um mês)</i>
15	41
15	32
11	33
10	42
10	28
10	21
10	17
10	14
9	18
9	14

## ***Chegando aos resultados com a regressão***

Depois de ter escolhido uma possível variável  $x$ , colete os pares de dados  $(x, y)$  em uma amostra aleatória de indivíduos da população e procure uma possível relação linear entre elas. Analisando o pequeno trecho de 10 dos 250 dados definidos na Tabela 12-1, você começa a notar que pode haver um padrão entre a escolaridade e o uso da internet. Parece que, à medida que a escolaridade aumenta, o uso da internet também cresce.

Para se aprofundar, construa um diagrama de dispersão dos dados e calcule a correlação ( $r$ ). Se os dados seguirem uma reta (como a mostrada no diagrama de dispersão), siga em frente e realize uma regressão linear simples da variável resposta  $y$  com base na variável  $x$ . O valor- $p$  da variável  $x$  na análise de regressão linear simples indica se a variável  $x$  faz ou não um bom trabalho para a previsão de  $y$ . (Para mais detalhes sobre regressão linear simples, consulte o Capítulo 4.)

Para fazer uma regressão linear simples usando o Minitab, insira seus dados em duas colunas: a primeira coluna para a variável  $x$  e a segunda coluna para a variável  $y$  (como na Tabela 12-1). Clique em Stat>Regression>Regression. Clique na variável  $y$  no campo à esquerda; ao fazer isso, a variável  $y$  aparece no campo Response ao lado direito. Clique na



variável  $x$  no campo à esquerda; ao fazer isso, a variável  $x$  aparece no campo Predictor à direita. Clique em OK e sua análise de regressão é realizada. Como parte de todas as análises de regressão, o Minitab também fornece os resultados correspondentes à ANOVA, encontrados na parte inferior da saída.

A saída da regressão linear simples que o Minitab oferece para o exemplo da relação escolaridade e Internet está na Figura 12-1. (Observe a saída da ANOVA na parte inferior; você poderá ver a ligação entre elas na seção “Regressão e ANOVA: o encontro dos modelos”).

Regression Analysis: Internet versus Education					
The regression equation is					
Internet = -8.29 + 3.15 Education					
Predictor	Coef	SE Coef	T	P	
Constant	-8.290	2.665	-3.11	0.002	
Education	3.1460	0.2387	13.18	0.000	
S = 7.23134      R-Sq = 41.2%      R-Sq(adj) = 41.0%					
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	9085.6	9085.6	173.75	0.000
Residual Error	248	12968.5	52.3		
Total	249	22054.0			

**Figura 12-1:** Saída para a regressão linear simples aplicada aos dados do exemplo Internet e escolaridade

Observando a Figura 12-1, você vê que o valor- $p$  na linha *Education* é 0,000, o que significa que o valor- $p$  é inferior a 0,001. Portanto, a relação entre os anos de estudo e o uso da Internet é estatisticamente significativa. Um diagrama de dispersão dos dados (não mostrado aqui) também indica que os dados parecem possuir uma relação linear positiva, sendo assim, conforme o número de anos de escolaridade aumenta, o uso da Internet também tende a aumentar (em média).

## ***Avaliando o ajuste do modelo de regressão***

Antes de seguir em frente e utilizar um modelo de regressão a fim de fazer previsões para  $y$  com base em uma variável  $x$ , primeiro você deve avaliar o ajuste de seu modelo. Para isso, use um diagrama de dispersão e a correlação, ou  $R^2$ .

### ***Usando o diagrama de dispersão e a correlação***

Uma forma de se ter uma ideia da qualidade do ajuste de seu modelo de regressão é usar um *diagrama de dispersão*, gráfico que mostra todos os pares de dados dispostos no plano  $x$   $y$ . Use o diagrama de dispersão para ver se os dados parecem seguir o padrão de uma reta. Se parecer que os dados seguem o padrão de uma reta (ou mesmo algo próximo a isso — qualquer coisa, menos uma curva ou uma dispersão de pontos que não siga padrão

algum), calcule a *correlação*,  $r$ , para ver o quão forte é a relação linear entre  $x$  e  $y$ . Quanto mais próximo de 1 ou  $-1$  estiver  $r$ , mais forte será a relação; quanto mais próximo ele estiver de zero, mais fraca será a relação. O Minitab pode construir os gráficos de dispersão e encontrar as correlações por você; consulte o Capítulo 4 para mais informações a respeito de regressão linear simples, inclusive sobre a construção de um diagrama de dispersão e como encontrar o valor de  $r$ .

Se os dados não apresentarem uma correlação significativa e/ou o diagrama de dispersão não parecer linear, interrompa a análise; você não pode continuar tentando encontrar uma reta que se ajuste a uma relação que não existe.

## Usando $R^2$

A forma mais comum de avaliar não só o ajuste de um modelo de regressão linear simples, mas também o de muitos outros modelos é usando o  $R^2$ , também conhecido como *coeficiente de determinação*. (Por exemplo, você pode aplicar esse método aos modelos de regressão múltipla, não linear e logística, abordados nos Capítulos 5, 7 e 8, apenas para citar alguns.) Na regressão linear simples, o valor de  $R^2$  (indicado pelo Minitab e pelos estatísticos como um  $R$  maiúsculo ao quadrado) é igual ao quadrado do coeficiente de correlação de Pearson,  $r$  (indicado pelo Minitab e pelos estatísticos como um  $r$  minúsculo). Em todas as outras situações, o  $R^2$  fornece uma medida mais geral do ajuste do modelo. (Note que  $r$  mede apenas o ajuste de uma relação linear entre uma variável  $x$  e uma variável  $y$ ; consulte o Capítulo 4.) Uma estatística ainda melhor, o  $R^2$  *ajustado*, modifica o  $R^2$  para compensar o número de variáveis no modelo. (Veja no Capítulo 6 mais informações sobre  $R^2$ , seu uso e sua interpretação.)

O valor de  $R^2$  *ajustado* para o modelo que usa a escolaridade para estimar o uso da Internet (veja a Figura 12-1) é igual a 41%. Tal valor reflete a porcentagem de variabilidade no uso da Internet que pode ser explicada pela escolaridade (em anos) de uma pessoa. Esse número não está próximo de um, mas note que  $r$ , a raiz quadrada de 41%, é 0,64, que, no caso da regressão linear, indica uma relação moderada.

Essa evidência lhe dá sinal verde para a utilização dos resultados da análise de regressão na estimativa do número de horas de utilização da internet (durante um mês) através do uso dos anos de escolaridade. A equação de regressão que aparece na parte superior da saída ilustrada na Figura 12-1 é  $\text{Uso da internet} = -8,29 + 3,15 * \text{escolaridade (em anos)}$ . Então, se você, por exemplo, tiver estudado 16 anos, a estimativa para seu uso da internet é  $-8,29 + 3,15 * 16 = 42,11$ , ou cerca de 42 horas por mês (cerca de 10h 30min por semana).

Mas, espere! Olhe novamente a Figura 12-1 e preste atenção na parte inferior. Eu não fiz nada de especial para obter essa informação na saída do Minitab, mas veja quem está lá, a tabela ANOVA. Parece um peixe fora d'água, não é? A próxima seção vai juntar as duas e lhe mostrar como uma tabela ANOVA pode descrever os resultados da regressão (mas de forma diferente, é claro).



# ***Regressão e ANOVA: O Encontro dos Modelos***

Depois de decompor a saída da regressão, o próximo passo para a compreensão da relação entre regressão e ANOVA é aplicar a soma de quadrados da ANOVA à regressão (algo que normalmente não é feito em uma análise de regressão). Antes de começar, imagine que esse processo seja como assistir a um filme em 3-D, quando você tem que usar óculos especiais para ver todos os efeitos!

Nesta seção, você vai ver a soma de quadrados da ANOVA sendo aplicada à regressão e como calcular os graus de liberdade. Também vai construir uma tabela ANOVA para a regressão e descobrir como o teste- $t$  para um coeficiente de regressão se relaciona ao teste- $F$  da ANOVA.

## ***Comparando as somas de quadrados***

A *soma de quadrados* é um termo da ANOVA (veja o Capítulo 9), mas que, com certeza, você não usaria no caso de regressão (veja Capítulo 4). No entanto, você pode decompor os dois tipos de modelos em somas de quadrados, e é essa semelhança que nos leva à verdadeira ligação entre a ANOVA e a regressão.

Se usarmos um procedimento passo a passo, primeiro você deve dividir a variabilidade na variável  $y$  usando as fórmulas para a soma de quadrados da ANOVA (somas de quadrados para o total, o tratamento e os resíduos). Depois, deve encontrar as somas de quadrados para a regressão — este é o segredo do processo, os dois procedimentos são comparados através de suas somas de quadrados, e esta seção vai explicar a você como tal comparação é feita.

## ***Decompondo a variabilidade com a SQTO, SQE e SQT.***

A ANOVA trata da decomposição da variabilidade total nos valores de  $y$  em somas de quadrados (encontre toda a informação que precisar sobre ANOVA com um fator no Capítulo 9). A ideia fundamental é que  $SQTO = SQT + SQE$ , onde  $SQTO$  é a variabilidade total nos valores de  $y$ ,  $SQT$  mede a variabilidade explicada pelo modelo (também conhecida como tratamento ou variável  $x$ , nesse caso) e  $SQE$  mede a variabilidade devido ao erro experimental (o que sobra depois que o modelo é ajustado).

Seguem-se as fórmulas correspondentes para  $SQTO$ ,  $SQE$  e  $SQT$ , onde  $\bar{y}$  a média dos valores de  $y$ ,  $y_i$  é cada valor observado de  $y$ , e  $\hat{y}_i$  é cada valor previsto para  $y$  a partir do modelo ANOVA:

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

$$SQE = \sum (y_i - \hat{y}_i)^2$$

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

Use essas fórmulas para calcular a soma de quadrados para a ANOVA (o Minitab faz isso



por você ao realizar a ANOVA). Guarde os valores para SQTO, SQE e SQT, pois vai compará-los aos resultados da regressão.

## ***Encontrando a soma de quadrados para a regressão***

Na regressão, para medir os desvios nos valores de  $y$ , subtraia cada  $y_i$  de sua média  $\bar{y}$ . Eleve cada resultado ao quadrado e os some para chegar ao valor de SQTO. Em seguida, pegue os resíduos, que representam a diferença entre cada  $y_i$  e seu valor estimado a partir do modelo,  $\bar{y}_i$ , eleve-os ao quadrado e os  $y$  some para chegar à fórmula da SQE.

Depois de calcular SQTO e SQE, é necessário fazer a ponte entre eles — ou seja, você precisa de uma fórmula que una a variabilidade nos valores  $y_i$  (SQTO) e a variabilidade dos resíduos, gerada depois que a reta de regressão é ajustada (SQE). Essa ponte é chamada de *soma de quadrados da regressão*, ou SQR (equivalente a SQT na ANOVA). Na regressão,  $\bar{y}$  representa o valor previsto para  $y_i$  a partir do modelo de regressão. Esses são os valores da reta de regressão. Para avaliar o quanto essa reta de regressão ajuda na previsão dos valores de  $y$ , você deve compará-la ao modelo que você teria obtido sem qualquer variável  $x$ .

Sem qualquer outra informação, a única coisa que você pode fazer para prever  $y$  é analisar sua média,  $\bar{y}$ . Então, SQR compara o valor previsto a partir da reta de regressão ao valor previsto pela reta contínua (a média dos  $y$ 's), subtraindo-os. O resultado é  $(\hat{y}_i - \bar{y})$ . Eleve cada resultado ao quadrado e os some. Assim, você obtém a fórmula para SQR que é igual à fórmula da SQT na ANOVA. *Voilà!*

Em vez de chamar a soma de quadrados do modelo de regressão de SQT como é feito na análise de variância, os estatísticos a chamam de SQR sigla para *soma de quadrados da regressão*. Considere a SQR equivalente à SQT da ANOVA. Mas você precisa saber a diferença, pois a saída do Minitab fornece as somas de quadrados para o modelo de regressão como SSR (em português, SQR), e não SST (em português, SQT).

Para resumir as somas de quadrados aplicadas à análise de regressão, temos  $SQTO = SQE + SQR$  onde

- ✓ **SQTO** mede a variabilidade dos valores de  $y$  observados em torno de sua média. Esse valor representa a variação dos valores de  $y$ .
- ✓ **SQE** representa a variabilidade entre os valores previstos para  $y$  (os valores na reta) e os valores de  $y$  observados. A SQE representa a variabilidade deixada depois que a reta é ajustada aos dados.
- ✓ **SQR** representa a variabilidade entre os valores previstos para  $y$  (os valores na reta) e os valores de  $y$  observados. A SQR é a soma de quadrados devido ao modelo de regressão (reta) em si.

O Minitab calcula todas as somas de quadrados para você como parte da análise de




regressão. Você pode ver esses cálculos na seção “Levando a regressão até a tabela ANOVA”.

## ***Dividindo os graus de liberdade***

Na ANOVA, você testa um modelo para a média do tratamento (população) usando um teste- $F$ , que é  $\frac{MQT}{MQE}$ . Para obter MQT (a média da soma de quadrados para o tratamento ou grupo), divida SQT (a soma de quadrados do tratamento) por seus graus de liberdade. Faça o mesmo para obter a MQE (ou seja, divida a soma de quadrados dos resíduos, SQE, por seus graus de liberdade). As perguntas agora são: o que esses graus de liberdade representam e como se relacionam com a regressão?

### ***Graus de liberdade na ANOVA***

Na ANOVA, o grau de liberdade para SQT é  $n - 1$ , que representa o tamanho amostral menos um. Na fórmula para SQT,  $\sum (y_i - \bar{y})^2$ , vemos  $n$  valores de  $y$  observados menos uma média. De modo geral, é daí que vem o  $n - 1$ .

 Observe que, se você dividir SQT por  $n - 1$ , terá  $\frac{\sum (y_i - \bar{y})^2}{n-1}$ , a variância nos valores de  $y$ . Esse cálculo faz sentido, pois a variância mede a variabilidade *total* nos valores de  $y$ .

### ***Graus de liberdade na regressão***

O valor para os graus de liberdade para a SQT na ANOVA é igual ao número de tratamentos menos um. Como a ideia de graus de liberdade se relaciona com a regressão? O número de tratamentos na regressão é equivalente ao número de parâmetros em um modelo (um *parâmetro* vem a ser uma constante desconhecida para a qual se está tentando fazer uma estimativa).

Quando você testa um modelo, sempre estará comparando-o a um modelo diferente (mais simples) para ver se ele se ajusta melhor aos dados. Na regressão linear, você compara sua reta de regressão  $y = a + bx$  à reta horizontal  $y = \bar{y}$ . Esse segundo modelo, mais simples, usa apenas a média de  $y$  para prevê-lo o tempo todo, independentemente de qualquer  $x$ . Na reta de regressão, tem-se dois coeficientes: um para estimar o parâmetro para a intersecção  $y$  ( $a$ ) e um para estimar o parâmetro para a inclinação ( $b$ ) no modelo. No modelo mais simples, tem-se apenas um parâmetro: o valor da média. O grau de liberdade para a SQR na regressão linear simples é a diferença entre o número de parâmetros dos dois modelos:  $2 - 1 = 1$ .

O grau de liberdade para SQE na ANOVA é  $n - k$ . Na fórmula para a SQE,  $\sum (\hat{y}_i - \bar{y})^2$ , vemos que existem  $n$  valores de  $y$  previstos e  $k$  é o número de tratamentos no modelo. Na regressão, o número de parâmetros do modelo é  $k = 2$  (a inclinação e a intersecção  $y$ ). Então, quando se trata de regressão, tem-se  $n - 2$  graus de liberdade associados à SQE.





Reunindo tudo isso, os graus de liberdade para a regressão devem se enquadrar na equação  $SQTO = SQR + SQE$ . Os graus de liberdade correspondentes a essa equação são  $(n - 1) = (2 - 1) + (n - 2)$ , o que é verdade se você fizer as contas. Assim, todos os graus de liberdade para a regressão, utilizando a abordagem ANOVA, devem bater. Ufa!

Na Figura 12-1, os graus de liberdade para cada soma de quadrados estão indicados na coluna DF da saída ANOVA. Assim, vemos que SQR tem  $2 - 1 = 1$  grau de liberdade, SQE tem  $250 - 2 = 248$  graus de liberdade (pois  $n = 250$  observações no conjunto de dados e  $k = 2$  e, ao fazer  $n - k$ , obtemos os graus de liberdade para SQE). Os graus de liberdade para SQTO são  $250 - 1 = 249$ .

## ***Levando a regressão até a tabela ANOVA***

Na Anova, você testa seu modelo  $H_0$ : todas  $k$  médias populacionais são iguais versus  $H_a$ : pelo menos duas médias populacionais são diferentes, usando um teste- $F$ . A estatística do teste- $F$  é formada através da relação entre a soma de quadrados do tratamento e a soma de quadrados dos resíduos. Para fazer isso, divida SQE e SQT por seus graus de liberdade ( $n - k$  e  $k - 1$ , respectivamente, onde  $n$  é o tamanho amostral e  $k$  é o número de tratamentos) a fim de chegar à média quadrática do resíduo (MQE) e à média quadrática do tratamento (MQT). De forma geral, a MQT tem que ser grande quando comparada à MQR, indicando um bom ajuste do modelo. Os resultados de toda essa ginástica estatística são resumidos pelo Minitab em uma tabela chamada (muito sabiamente) de tabela ANOVA.

A tabela ANOVA mostrada na parte inferior da Figura 12-1, que ilustra o exemplo do uso da Internet, representa a tabela ANOVA obtida a partir da reta de regressão. Embaixo da coluna Source, onde você está acostumado a ver treatment, error e total, no caso da regressão, como o tratamento é a reta de regressão, você vai ver *regression* em vez de treatment. O termo de erro na análise de variância recebe o nome de *residual error*, pois, em regressão, os erros são medidos em termos de resíduos. Por último, você tem o *total*, que é a mesma coisa no mundo todo.

A coluna SS representa a soma dos quadrados para o modelo de regressão. As três somas de quadrados listadas na coluna SS são SQR (da regressão), SQE (de resíduos) e SQTO (do total). Essas somas de quadrados são calculadas com as fórmulas da seção anterior; os graus de liberdade, gl na tabela, também são encontrados através das fórmulas da seção anterior.

A coluna de MS (média quadrática, no Minitab) assume os valores da SS (você a preenche) divididos por seus respectivos graus de liberdade, assim como na ANOVA. Por exemplo, na Figura 12-1, SSE (SQE) é 12.968,5, e o valor dos graus de liberdade é 248. Divida o primeiro valor pelo segundo para obter 52,29 ou 52,3, o valor dado na tabela ANOVA para MSE (MQE).

O valor da estatística- $F$ , utilizando o método ANOVA, é:  $F = \frac{MQT}{MQE} = \frac{9085,6}{52,3} = 173,7$  no exemplo sobre a utilização da Internet, que pode ser visto na coluna cinco da parte referente à



ANOVA na Figura 12-1 (sujeita a arredondamento). O valor- $p$  da estatística- $F$  é calculado com base em uma distribuição- $F$  com  $k - 1 = 2 - 1 = 1$  e  $n - k = 250 - 2 = 248$  graus de liberdade, respectivamente. (No exemplo sobre a utilização da Internet, o valor- $p$  listado na última coluna da tabela ANOVA é 0,000, significando que o modelo de regressão realmente se ajusta.) Mas, lembre-se, em regressão, não se usa nem a estatística- $F$  nem o teste- $F$ , mas, sim, a estatística- $t$  e o teste- $t$ . (Uau...)

## ***Relacionando as estatísticas $F$ e $t$ : A última fronteira***

Em regressão, uma forma de testar se a reta de regressão é estatisticamente significativa é testando a  $H_0$ : inclinação (coeficiente angular) = 0 versus  $H_a$ : inclinação  $\neq 0$ . Para fazer isso, é preciso usar um teste- $t$  (consulte o Capítulo 3). O coeficiente angular é o coração e a alma da reta de regressão, pois descreve a principal parte da relação entre  $x$  e  $y$ . Se a inclinação da reta for igual a zero ( $H_0$  não poderá ser rejeitada), para você restará apenas  $y = a$ , uma reta horizontal, e seu modelo  $y = a + bx$  não vai lhe servir de nada.

Na ANOVA, para ver se o modelo se ajusta, você deve testar a  $H_0$ : as médias das populações são todas iguais versus a  $H_a$ : pelo menos duas das médias populacionais são diferentes. Para fazer isso, é preciso usar um teste- $F$  (ou seja, dividir a MQT pela MQR; consulte o Capítulo 9).

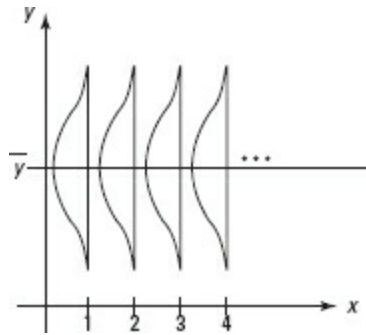
Os conjuntos de hipóteses em regressão e na ANOVA parecem ser totalmente diferentes, mas, em essência, ambos têm o mesmo objetivo: testar o ajuste de um determinado modelo. No caso da regressão, o modelo em questão é uma reta e no caso da ANOVA o modelo de interesse é um conjunto de populações (normalmente, distribuídas) com pelo menos duas médias diferentes (e mesma variância). Além disso, na ANOVA cada população é chamada de tratamento.

Mas, mais do que isso, você pode pensar desta maneira: suponha que você tenha pegado todas as populações da ANOVA e as tenha colocado lado a lado em um plano  $x y$  (veja a Figura 12-2). Se as médias dessas distribuições forem todas ligadas por uma reta contínua (representando a média dos  $y$ 's), então você não terá provas contra a  $H_0$  do teste- $F$  e, portanto, não poderá rejeitá-la — seu modelo não lhe serviu de nada (ou seja, ele não se ajusta). Essa ideia é semelhante à ideia de ajustar uma reta plana e horizontal aos valores de  $y$  na regressão; um modelo de reta com inclinação diferente de zero que também indica a ausência de uma relação entre  $x$  e  $y$ .

A boa notícia é que os estatísticos podem provar (e, assim, você não tem que fazer isso) que uma estatística- $F$  é equivalente ao quadrado da estatística- $t$  e que a distribuição- $F$  é equivalente ao quadrado de uma distribuição- $t$ , quando SSR (SQR) tem gl (graus de liberdade) =  $2 - 1 = 1$ . E, quando você tem um modelo de regressão linear simples, o valor dos graus de liberdade é exatamente igual a 1! (Note que  $F$  é sempre maior ou igual a zero, condição necessária para quando se está fazendo com que ele seja o quadrado de alguma coisa.) Então, aí está ele! A estatística- $t$  para testar o modelo de regressão é equivalente a



uma estatística- $F$  para a ANOVA quando a tabela ANOVA é construída para o modelo de regressão simples.



---

**Figura 12-2:** Ligando as médias populacionais à inclinação de uma reta.

---

Na verdade (como os professores de Estatística costumam dizer: “e esse que é o barato!”), se você olhar o valor da estatística- $t$  para o teste da inclinação da variável escolaridade na Figura 12-1, vai ver que ele é 13,18 (procure na interseção entre a linha *Education* e a coluna *T*). Note que esse valor elevado ao quadrado é igual a 173,71, a estatística- $F$  da tabela ANOVA ilustrada na Figura 12-1. A estatística- $F$  da ANOVA e o quadrado da estatística- $t$  da regressão são iguais, sujeitos apenas a um pequeno erro de arredondamento cometido pelo Minitab. (Parece mágica! Fico arrepiada só de pensar.)

## Parte IV:

# Construindo Fortes Ligações com os Testes Qui-quadrado

A 5ª Onda

Por Rich Tennant



"Sou eu ou esses '50% de satisfação' parecem um pouco inexpressivos?"



## *Nesta parte...*

**A**lguma vez, você já se perguntou se a porcentagem de M&M's de cada cor é a mesma em cada embalagem? Ou se o voto de alguém em uma eleição tem relação com seu sexo? Você já quis saber se os bancos realmente têm um método de negar empréstimos que se baseia na classificação de crédito do cliente? Esta parte vai responder a todas essas perguntas e a muitas outras utilizando a distribuição Quiquadrado. Em geral, usa-se o Qui-quadrado para testar a independência e conduzir testes de qualidade de ajuste.

# Capítulo 13

## Fazendo Associações com Tabelas de Dupla Entrada

---

### *Neste Capítulo*

- ▶ Interpretando as tabelas de dupla entrada
  - ▶ Descobrindo probabilidades e verificando a independência
  - ▶ Tomando cuidado com o Paradoxo de Simpson
- 

**A** busca por relações entre duas variáveis categóricas é um objetivo muito comum dos pesquisadores. Por exemplo, muitos estudos médicos focam em como uma característica do indivíduo pode tanto aumentar quanto diminuir suas chances de ter uma determinada doença. O pessoal do marketing faz perguntas do tipo: "Quem provavelmente compraria nossos produtos: homens ou mulheres?" Já os fanáticos por estatística esportiva se perguntam coisas do tipo: "Será que ganhar no cara e coroa no início do jogo aumenta as chances de um time vencer a partida?" (Eu acredito que sim!)

Para responder a cada uma das perguntas anteriores, você deve primeiro coletar dados (a partir de uma amostra aleatória) referentes às duas variáveis categóricas que estão sendo comparadas — vamos chamá-las de  $x$  e  $y$ . Depois, organize os dados em uma tabela que contenha linhas e colunas, mostrando a quantidade de indivíduos da amostra que aparece em cada combinação entre  $x$  e  $y$ . Por último, use as informações da tabela para realizar um teste de hipótese (chamado *teste do Qui-quadrado*). Com o teste do Qui-quadrado, é possível ver se existe uma relação entre  $x$  e  $y$  na população de onde os dados foram extraídos. No entanto, você vai precisar do maquinário do Capítulo 14 para realizar esta última etapa.

Os objetivos deste capítulo são ajudá-lo a entender o que significa a associação de duas variáveis categóricas ( $x$ ,  $y$ ) e a descobrir como usar a porcentagem para determinar se um conjunto de dados amostrais aparenta demonstrar uma relação entre  $x$  e  $y$ .

Suponha que você esteja coletando dados referentes a usuários de celulares e queira descobrir se as mulheres usam mais os celulares para uso pessoal do que os homens. Um estudo com 508 usuárias e 508 usuários selecionados aleatoriamente realizado por uma empresa de telefonia celular descobriu que as mulheres tendem a usar seus celulares mais para chamadas pessoais do que os homens (espantoso). O levantamento demonstrou que 427 das mulheres disseram usar seus celulares principalmente para conversar com amigos e familiares, enquanto apenas 325 dos homens admitiram fazê-lo.

Mas você não pode parar por aí. É preciso desmembrar essa informação, calcular algumas porcentagens e compará-las para ver o quão perto da realidade elas realmente estão. Os resultados amostrais variam de amostra para amostra e as diferenças podem aparecer por acaso.

Neste capítulo, você vai descobrir como organizar os dados obtidos das variáveis categóricas (ou seja, dados baseados em categorias em vez de medições) em uma tabela. Essa habilidade é especialmente útil quando você está à procura de relações entre duas variáveis categóricas, como a utilização de celulares para chamadas pessoais (de categoria sim ou não) e sexo (masculino ou feminino). Além disso, também vai aprender a resumir os dados para responder às suas perguntas. E, por último, vai descobrir, de uma vez por todas, o que é o tal Paradoxo de Simpson.

# Decompondo uma Tabela de Dupla Entrada

A *tabela de dupla entrada* contém linhas e colunas que ajudam a organizar os dados de variáveis categóricas do seguinte modo:

- ✓ As **linhas** representam as possíveis categorias para uma variável categórica, como homens e mulheres.
- ✓ As **colunas** representam as possíveis categorias para uma segunda variável categórica, como o uso do celular para chamadas pessoais ou não.

## Organizando dados em uma tabela de dupla entrada

Para organizar seus dados em uma tabela de dupla entrada, primeiro, configure as linhas e colunas. A Tabela 13-1 mostra a configuração dos dados para o uso do celular apresentados na introdução do capítulo.

Tabela 13-1		Tabela de Dupla Entrada para Dados de Celular	
		Ligações Pessoais: Sim	Ligações Pessoais: Não
Homens			
Mulheres			

Observe que a Tabela 13-1 possui quatro células vazias (sem contar o espaço vazio no canto superior esquerdo). Uma vez que a variável sexo tem duas opções (masculino e feminino) e a variável ligações pessoais também (sim e não), a tabela de dupla entrada resultante possui  $2 * 2 = 4$  células.

Para calcular o número de células de qualquer tabela de dupla entrada, multiplique o número de possíveis categorias para a variável da linha pelo número de possíveis categorias da variável da coluna.



## Preenchendo as células

Depois de configurar a tabela com o número adequado de linhas e colunas, é preciso preencher as células da tabela com os números adequados. O número em cada célula de uma tabela de dupla entrada é chamado de frequência da célula. Das quatro células da Tabela 13-1, a célula superior esquerda representa o número de homens que usam seus celulares para chamadas pessoais. Segundo as informações do exemplo em questão, a frequência da célula para essa célula em particular é 325. Mas você também sabe que 427 mulheres usam seus celulares para chamadas pessoais e, portanto, esse número vai para a célula inferior esquerda.

Para descobrir os números das duas células restantes basta fazer uma pequena subtração. A partir da informação dada, você sabe que o número total de usuários de celulares do sexo



masculino na pesquisa é 508. Cada homem ou usa seu celular para chamadas pessoais (grupo sim) ou não (grupo não). Uma vez que há 325 homens no grupo sim, e o total de homens é 508, temos 183 homens ( $508 - 325 = 183$ ) que dizem não usar seus celulares para fazer ligações pessoais. Assim, esse número é a frequência da célula superior direita da tabela. Por fim, uma vez que 508 é o total de mulheres que participaram da pesquisa, e 427 delas disseram usar os celulares para fazer ligações pessoais, sabemos que o restante ( $508 - 427 = 81$ ) não o faz. Portanto, 81 é a frequência da célula para a célula inferior direita. A Tabela 13-2 mostra a tabela completa para o exemplo do uso de celular, com as quatro frequências preenchidas.

**Tabela 13-2**                      **Tabela de Dupla Entrada Completa para Dados de Celular**

	<i>Ligações Pessoais: Sim</i>	<i>Ligações Pessoais: Não</i>
Homens	325	$183 = (508 - 325)$
Mulheres	427	$81 = (508 - 427)$

Só para poupar um pouco de tempo, se você tiver o número total de um grupo e o número de indivíduos que se enquadram em uma das categorias da tabela de dupla entrada, pode determinar o número de indivíduos que se enquadram na categoria restante subtraindo o número total do grupo pelo número desta categoria. Você pode concluir esse processo para cada grupo restante da tabela.

## ***Totais marginais***

Uma das características mais importantes de uma tabela de dupla entrada é o fato de que ela facilita o acesso a todos os totais pertinentes. Como todas as tabelas de dupla entrada são compostas por linhas e colunas, você já deve ter imaginado que os totais de cada linha e coluna são importantes. Além disso, também é importante saber o grande total.

Se você somar todas as frequências das células de uma única linha terá o *total marginal da linha* em questão. Mas onde é que fica esse total marginal da linha? Aposto que você já adivinhou — na margem no final da linha. Você pode encontrar os totais marginais de cada linha da tabela e colocá-los nas margens das linhas. Este grupo de totais marginais de cada linha representa o que os estatísticos chamam de *distribuição marginal* da variável linha.

A soma dos marginais totais das linhas resulta no *total geral*, o número total de indivíduos no estudo. (Os indivíduos podem ser pessoas, cidades, cães, empresas, e assim por diante, dependendo do contexto em mãos.)

Da mesma forma, se você somar todas as frequências das células de uma única coluna terá o *total marginal da coluna* em questão. Esse número fica na margem da coluna, isto é, na parte inferior. Siga esse padrão para cada coluna da tabela e você obterá a distribuição marginal da variável coluna. Novamente, a soma de todos os totais marginais de cada coluna é igual ao total geral, que sempre fica no canto inferior direito da tabela.

O total marginal das linhas, das colunas e o total geral para o exemplo do celular podem ser vistos na Tabela 13-3.

**Tabela 13-3**

**Totais Marginais e Total Geral para Dados de Celular**

<i>Ligações Pessoais: Sim</i>	<i>Ligações Pessoais: Sim</i>	<i>Ligações Pessoais: Não</i>	<i>Totais marginais das Linhas</i>
Homens	325	$183 = (508 - 325)$	508
Mulheres	427	$81 = (508 - 427)$	508
<i>Totais marginais das colunas</i>	752	264	1.016 <i>Total Geral</i>



Os totais marginais das linhas são a soma das frequências das células de cada linha; embora apareçam na tabela em uma coluna. Esse fenômeno ocorre, pois, ao somarmos as frequências das células de uma linha, colocamos o resultado no final da linha e, ao fazermos isso para cada linha, acabamos empilhando os totais das linhas, formando assim uma coluna. Da mesma forma, os totais marginais das colunas são a soma das frequências das células de cada coluna; embora apareçam na tabela em uma linha. Não deixe que isso o confunda quando estiver tentando navegar ou criar uma tabela de dupla entrada. Recomendo que você nomeie seus totais como marginal da linha, marginal da coluna e total geral para ajudá-lo a manter tudo organizado.

# Desmembrando as Probabilidades

Em uma tabela de dupla entrada, a porcentagem pode ser interpretada de duas formas — em termos de grupo ou de indivíduo. No que diz respeito a grupo, a porcentagem representa a parcela do grupo que se enquadra em uma determinada categoria. No entanto, a porcentagem também representa a probabilidade de que um indivíduo aleatoriamente selecionado de um grupo se enquadre em uma determinada categoria.

Sendo assim, a tabela de dupla entrada lhe dá a oportunidade de encontrar muitos tipos diferentes de probabilidades, que, por sua vez, o ajudam a encontrar as respostas às diferentes perguntas sobre seus dados ou a ver os dados com outros olhos. Nesta seção, abranjo os três tipos mais importantes de probabilidades em uma tabela de dupla entrada: as probabilidades marginais, as probabilidades conjuntas e as probabilidades condicionais. (Para uma abordagem mais completa sobre esses tipos de probabilidade, consulte o livro *Probability For Dummies*, escrito por esta que vos fala e publicado pela Wiley.)



Quando encontramos probabilidades a partir de uma amostra, como fizemos no presente capítulo, temos que nos dar conta de que tais probabilidades pertencem apenas a essa amostra em questão. Tenha em mente que elas não são transferidas automaticamente para a população que está sendo estudada. Por exemplo, se você coletar uma amostra aleatória de mil adultos e descobrir que 55% deles assistem a reality shows, isso não significa que 55% de todos os adultos de toda a população assistem a esse tipo de programa (e a mídia comete esse erro todos os dias). É preciso levar em conta o fato de que os resultados amostrais variam. Nos Capítulos 14 e 15, vamos fazer exatamente isso. Mas este capítulo se detém apenas em resumir as informações contidas na amostra, o primeiro passo para essa finalidade (mas não o último em relação à tirada de conclusões sobre a população correspondente).

## Probabilidades marginais

A *probabilidade marginal* tira uma probabilidade de um total marginal, tanto para as linhas quanto para as colunas. A probabilidade marginal representa a proporção de todo o grupo que pertence a uma única categoria de linha ou de coluna. Cada probabilidade marginal representa apenas uma categoria de apenas uma variável — ela não considera as outras variáveis. No exemplo do celular, tem-se quatro possíveis probabilidades marginais (consulte a Tabela 13-3):

- ✓ A probabilidade marginal das mulheres  $\left(\frac{508}{1.016} = 0,50\right)$ , o que significa que 50% de todos os usuários de celular nesta amostra eram do sexo feminino.
- ✓ A probabilidade marginal do sexo masculino  $\left(\frac{508}{1.016} = 0,50\right)$ , o que significa que 50% de todos os usuários de celular nesta amostra eram do sexo masculino.

- ✓ A probabilidade marginal de usar um celular para chamadas pessoais  $\left(\frac{752}{1.016} = 0,74\right)$ , o que significa que 74% de todos os usuários de celular desse exemplo utilizam seus celulares para ligações pessoais.
- ✓ A probabilidade marginal de não usar o celular para fazer ligações pessoais  $\left(\frac{264}{1.016} = 0,26\right)$ , o que significa que 26% de todos os usuários de celulares dessa amostra não utilizam seus celulares para ligações pessoais.

Porém, os estatísticos utilizam uma notação abreviada para todas essas probabilidades. Se considerarmos M = masculino, F = feminino, Sim = uso pessoal do celular e Não = uso impessoal do celular, então, as probabilidades marginais acima poderão ser escritas da seguinte forma:

- ✓  $P(F) = 0,50$
- ✓  $P(M) = 0,50$
- ✓  $P(\text{Sim}) = 0,74$
- ✓  $P(\text{Não}) = 0,26$

Observe que a soma de  $P(F)$  e  $P(M)$  é igual a 1,00. Porém, esse resultado não é uma coincidência, pois essas duas categorias compõem toda a variável sexo. Da mesma forma, a soma de  $P(\text{Sim})$  e  $P(\text{Não})$  também é 1,00, pois essas escolhas são as duas únicas para a variável uso pessoal do celular. Todos têm que se enquadrar em algum lugar.

Esteja ciente de que algumas probabilidades não são úteis para a descoberta de informações sobre a população em geral. Por exemplo,  $P(F) = 0,50$ , pois os pesquisadores já haviam determinado com antecedência que queriam uma amostra com exatamente 508 mulheres e 508 homens. O fato de que 50% da amostra é do sexo feminino e os outros 50%, do sexo masculino não significa que, em toda a população de usuários de celular, 50% são homens e 50% são mulheres. Se quiser saber qual a proporção de homens e mulheres na população total de usuários de celulares vai precisar coletar uma amostra combinada, em vez de duas separadas, e ver quantos homens e quantas mulheres aparecem nela.

## ***Probabilidades conjuntas***

A probabilidade conjunta nos fornece a probabilidade da interseção entre duas categorias, uma da variável linha e a outra da variável coluna. É a probabilidade de que alguém selecionado do grupo inteiro tenha duas características especiais ao mesmo tempo. Em outras palavras, as características aparecem juntas ou em conjunto. Para calcular a probabilidade conjunta, divida a frequência das células dos indivíduos que tiverem as duas características pelo total geral.

Aí, vão as quatro probabilidades conjuntas para o exemplo do celular:



- ✓ A probabilidade de que alguém no grupo todo seja do sexo masculino e use seu celular para chamadas pessoais é  $\frac{325}{1.016} = 0,32$ , o que significa que 32% de todos os usuários de celulares nessa amostra são do sexo masculino e usam seus celulares para fazer ligações pessoais.
- ✓ A probabilidade de que alguém no grupo inteiro seja do sexo masculino e não use seu celular para fazer chamadas pessoais é  $\frac{183}{1.016} = 0,18$ .
- ✓ A probabilidade de que alguém no grupo inteiro seja do sexo feminino e faça chamadas pessoais com seu celular é  $\frac{427}{1.016} = 0,42$ .
- ✓ A probabilidade de que alguém no grupo inteiro seja do sexo feminino e não faça ligações pessoais com seu celular é  $\frac{81}{1.016} = 0,08$ .

A notação para as probabilidades conjuntas mencionadas acima é a seguinte, onde representa a interseção entre as duas categorias indicadas:

- ✓  $P(M \cap \text{Sim}) = 0,32$
- ✓  $P(M \cap \text{Não}) = 0,18$
- ✓  $P(F \cap \text{Sim}) = 0,42$
- ✓  $P(F \cap \text{Não}) = 0,08$



A soma de todas as probabilidades conjuntas de qualquer tabela de dupla entrada deve ser igual a 1,00, a não ser que você tenha um pequeno erro de arredondamento, o que o deixa muito próximo a 1,00, ainda que não exatamente. O montante é 1,00, pois todos no grupo estão classificados de acordo com alguma categoria de ambas variáveis. É como se dividíssemos todo o grupo em quatro partes e mostrássemos quais proporções se enquadram em cada parte.

## ***Probabilidades condicionais***

A *probabilidade condicional* é o que você usa para comparar os subgrupos de uma amostra. Em outras palavras, se quiser decompor ainda mais a tabela, transforme-a em uma probabilidade condicional. Cada linha tem uma probabilidade condicional para cada uma de suas células, assim como também cada coluna tem uma probabilidade condicional para cada célula dentro dela.

**Observação:** uma vez que a probabilidade condicional é um dos pontos críticos para vários alunos, vou gastar um pouco mais de tempo falando sobre ela. Meu objetivo com esta seção é que você realmente entenda o que é a probabilidade condicional e como pode usá-la no mundo real (algo que muitos livros de Estatística negligenciam).

### ***Calculando as probabilidades condicionais***

Para encontrar uma probabilidade condicional, primeiro você deve analisar uma única linha ou coluna da tabela que represente a característica conhecida dos indivíduos. O total marginal para essa linha (coluna) representa agora o seu novo total geral, pois este grupo passa a ser o seu universo inteiro. Em seguida, divida a soma das frequências das células dessa linha (coluna) pelo total marginal dessa linha (coluna).

Considere o exemplo do celular na Tabela 13-3. Suponha que você queira examinar apenas os homens que participaram da pesquisa. O número total de homens é 508. É possível decompor esse grupo em dois subgrupos, utilizando a probabilidade condicional: você pode encontrar a probabilidade do uso de celulares para chamadas pessoais (somente para os homens) e a probabilidade de não se usar o celular para chamadas pessoais (somente para os homens). Da mesma forma, você pode decompor os dados referentes às mulheres em mulheres que utilizam o celular para chamadas pessoais e mulheres que não o utilizam para essa finalidade.

No exemplo do celular, ao decompor a tabela segundo o sexo dos participantes, tem-se as seguintes probabilidades condicionais:

- ✓ A probabilidade condicional de que um homem use o celular para fazer chamadas pessoais  $\frac{325}{508} = 0,64$ .
- ✓ A probabilidade condicional de que um homem não use o celular para chamadas pessoais  $\frac{183}{508} = 0,36$ .
- ✓ A probabilidade condicional de que uma mulher use o celular para fazer chamadas pessoais  $\frac{427}{508} = 0,84$ .
- ✓ A probabilidade condicional de que uma mulher não use o celular para chamadas pessoais  $\frac{81}{508} = 0,16$ .

Para interpretar esses resultados, pode-se dizer que dentro desta amostra se você é homem, o mais provável é que não use seu celular para chamadas pessoais (64% em comparação a 36%). No entanto, o percentual dos que fazem chamadas pessoais é maior entre os indivíduos do sexo feminino (84% versus 16%).



Observe que para os participantes do sexo masculino do exemplo anterior, as duas probabilidades condicionais (0,64 e 0,36) somam 1,00. Mas isso não é uma coincidência. O grupo de participantes do sexo masculino foi classificado pelo uso ou não do celular para chamadas pessoais, e, já que todos no estudo são usuários de celular, cada participante teve que ser classificado em um dos grupos. Da mesma forma, a soma das duas probabilidades condicionais para as participantes do sexo feminino também é 1,00.

### ***Notação para as probabilidades condicionais***

As probabilidades condicionais são designadas por uma linha reta vertical que lista e separa o evento ocorrido (o que é dado) e o evento para o qual se quer encontrar a

probabilidade. A notação pode ser escrita da seguinte forma:  $P(XX | XX)$ . Coloque o evento dado à direita da linha e o evento para o qual se quer encontrar a probabilidade à esquerda da linha. Por exemplo, suponha que você saiba que alguém é do sexo feminino (F) e quer descobrir as chances de que ela seja uma democrata (D). Neste caso, você está procurando  $P(D | F)$ . Por outro lado, digamos que você saiba que uma pessoa é democrata e deseja a probabilidade de que essa pessoa seja do sexo feminino — ou seja, está buscando  $P(F | D)$ .

A linha vertical na notação da probabilidade condicional não é um sinal de divisão, mas apenas uma linha separando os eventos A e B. Além disso, tenha cuidado com a ordem em que você coloca A e B na notação da probabilidade condicional. Em geral,  $P(A | B) \neq P(B | A)$ .

Na sequência, veja a notação utilizada para as probabilidades condicionais do exemplo do celular:

- ✓  **$P(\text{Sim} | M) = 0,64$ .** Você pode dizer desta maneira: “A probabilidade de Sim dado o Sexo Masculino é de 0,64.”
- ✓  **$P(\text{Não} | M) = 0,36$ .** Em termos humanos, dizemos: “A probabilidade de Não dado o Sexo Masculino é de 0,36.”
- ✓  **$P(\text{Sim} | F) = 0,84$ .** Encha a boca para dizer esta: “A probabilidade de Sim dado o Sexo Feminino é de 0,84.”
- ✓  **$P(\text{Não} | F) = 0,16$ .** Traduzindo: “A probabilidade de Não dado o Sexo Feminino é de 0,16.”

Você pode ver que  $P(\text{Sim} | M) + P(\text{Não} | M) = 1,00$ , pois estamos dividindo todos os homens em dois grupos: o dos que usam o celular para chamadas pessoais (S) e o dos que não usam (N). Observe, entretanto, que a soma de  $P(\text{Sim} | M) + P(\text{Sim} | F)$  não é 1,00. No primeiro caso, estamos analisando apenas o sexo masculino e, no segundo caso, apenas o sexo feminino.

### ***Comparando dois grupos através das probabilidades condicionais***

Uma das perguntas mais comuns a respeito de duas variáveis categóricas é esta: elas se relacionam? Para responder a essa pergunta, é preciso comparar suas probabilidades condicionais.

Para comparar as probabilidades condicionais, siga estes passos:

1. Tome uma variável e encontre suas probabilidades condicionais baseando-se nas outras variáveis.
2. Repita o primeiro passo para cada categoria da primeira variável.




**3. Compare as probabilidades condicionais (você pode até mesmo colocá-las em um gráfico para os dois grupos) e veja se elas são iguais ou diferentes.**

Se as probabilidades condicionais forem iguais para todos os grupos, isso significa que na amostra as variáveis não se relacionam. Se elas forem diferentes, isso significa que na amostra as variáveis se relacionam.

**4. Expanda os resultados para toda a população envolvida, utilizando os resultados amostrais para fazer um teste do Quiquadrado (veja o Capítulo 14).**

Voltando ao exemplo do celular da seção anterior, suas perguntas podem ser: o uso pessoal do celular se relaciona ao sexo do usuário? Sendo assim, você sabe que deve comparar o uso do celular por homens e mulheres para descobrir se ele está relacionado ao sexo. No entanto, é muito difícil comparar as frequências das células; por exemplo, 325 homens usam seus telefones para chamadas pessoais, comparados a 427 mulheres. Na verdade, é impossível comparar esses números sem usar um total como perspectiva. 325 de quantos?



Não há como comparar as frequências das células em dois grupos sem calcular as porcentagens (obtidas pela divisão de cada frequência das células pelo total adequado). A porcentagem lhe proporciona um método para a comparação de dois números em condições iguais. Por exemplo, suponha que você tenha feito uma pesquisa de opinião com apenas uma pergunta (cujas respostas eram sim, não e sem opinião) com uma amostra aleatória de 1.099 pessoas: 465 entrevistados responderam sim, 357 responderam não e 277 não souberam opinar. Para realmente conseguir interpretar essas informações, você provavelmente está tentando comparar esses números a outros. E é isso que a porcentagem faz por você. Colocada lado a lado, a porcentagem de cada grupo lhe proporciona a comparação relativa dos grupos.

Mas, primeiro, você precisa adicionar as probabilidades condicionais à mistura. No exemplo do celular, se quiser a porcentagem de mulheres que usam seus celulares para chamadas pessoais, divida 427 pelo número total de mulheres (508) e obtenha 84%. Da mesma forma, para obter a porcentagem de homens que usam seus celulares para chamadas pessoais, divida a frequência (325) pelo total das linhas para o sexo masculino (508), que lhe dá 64%. Esse percentual é a probabilidade condicional de usar o celular para chamadas pessoais, tendo em conta que a pessoa em questão é do sexo masculino.

Agora, você está pronto para comparar homens e mulheres usando as probabilidades condicionais. Compare o percentual de mulheres que usam seus celulares para chamadas pessoais à porcentagem de homens que usam seus celulares para esse mesmo fim. Ao encontrar essas probabilidades condicionais é possível comparar sem grande esforço os dois grupos e dizer que, pelo menos nesta amostra, as mulheres usam seus celulares (84%) para chamadas pessoais mais do que os homens (64%).

### ***Usando gráficos para mostrar as probabilidades condicionais***

Uma forma de enfatizar as probabilidades condicionais como uma ferramenta para

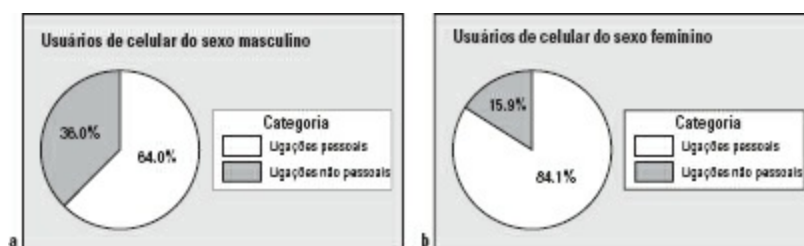


comparar dois grupos é a utilização de gráficos, tal como um gráfico de pizza comparando os resultados da outra variável para cada grupo ou um gráfico de barras comparando os resultados da outra variável para cada grupo. (Consulte um livro de Estatística I ou o *Estatística Para Leigos*, da Alta Books, para mais informações sobre gráficos de pizza e gráficos de barras).



Você pode estar se perguntando o quanto os dois gráficos de pizza precisam se parecer (em se tratando da semelhança entre as fatias de um gráfico em relação às do outro), a fim de que você possa dizer que as variáveis são independentes. Esse é o tipo de questão que você não poderá responder totalmente até que realize um teste de hipótese para as proporções (veja o teste do Qui-quadrado no Capítulo 14). Por enquanto, no que diz respeito aos dados de sua amostra, se a diferença na aparência das fatias dos dois gráficos for suficiente para que você escreva um artigo de jornal sobre ela, então, fique com a dependência. Caso contrário, conclua pela independência.

As Figuras 13-1a e 13-1b usam dois gráficos de pizza para comparar o uso do celular por homens e mulheres. A Figura 13-1a mostra a distribuição condicional do uso do celular para os participantes do sexo masculino (dado). A Figura 13-1b mostra a distribuição condicional do uso do celular para os participantes do sexo feminino (dado). A comparação entre as Figuras 13-1a e 13-1b revela que as fatias do gráfico para o uso de celulares não são iguais (tampouco semelhantes), o que significa que o sexo e uso do celular para chamadas pessoais são dependentes nesta amostra. Isso apenas confirma as conclusões anteriores.



**Figura 13-1:** Gráficos de pizza comparando homens versus mulheres quanto ao uso pessoal do celular.

Outra forma de fazer comparações é decompor a tabela de dupla entrada a partir da variável da coluna (não precisa usar sempre a variável da linha para fazer as comparações). No exemplo do celular (Tabela 13-3), você pode comparar o grupo dos que fazem chamadas pessoais com o grupo dos que não fazem e verificar qual é a porcentagem de homens e mulheres em cada grupo. Esse tipo de comparação dá um toque diferente à informação, pois compara os comportamentos em relação ao sexo dos participantes.

Com esta nova decomposição da tabela de dupla entrada, obtém-se o seguinte:

- ✓ A probabilidade condicional de que você seja um homem, dado que você utiliza o celular para fazer chamadas pessoais  $\frac{427}{752} = 0,57$ . **Observação:** o denominador é 752,

o número<sup>752</sup> total de pessoas que usam seus celulares para fazer ligações pessoais.

- ✓ A probabilidade condicional de que você seja uma mulher, dado que você utiliza o celular para fazer chamadas pessoais  $\frac{427}{752} = 0,57$ .

Mais uma vez, a soma dessas duas probabilidades é 1,00, pois você está decompondo os que fazem ligações pessoais de acordo com seu sexo (masculino ou feminino). As probabilidades condicionais para os usuários de celulares que não fazem ligações pessoais são  $P(M | \text{Não}) = \frac{183}{264} = 0,69$  e  $P(M | \text{Não}) = \frac{81}{264} = 0,31$ . A soma dessas duas probabilidades também é igual a 1,00, pois você está decompondo os que não fazem ligações pessoais segundo seu sexo (masculino e feminino).

As conclusões gerais se assemelham às encontradas na seção anterior, mas as porcentagens específicas e a interpretação são diferentes. Interpretando os dados dessa forma, pode-se dizer que, dentro desta amostra, se você usa seu celular para fazer chamadas pessoais, é mais provável que seja uma mulher do que um homem (57% em comparação a 43%). Mas se você não usa seu celular para fazer chamadas pessoais, é mais provável que seja um homem (69% em comparação a 31%).

## Dividir por quanto? Eis a questão!

Para obter a resposta correta para qualquer probabilidade de uma tabela de dupla entrada, aqui vai a dica: sempre identifique o grupo examinado. “Quanto de quanto?”. No exemplo do celular (veja a Tabela 13-3),

- ✓ Se quiser o percentual de todos os usuários que são homens e usam seus telefones para chamadas pessoais, divida a frequência 325 por 1.016, o total geral.
- ✓ Se quiser o percentual de homens que usam seus celulares para chamadas pessoais, divida 325 por 508, o número total de homens.
- ✓ Se quiser o percentual dos que fazem ligações pessoais e que sejam do sexo masculino, divida 325 por 752, o número total de pessoas que fazem ligações pessoais de seus celulares.

Em cada um desses três casos, o numerador é o mesmo, mas os denominadores são diferentes, o que o leva a obter respostas muito diferentes. Decidir que denominador usar geralmente é um motivo de confusão para as pessoas, e esse truque pode realmente lhe trazer alguma vantagem.

# *Tentando Ser Independente*

Em Estatística, a *independência* é um grande negócio. O termo geralmente indica que dois itens têm resultados cujas probabilidades não se influenciam. Os itens podem ser os eventos A e B, as variáveis x e y, os resultados de uma pesquisa feita com duas pessoas selecionadas aleatoriamente de uma população, e assim por diante. Porém, se os resultados de dois itens se influenciam, os estatísticos os chamam de *dependentes* (ou não independentes). Nesta seção, você vai verificar e interpretar a independência de categorias individuais, uma de cada variável categórica de uma amostra, mas também vai verificar e interpretar a independência de duas variáveis categóricas de uma amostra.

## *Verificando a independência entre duas categorias*

Os instrutores de Estatística muitas vezes ensinam seus alunos a verificar se duas categorias (uma da variável categórica x e outra da variável categórica y) são independentes. No entanto, prefiro só comparar os dois grupos e conversar sobre as semelhanças ou diferenças entre as porcentagens, decompondo-os através de outra variável. No entanto, para abordar todas as bases e me certificar de que você possa responder a essa pergunta tão popular, aqui vai a definição oficial de independência, diretamente da boca de uma estatística: duas categorias são *independentes* se sua probabilidade conjunta for igual ao produto das suas probabilidades marginais. O único problema aqui é que nenhuma das categorias pode estar completamente vazia.


Por exemplo, se o fato de ser mulher é independente do fato de ser democrata, então  $P(F \cap D) = P(F) * P(D)$ , onde D = Democratas e F = Feminino. Então, para mostrar que duas categorias são independentes, determine a probabilidade conjunta e a compare ao produto das duas probabilidades marginais. Se você obtiver a mesma resposta duas vezes, as categorias são independentes. Se não, as categorias são dependentes.

Porém, você pode estar se perguntando: mas todas as probabilidades não funcionam assim? Então, onde a probabilidade conjunta é igual ao produto dos marginais? Não. Por exemplo, se você tirar uma carta de um baralho padrão com 52 cartas, a probabilidade de tirar uma carta vermelha é  $\frac{1}{2}$ , e a probabilidade de tirar uma carta de copas é  $\frac{1}{4}$ . A chance de tirar tanto uma carta de copas quanto uma carta vermelha ainda é  $\frac{1}{4}$  (pois todas cartas de copas são vermelhas). No entanto, o produto das probabilidades individuais para a carta de copas e a carta vermelha acaba sendo,  $\frac{1}{4} * \frac{1}{2} = \frac{1}{8}$ , que é diferente de  $\frac{1}{4}$ . Isso indica que as categorias "copas" e "vermelha" não são independentes (ou seja, são dependentes). Agora, a probabilidade conjunta de um dois vermelho é  $\frac{2}{52}$  ou  $\frac{1}{26}$ , que equivale à probabilidade de uma carta vermelha,  $\frac{1}{2}$ , vezes a probabilidade de um dois (porque  $\frac{1}{2} * \frac{4}{52} = \frac{1}{26}$ ). Isso indica que as categorias "vermelho" e "dois" são independentes.


Outra forma de verificar a independência é comparar a probabilidade condicional à marginal. Para ser mais específica, se você quiser verificar se o fato de ser mulher é

independente do fato de ser democrata, verifique qualquer das duas situações a seguir (ambas funcionarão se as variáveis forem independentes):

- ✓  $P(F|D) = P(F)$ ? Ou seja, se você sabe que alguém é democrata, isso afeta a chance de que esse alguém também seja uma mulher? Se sim, F e D são dependentes. Se não, F e D são independentes.
- ✓  $P(D|F) = P(D)$ ? Essa pergunta quer saber se o fato de você ser uma mulher altera suas chances de ser uma democrata. Se sim, D e F são dependentes. Se não, D e F são independentes.

 O fato de saber que está em uma categoria vai mudar a probabilidade de você estar em outra categoria? Se sim, as duas categorias não são independentes. Se o fato de você saber isso não afeta a probabilidade, então, as duas categorias são independentes.

## *Verificando a independência entre duas variáveis*

 A seção anterior abordou a verificação da independência entre duas categorias específicas em uma amostra. Caso queira expandir essa ideia para mostrar que duas variáveis categóricas inteiras são independentes, verifique as condições de independência para cada combinação de categorias dessas variáveis. Todas devem funcionar, ou a independência estará perdida. O primeiro caso em que se encontrar dependência entre as duas categorias já é o suficiente para indicar que as duas variáveis são dependentes. Porém, se o primeiro caso mostrar independência, você deve continuar verificando todas as combinações antes de declarar independência.

Suponha que um consultório médico queira saber se ligar para os pacientes confirmando seus horários relaciona-se ao fato de eles realmente comparecerem à consulta. As variáveis são  $x$  = ligação para o paciente (recebeu ligação ou não recebeu ligação) e  $y$  = compareceu à consulta (ou não compareceu). Aqui, estão as quatro condições que precisam ser satisfeitas antes de declarar a independência:

- ✓  $P(\text{compareceu}) = P(\text{compareceu} | \text{recebeu ligação})$
- ✓  $P(\text{compareceu}) = P(\text{compareceu} | \text{não recebeu ligação})$
- ✓  $P(\text{não compareceu}) = P(\text{não compareceu} | \text{recebeu ligação})$
- ✓  $P(\text{não compareceu}) = P(\text{não compareceu} | \text{não recebeu ligação})$

Se qualquer uma dessas condições não for atendida, pare por aqui e declare as duas variáveis como sendo dependentes na amostra. Somente se todas as condições forem satisfeitas você pode declarar que as duas variáveis são independentes na amostra.

Veja, na Tabela 13-4, os resultados de uma amostra com 100 pacientes selecionados aleatoriamente para a situação ilustrada nesse exemplo.

Tabela 13-4

**Ligação de confirmação relacionada com o comparecimento à consulta**

	<i>Recebeu ligação</i>	<i>Não recebeu ligação</i>	<i>Total das linhas</i>
Compareceu	57	33	90
Não compareceu	3	7	10
<b><i>Total das colunas</i></b>	60	40	100

Para verificar as condições de independência, você pode começar na primeira condição e verificar se  $P(\text{compareceu}) = P(\text{compareceu} \mid \text{recebeu ligação})$ . Na última coluna da Tabela 13-4, você pode ver que  $P(\text{compareceu})$  é igual a  $\frac{90}{100} = 90\%$ . Em seguida, olhe na primeira coluna para encontrar  $P(\text{compareceu recebeu ligação})$ ; essa probabilidade é  $\frac{57}{60} = 95\%$ . Como essas duas probabilidades não são iguais (embora próximas), conclui-se que comparecer e receber a ligação são dependentes nesta amostra. Também se pode concluir que o comparecimento dos pacientes é um pouco mais alto quando o consultório liga confirmando. (Para determinar se esses resultados amostrais se expandem à população, que também se atenta à questão da proximidade necessária às probabilidades, a fim de concluir pela independência, consulte o Capítulo 14.)

# ***Desmistificando o Paradoxo de Simpson***

O *Paradoxo de Simpson* é um fenômeno em que os resultados parecem estar em contradição direta um com o outro, fazendo com que até mesmo o melhor aluno entre em desespero. Tal situação pode passar despercebida, a menos que três variáveis (ou mais) sejam examinadas, caso em que os resultados são organizados em uma *tabela de tripla entrada*, com colunas dentro de colunas ou linhas dentro de linhas.

O Paradoxo de Simpson é um dos favoritos entre os instrutores de Estatística (pois é tão mágico e místico — e os números ficam tão pegajosos e complexos), mas é um dos mais odiados entre os alunos, principalmente por causa dos seguintes motivos (na minha opinião):

- ✓ Devido à forma como o Paradoxo de Simpson é apresentado na maioria dos cursos de Estatística, onde você facilmente se vê enterrado nos detalhes e acaba não tendo nenhuma esperança de conseguir entender o todo. O Paradoxo de Simpson chama a atenção para um grande problema relacionado à interpretação dos dados e, portanto, é preciso entendê-lo plenamente se quiser evitá-lo.
- ✓ A maioria dos livros faz um bom trabalho ao mostrar exemplos do Paradoxo de Simpson, no entanto, falham ao tentar explicar por que ele ocorre. Alguns nem sequer tentam explicar isso! Esta seção vai ajudá-lo a ter uma ideia melhor do que é o Paradoxo de Simpson, a compreender melhor por que e como ele acontece, além de ensiná-lo a se precaver, para que esse fenômeno não ocorra.

## ***Experimentando o Paradoxo de Simpson***

O Paradoxo de Simpson foi descoberto em 1951 pelo estatístico americano E. H Simpson. Ele percebeu que, quando você analisava alguns conjuntos de dados dividindo-os em apenas duas variáveis, conseguia obter um resultado, mas, quando os dividia em três variáveis, os resultados mudavam de direção. Por isso, seu resultado leva o nome de *Paradoxo de Simpson* — sendo o paradoxo uma aparente contradição nos resultados.

### ***Paradoxo de Simpson em ação: Video games e a diferença entre os sexos***

A melhor maneira de entender o Paradoxo de Simpson é vê-lo em cena por meio de um exemplo e explicar todos os porquês ao longo do caminho. Suponha que eu esteja interessada em descobrir quem é melhor no video game, homens ou mulheres. Então, assisto a homens e mulheres escolherem e jogarem uma variedade de jogos e registro se ganharam ou perderam. Suponha que eu tenha registrado os resultados de 200 jogos, como visto na Tabela 13-5 (note que as mulheres jogaram 120 jogos, e os homens, 80).

---

**Tabela 13-5**

**Vitória ou Derrota em Jogos de Video Games para Homens  
versus Mulheres**

---

	<i>Ganharam</i>	<i>Perderam</i>	<i>Total marginal das linhas</i>
Homens	44	36	80
Mulheres	84	36	120
Total marginal das colunas	128	72	200 (Total geral)

Segundo a Tabela 13-5, a proporção de homens que ganharam no video game,  $P(\text{Ganhou}|\text{Homem})$ , é  $\frac{44}{80} = 0,55$ . A proporção de mulheres que ganharam os jogos,  $P(\text{Ganhou}|\text{Mulher})$ , é  $\frac{84}{120} = 0,70$ . Portanto, de modo geral, as mulheres ganharam mais do que os homens. Mas será que esta descoberta significa que, em se tratando de video game, as mulheres são melhores do que os homens na amostra?

Vá com calma, meu amigo. Observe que as pessoas no estudo podiam escolher os jogos, fator que acaba deixando o estudo em aberto. Suponha que homens e mulheres tenham escolhido tipos diferentes de jogos: será que isso pode afetar os resultados? Talvez, a resposta seja sim. É importante considerar outras variáveis que poderiam estar relacionadas aos resultados, mas que não foram incluídas no estudo original (ou, pelo menos, não na análise de dados original). Essas variáveis extras que ocultam os resultados são chamadas de *variáveis de confusão*.

### ***Fatorando no nível de dificuldade***

Muitos de vocês devem estar esperando que os resultados do exemplo da seção anterior virem o jogo e passe a indicar que, em se tratando de video game, os homens são melhores do que as mulheres. Segundo a pesquisa, em média, os homens passam mais tempo jogando e são, de longe, os principais compradores desse tipo de aparelho em comparação com as mulheres. Então, o que explica os resultados surpreendentes deste estudo? Será que existe outra explicação possível? Será que está faltando informações importantes, relevantes para este caso?

Uma das variáveis não consideradas quando fiz a Tabela 13-5 foi o nível de dificuldade do jogo escolhido. Suponha que eu volte e inclua o nível de dificuldade do jogo escolhido a cada vez com cada resultado (ganhou ou perdeu). O nível um representa os jogos fáceis, comparáveis ao nível do Come-Come (meu tipo de jogo) e o nível dois representa os jogos mais difíceis (como os jogos de guerra ou sofisticados jogos de estratégia).

A Tabela 13-6 representa os resultados obtidos com a inclusão da informação sobre o nível de dificuldade dos jogos. Agora, temos três variáveis: o nível de dificuldade (um ou dois), o sexo (masculino ou feminino) e os resultados (ganhou ou perdeu). Isso faz com que a Tabela 13-6 seja uma tabela de tripla entrada.

**Tabela 13-6 Tabela de Tripla Entrada para Sexo, Nível do Jogo e Resultado**

<i>Jogos de Nível Um</i>		<i>Jogos de Nível Dois</i>	
Ganhou	Perdeu	Ganhou	Perdeu



Homens	9	1	35	35
Mulheres	72	18	12	18

Observe na tabela 13-6 que o número escolhido de jogos de nível um foi  $9 + 1 + 72 + 18 = 100$ , e o número escolhido de jogos de nível dois foi  $35 + 35 + 12 + 18 = 100$ . A fim de reavaliar os dados com base nas informações sobre nível de dificuldade do jogo, você precisa examinar quem escolheu cada tipo de nível do jogo. A próxima seção vai sondar esta questão.

### ***Comparando as taxas de sucesso através das probabilidades condicionais***

Para comparar as taxas de sucesso para o sexo masculino versus o feminino, utilizando a Tabela 13-6, você deve descobrir as probabilidades condicionais apropriadas, primeiro, para os jogos de nível um e, em seguida, para os de nível dois. Para os jogos de nível um (apenas), a probabilidade condicional de ganhar dado o sexo masculino é

$P(\text{Venceu}|\text{Masc}) = \frac{9}{10} = 0,90$ . Assim, para os jogos de nível um, os homens ganharam em 90% das vezes que jogaram. Para os jogos de nível um, a porcentagem de jogos vencidos pelas mulheres é  $P(\text{Venceu}|\text{Fem}) = \frac{72}{90} = 0,80$ . Esses resultados indicam que, no nível um, os homens se saíram 10% melhores do que as mulheres. No entanto, tal porcentagem parece contradizer os resultados encontrados na Tabela 13-5. (Mas espere um pouco — a contradição não para por aqui!)

Agora, calcule as probabilidades condicionais para os jogos de nível dois. Nesse caso, a porcentagem de homens que ganharam é  $\frac{35}{70} = 0,50$  0,50, ou 50%. Já a porcentagem de mulheres que ganharam os jogos de nível dois foi  $\frac{12}{30} = 0,40 = 0,40$ , ou 40%. E, mais uma vez, os homens superaram as mulheres! Pare e pense nesta situação por um minuto. A Tabela 13-5 mostra que as mulheres ganharam uma maior porcentagem dos jogos em geral. Mas a Tabela 13-6 mostra que os homens ganharam mais tanto nos jogos de nível um quanto nos de nível dois. Afinal, o que está acontecendo? Não precisa verificar suas contas. Não tem erro nenhum aqui — nem pegadinhas. De vez em quando, esta inconsistência nos resultados acontece na vida real, em situações em que uma terceira variável importante é deixada de fora de um estudo, situação conhecida como *Paradoxo de Simpson*. (Entendeu por que ele se chama paradoxo?)

### ***Descobrendo o porquê do Paradoxo de Simpson***

As variáveis de confusão são a causa subjacente do Paradoxo de Simpson. A *variável de confusão* é uma terceira variável que se relaciona com cada uma das outras duas variáveis e pode influenciar os resultados quando não levadas em consideração.

No exemplo do video game, quando analisamos os resultados dos jogos (ganhado ou perdido), divididos apenas segundo o sexo do jogador (Tabela 13-5), vemos que as mulheres ganharam uma maior porcentagem dos jogos em geral do que os homens (70%





para as mulheres contra 55% para os homens). No entanto, quando se dividem os resultados segundo o nível de dificuldade (nível um ou nível dois; veja a Tabela 13-6), os resultados se invertem, e vemos que os homens se saíram melhor do que as mulheres nos jogos de nível um (90% contra 80%) e também se saíram melhor nos de nível dois (50% contra 40%).

Para entender o porquê desse resultado aparentemente impossível, analise as *probabilidades marginais das linhas versus o total marginal das linhas* dos jogos de nível um na Tabela 13-6. A porcentagem de vezes que um homem ganhou quando jogou um jogo fácil foi de 90%. No entanto, os homens escolheram os jogos de nível um apenas 10 vezes das 80 vezes que jogaram, ou seja, em apenas 12,5% das vezes que jogaram.

Para decompor ainda mais essa ideia, o desempenho meia boca dos homens nos jogos mais difíceis (50% — mas, ainda assim, melhor do que o das mulheres), aliado ao fato de que os homens optaram por jogos mais difíceis em 87,5% das vezes (ou seja, em 70 das 80 vezes) derrubou sua porcentagem de vitórias (55%). Embora os homens tenham se saído muito bem nos jogos de nível um, eles não jogaram muitas vezes (em comparação com as mulheres) e, portanto, sua elevada porcentagem de vitória nos jogos de nível um (90%) não contou muito para sua porcentagem total de vitórias.

Enquanto isso, na Tabela 13-6, vemos que as mulheres escolheram os jogos de nível um em 90 vezes (das 120 que jogaram). Mesmo que as mulheres tenham ganhado só 72 dos 90 jogos (ou seja, 80%, uma porcentagem mais baixa do que a dos homens, que venceram 9 dos 10 jogos), elas optaram por jogar muito mais os jogos de nível um e, portanto, aumentaram seus percentuais de vitórias.

Agora, a situação inversa acontece quando você analisa os jogos de nível dois na Tabela 13-6. Os homens optaram pelos jogos mais difíceis em 70 das 80 vezes que jogaram, enquanto as mulheres só escolheram os mais difíceis em apenas 30 de 120 vezes. Os homens também se saíram melhor do que as mulheres nos jogos de nível dois (ganhando 50% deles versus 40% para as mulheres). Entretanto, os jogos de nível dois são mais difíceis de vencer do que os jogos de nível um. Isso significa que a porcentagem de vitórias nos jogos de nível dois dos homens, sendo apenas de 50%, não contribuiu muito para a sua porcentagem total de vitórias. Porém, a baixa porcentagem de vitória nos jogos de nível dois das mulheres não prejudicou muito sua porcentagem total de vitórias, pois elas não jogaram muito os jogos de nível dois.

A verdade é que a ocorrência ou não do Paradoxo de Simpson é uma questão de pesos. No total geral da Tabela 13-5, os homens não parecem estar tão bem quanto as mulheres. Mas, quando você adiciona a dificuldade dos jogos, verá que a maioria das vitórias dos homens se dá nos jogos mais difíceis (que têm menor porcentagem de vitórias). Na média, as mulheres jogaram muito mais os jogos mais fáceis, jogos cuja chance de vitória é maior, independentemente de quem os joga. Então, tudo se resume a isto: quais os jogos que os homens escolheram para jogar e quais os que as mulheres escolheram? Os homens escolheram jogos mais difíceis, o que contribuiu de forma negativa para sua porcentagem



total de vitórias e fez com que parecesse que as mulheres se saíram melhor.

## ***De olho no Paradoxo de Simpson***

O Paradoxo de Simpson mostra a importância de incluir dados sobre possíveis variáveis de confusão ao se tentar analisar as relações entre variáveis categóricas.

O nível do jogo não foi incluído no resumo original, a Tabela 13-5, mas deveria ter sido, pois é uma variável que afeta os resultados. O nível do jogo, neste caso, é uma variável de confusão. Uma quantidade maior de homens escolheu jogar os jogos mais difíceis, cujas chances de vencer são menores, e, portanto, diminuíram sua taxa total de sucesso.



Você pode evitar o Paradoxo de Simpson certificando-se de que as variáveis de confusão foram incluídas no estudo; dessa forma, quando for examinar os dados, vai obter a relação correta logo na primeira vez e não correrá o menor risco de obter resultados contraditórios. E, assim como em todos os outros resultados estatísticos, se parecer bom demais para ser verdade ou simples demais para ser correto, então, provavelmente, é mesmo! Cuidado com quem tenta simplificar demais qualquer resultado. Embora as tabelas de tripla entrada sejam um pouco mais difíceis de serem analisadas, em muitos casos, vale a pena usá-las.

# Capítulo 14

## Independente o Suficiente para o Teste do Qui-quadrado

---

### *Neste Capítulo*

- ▶ Testando a independência na população (não apenas na amostra)
  - ▶ Usando a distribuição Qui-quadrada
  - ▶ Descobrindo a ligação entre o teste-Z e o teste do Qui-quadrado
- 

**T**enho certeza de que você já viu um desses julgamentos precipitados — pessoas que coletam uma amostra de dados e tentam usá-la para tirar conclusões a respeito de toda uma população. Quando tratamos de duas variáveis categóricas (em que os dados estão divididos em categorias e não representam medições), o problema parece estar ainda mais difundido.

Por exemplo, um noticiário de TV descobre que de 1.000 eleitores, 200 mulheres vão votar no Partido Republicano, 300 mulheres vão votar no Partido Democrata, 300 homens vão votar no Partido Republicano e 200 homens vão votar nos democratas. A âncora do telejornal, então, mostra os dados e afirma que 30% ( $300 \div 1.000$ ) dos votos de *todos* os eleitores do sexo feminino vai para os democratas (e assim sucessivamente para as outras frequências).

No entanto, tal conclusão é enganosa. É verdade que, nesta amostra de 1.000 eleitores, 30% deles são do sexo feminino e votam nos democratas. No entanto, este resultado não significa automaticamente que 30% de toda a população de eleitores do sexo feminino vota nos democratas. Afinal de contas, os resultados amostrais variam de amostra para amostra.

Neste capítulo, você verá como ir além da simples organização dos resultados da amostra em uma tabela de dupla entrada (discutida no Capítulo 13) e usar esses resultados em um teste de hipótese para tirar conclusões a respeito de uma população inteira. Esse processo exige uma nova distribuição de probabilidade chamada *distribuição Qui-quadrado*. Você também vai descobrir como responder a uma pergunta muito popular entre os pesquisadores: estas duas variáveis categóricas são independentes (não se relacionam) para toda a população?

# O Teste do Qui-quadrado para a Independência

Uma das razões mais comuns para a coleta de dados é a busca por relações entre variáveis. A análise isolada de uma variável geralmente não funciona para esse fim. Os métodos utilizados para analisar dados em busca de relações são diferentes, dependendo do tipo de dados coletado. Se as duas variáveis forem quantitativas (por exemplo, tempo de estudo e nota da prova), usam-se a correlação e a regressão (veja o Capítulo 4). No entanto, se as duas variáveis forem categóricas (por exemplo, sexo e filiação política), usa-se o teste do Qui-quadrado para examinar tal relação. Nesta seção, você verá como usar um teste do Qui-quadrado para verificar as relações entre duas variáveis categóricas.



Se duas variáveis categóricas não se relacionarem, são consideradas *independentes*. Porém, se apresentarem algum tipo de relação, são chamadas de *variáveis dependentes*. Muitas pessoas se confundem com esses termos, por isso, é importante entender bem a distinção entre eles.

Para testar se duas variáveis categóricas são independentes, é preciso usar o teste do Qui-quadrado. Veja a seguir as etapas para a realização do teste do Qui-quadrado. (O Minitab pode realizar este teste por você a partir do terceiro passo descrito abaixo.)

## 1. Colete os dados e resuma-os em uma tabela de dupla entrada.

Esses números representam as frequências observadas. (Para mais informações sobre as tabelas de dupla entrada, veja o Capítulo 13.)

## 2. Estabeleça sua hipótese nula, $H_0$ : as variáveis são independentes; e a hipótese alternativa, $H_a$ : as variáveis são dependentes.

## 3. Calcule as frequências das células segundo a hipótese de independência.

A frequência para uma célula é o produto entre o total da linha e o total da coluna divididos pelo total geral.

## 4. Verifique as condições necessárias para a realização do teste do Qui-quadrado antes de prosseguir; cada frequência de célula esperada deve ser maior ou igual a cinco.

## 5. Calcule a estatística Qui-quadrado.

Para isso, subtraia as células observadas pelas células esperadas, eleve essa diferença ao quadrado e divida-a pela célula esperada. Faça esse passo para cada célula e, depois, some-as.

## 6. Procure a estatística de teste na tabela do Qui-quadrado (Tabela A-3 no apêndice) e encontre o valor- $p$ (ou um que seja próximo).

## 7. Se o resultado for menor que o valor de corte predeterminado (o nível $\alpha$ ), geralmente igual a 0,05, rejeitamos a $H_0$ e concluímos a dependência das duas variáveis.



O relatório da American Demographics concluiu que, a partir desses dados "... Em geral, homens e mulheres concordam sobre a cor da fachada da casa, sendo a principal exceção a primeira escolha dos homens, a cor branca (36% pintaria a casa de branco versus 25% das mulheres)". Este tipo de conclusão é comum, mas não deixa de ser uma generalização dos resultados obtidos neste momento.

Você sabe que nesta amostra a quantidade de homens que pintariam suas casas de branco é maior do que a de mulheres, mas será que 180 é realmente diferente de 125, quando se está lidando com uma amostra de 1.000 pessoas, cujos resultados irão variar na próxima vez em que a pesquisa for feita? Como você sabe que estes resultados podem ser transferidos para a população de todos os homens e todas as mulheres? Essa questão não pode ser respondida sem a realização de um procedimento estatístico formal chamado teste de *hipótese*. (Consulte o Capítulo 3 para o básico sobre testes de hipótese.)

Para mostrar que homens e mulheres na população se diferem em relação à cor que pintariam suas casas, primeiro note que você tem duas variáveis categóricas:

- ✓ Sexo (masculino ou feminino)
- ✓ Cor da pintura (branca ou outra)



Tirar conclusões a respeito da população com base em uma amostra de dados (observação) organizada em uma tabela de dupla entrada é como dar um salto maior do que a perna. Você precisa realizar um teste do Quiquadrado a fim de ampliar suas conclusões a toda a população. A mídia (e até mesmo alguns pesquisadores) pode acabar encrocada por ignorar o fato de que os resultados da amostra variam. Contentar-se com os resultados amostrais pode levá-lo a conclusões que outros não poderão confirmar quando coletarem novas amostras.



Você consegue manter a conexão entre as duas informações se organizar os dados em uma tabela de dupla entrada em vez de usar duas tabelas individuais — uma para homens e mulheres e outra para a cor de preferência para a pintura da casa. Com uma tabela de dupla entrada, você pode examinar melhor a relação entre as duas variáveis. (Para detalhes completos sobre a organização e interpretação dos resultados de uma tabela de dupla entrada, consulte o Capítulo 13.)

## ***Determinando as hipóteses***

Todo teste de hipótese (seja um teste do Qui-quadrado ou qualquer outro) possui duas hipóteses:

- ✓ **Hipótese nula:** Na qual você tem que acreditar até que alguém prove o contrário. A notação para essa hipótese é  $H_0$ .
- ✓ **Hipótese alternativa:** Você deverá concluir esta no caso de não poder mais sustentar a hipótese nula. A notação para essa hipótese é  $H_a$ .

No caso em que se está testando a independência de duas variáveis categóricas, a hipótese nula é a não existência de uma relação entre elas. Em outras palavras, elas são independentes. A hipótese alternativa é a existência de uma relação entre as duas variáveis, ou seja, elas são dependentes.

Para o exemplo da cor da casa apresentado na seção anterior, escreva  $H_0$ : sexo e a cor de preferência são independentes versus  $H_a$ : sexo e cor de preferência são dependentes. E aí está o segundo passo do teste do Quiquadrado.

Para mais sobre testes de hipótese, veja o Capítulo 3. (Consulte um livro de Estatística I ou o *Estatística Para Leigos*, da Alta Books, para uma discussão completa sobre o tópico.)

## Calculando as frequências esperadas

Quando tiver coletado os dados e criado sua tabela de dupla entrada (por exemplo, veja a Tabela 14-1), já saberá os valores observados para cada célula na tabela. Agora, precisa compará-los a algo. Você, então, está pronto para o terceiro passo do teste do Quiquadrado — encontrar as frequências esperadas para as células.

A hipótese nula afirma que as duas variáveis  $x$  e  $y$  são independentes, ou seja, isso é o mesmo que dizer que  $x$  e  $y$  não têm nenhuma relação. Assumindo a hipótese de independência, é possível determinar quais os números que devem estar em cada célula da tabela através de uma fórmula para o que chamamos de *frequência esperada para as células*. (Cada quadrado em uma tabela de dupla entrada é chamado de célula, e o número encontrado em cada célula recebe o nome de *frequência da célula*; veja o Capítulo 13.)

A Tabela 14-1 mostra as frequências observadas para o exemplo da relação entre sexo e a cor de preferência. Para encontrar a frequência esperada, divida o produto entre o total da linha e o total da coluna pelo total geral, e faça isso para cada célula da tabela. A Tabela 14-2 mostra os cálculos para as frequências de célula esperadas para os dados da relação entre a variável sexo e a cor de preferência.

**Tabela 14-2 Sexo e Preferência pela Cor da Casa: Frequências esperadas**

	<i>Branca</i>	<i>Outra Cor</i>	<i>Total Marginal das Linhas</i>
Homens	$(500 * 305) \div 1000 = 152,5$	$(500 * 695) \div 1000 = 347,5$	500
Mulheres	$(500 * 305) \div 1000 = 152,5$	$(500 * 695) \div 1000 = 347,5$	500
<i>Total Marginal das Colunas</i>	305	695	1000 ( <i>Total Geral</i> )

Em seguida, compare as frequências de células observadas na Tabela 14-1 com as frequências de célula esperadas na Tabela 14-2, analisando suas diferenças. As diferenças

entre as frequências de célula observadas e esperadas apresentadas nessas tabelas são as seguintes:

$$180 - 152,5 = 27,5$$

$$320 - 347,5 = -27,5$$

$$125 - 152,5 = -27,5$$

$$375 - 347,5 = 27,5$$

Em seguida, faça um teste do Qui-quadrado para verificar a independência (consulte o Capítulo 15) a fim de determinar se as diferenças encontradas na amostra entre as frequências observadas e esperadas se devem simplesmente ao acaso ou se podem ser expandidas para a população.

Ao assumir a hipótese de independência, conclui-se que não há uma diferença significativa entre o que foi observado e o esperado.

## *Verificando as condições para o teste*

O quarto passo do teste do Qui-quadrado é verificar as condições. O teste do Qui-quadrado tem uma condição essencial que deve ser satisfeita para que a independência em uma tabela de dupla entrada possa ser testada: a frequência esperada para cada célula deve ser pelo menos cinco — ou seja, maior ou igual a cinco. As frequências esperadas que forem menores do que cinco não são confiáveis em termos da variabilidade que pode ocorrer.

No exemplo da relação entre cor de preferência e sexo, a Tabela 14-2 mostra que todas as frequências de células esperadas são pelo menos iguais a cinco, assim, a condição do teste do Qui-quadrado foi satisfeita.

Se estiver analisando os dados e descobrir que seu conjunto de dados não atende à condição para a frequência esperada (de ser igual a pelo menos cinco para uma ou mais células), você pode combinar algumas de suas linhas e/ou colunas. Essa combinação faz com que sua tabela fique menor, mas aumenta as frequências das células que você tem, o que ajuda a satisfazer essa condição.

## *Calculando a estatística Qui-quadrado*

Todo teste de hipóteses utiliza os dados para decidir se deve ou não rejeitar  $H_0$  em favor da  $H_a$ . No caso dos testes para independência em uma tabela de dupla entrada, utiliza-se um teste de hipóteses com base na estatística de teste Qui-quadrado. Nas seções seguintes, você poderá ver as etapas para o cálculo e a interpretação da estatística de teste Qui-quadrado, o quinto passo do teste.



## Calculando a fórmula

O principal componente da estatística de teste do Qui-quadrado é a frequência esperada para cada célula da tabela. A fórmula para encontrar a frequência esperada,  $e_{ij}$ , para a célula na linha  $i$  e coluna  $j$  é  $e_{ij} = \frac{\text{total de linha } i * \text{total de coluna } j}{\text{total geral}}$

$$e_{ij} = \frac{\text{total de linha } i * \text{total de coluna } j}{\text{total geral}}$$

Note que os valores de  $i$  e  $j$  variam para cada célula da tabela. Em uma tabela de dupla entrada, a célula superior esquerda da tabela está na linha um, coluna um. A célula no canto superior direito está na linha um, coluna dois. A célula no canto inferior esquerdo está na linha dois, coluna um, e a célula inferior direita está na linha dois, coluna dois.

A fórmula para a estatística de teste do Qui-quadrado é  $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$ , onde  $O_{ij}$  é a frequência observada para a célula na linha  $i$ , coluna  $j$ , e  $e_{ij}$  é a frequência esperada para a célula na linha  $i$ , coluna  $j$ .

Quando calculamos as frequências esperadas para algumas células, normalmente, obtemos um número com alguns dígitos depois da vírgula (em outras palavras, um número não inteiro). Não o arredonde, apesar da tentação em fazê-lo. A frequência esperada é, na verdade, um valor geral da média esperada, portanto, você deve mantê-la com os decimais.

## Calculando a estatística de teste

Aqui, estão os principais passos para calcular a estatística de teste do Quiquadrado para independência (o Minitab também executa estes passos para você):

1. Subtraia a frequência observada da esperada para a célula superior esquerda da tabela.
2. Eleve o resultado do primeiro passo ao quadrado para torná-lo um número positivo.
3. Divida o resultado do segundo passo pela frequência esperada.
4. Repita esse processo para todas as células da tabela e, depois, some todos os resultados para chegar à estatística de teste do Qui-quadrado.

A razão para a divisão pela frequência esperada é considerar os tamanhos das frequências. Se você espera uma frequência grande, digamos, 100, e se engana por apenas 5 em relação à frequência observada nessa célula, tal diferença não deve contar tanto quanto contaria caso você tivesse esperando uma frequência pequena (como 10) e tivesse se enganado por 5 em relação à frequência observada. A divisão pela frequência esperada coloca um peso mais justo sobre as diferenças usadas na estatística do teste do Qui-quadrado.

Para realizar o teste do Qui-quadrado no Minitab, primeiro você deve inserir os dados brutos (os dados sobre cada pessoa) em duas colunas. A primeira coluna contém os valores

da primeira variável do conjunto de dados. (Por exemplo, se a sua primeira variável é o sexo, preencha a primeira coluna do Minitab com o sexo de cada pessoa.) Em seguida, insira os dados da segunda variável na segunda coluna, onde cada linha representa uma única pessoa do conjunto de dados. (Se a sua segunda variável é a cor de preferência para a pintura da casa, por exemplo, insira a preferência de cada pessoa na coluna dois, mantendo juntos os dados de cada pessoa em cada linha.) Vá ao menu Stat>Tables>Cross Tabulation and Chi-square.

Agora, o Minitab precisa saber qual é a variável linha e qual é a variável coluna em sua tabela. No lado esquerdo, clique sobre a variável que você quer que represente as linhas de sua tabela de dupla entrada (você pode clicar sobre a primeira variável). Clique em Select e o nome da variável aparece na parte variável linha da tabela à direita. Agora, localize a variável coluna em branco, ao lado direito e clique sobre ela. Vá ao lado esquerdo e clique no nome de sua segunda variável. Clique em Select. Em seguida, clique no botão do Chisquare e escolha Chi-square analysis, marcando a opção. Se você deseja que as frequências esperadas sejam incluídas, marque também essa opção. Depois, clique em OK. Por último, clique em OK de novo para limpar todas as janelas.

### ***Analizando a saída***

A saída do Minitab para a análise do Qui-quadrado para o exemplo em questão (Tabela 14-1) está na Figura 14-1. Você pode escolher alguns números presentes na saída ilustrada pela Figura 14-1, que são especialmente importantes. Os seguintes três números estão listados em cada célula:

- ✓ O primeiro número é a frequência observada para a célula; ele coincide com a frequência observada para cada célula mostrada na Tabela 14-1. (Note que os totais marginais de linha e de coluna da Figura 14-1 também coincidem com os da Tabela 14-1.)
- ✓ O segundo número em cada célula da Figura 14-1 é a frequência esperada para a célula; para encontrá-la, divida o produto entre o total da linha e o total da coluna pelo total geral (consulte a seção “Calculando as frequências esperadas”). Por exemplo, a frequência esperada para a célula superior esquerda (homens que preferem pintar suas casa de branco),  $(500 * 305) \div 1.000 = 152,50$ .
- ✓ O terceiro número em cada célula da Figura 14-1 é a parte da estatística de teste Qui-quadrado que vem da célula. (Consulte os passos de um a três da seção anterior “Calculando a fórmula”.) A soma dos terceiros números de cada célula é igual ao valor da estatística Qui-quadrado que aparece na última linha da saída. (No exemplo da cor de preferência para a pintura da casa, a estatística de teste do Qui-quadrado é 14,27.)

Chi-Square Test: Gender, House-Paint Preference			
Expected counts are printed below observed counts			
Chi-Square contributions are printed below expected counts			
	White Paint	Nonwhite Paint	Total
M	180	320	500
	152.50	347.50	
	4.959	2.176	
F	125	375	500
	152.50	347.50	
	4.959	2.176	
Total	305	695	1000
Chi-Sq = 14.271, DF = 1, P-Value = 0.000			

**Figura 14-1:** Saída do Minitab para dados referentes à cor de preferência para a pintura da casa.

## ***Encontrando seus resultados na tabela do Qui-quadrado***

A única maneira de avaliar sua estatística de teste do Qui-quadrado é compará-la a todas as possíveis estatísticas de teste do Qui-quadrado que pudessem ser obtidas caso você tivesse uma tabela de duas entradas com os mesmos totais para linhas e colunas, ainda que os números nas células tivessem sido distribuídos de todas as maneiras possíveis. (Isso você faz com o pé nas costas, não é?) Algumas tabelas resultam em estatísticas de teste do Qui-quadrado grandes, enquanto outras resultam em pequenas.

A junção de todas essas estatísticas de teste do Qui-quadrado resulta no que chamamos de *distribuição do Qui-quadrado*. Você deve encontrar sua estatística de teste nessa distribuição (sexto passo do teste do Qui-quadrado) e ver sua posição em relação às demais.

Se sua estatística de teste for grande a ponto de parecer sair do alcance da ponta direita da distribuição Qui-quadrado (audaciosamente indo aonde nenhuma estatística de teste esteve antes), rejeite  $H_0$  e conclua que as duas variáveis não são independentes. Porém, se a estatística de teste não for tão longe assim, não se pode rejeitar  $H_0$ .

Nas próximas seções, você vai encontrar mais informações sobre a distribuição Qui-quadrado e como ela se comporta, podendo, assim, tomar uma decisão sobre a independência das duas variáveis com base em sua estatística do Qui-quadrado.

## ***Determinando os graus de liberdade***

Cada tipo de tabela de dupla entrada tem sua própria distribuição Quiquadrado, dependendo de seu número de linhas e colunas, e, por sua vez, cada distribuição Qui-quadrado é identificada por seus *graus de liberdade*.

Em geral, uma tabela de dupla entrada com  $r$  linhas e  $c$  colunas usa uma distribuição Qui-quadrado com  $(r - 1) * (c - 1)$  graus de liberdade. Assim, uma tabela de dupla entrada,

com duas linhas e duas colunas, usa uma distribuição do Qui-quadrado com um grau de liberdade. Observe que  $1 = (2 - 1) * (2 - 1)$ . Já uma tabela de dupla entrada com três linhas e duas colunas usa uma distribuição do Qui-quadrado com  $(3 - 1) * (2 - 1) = 2$  graus de liberdade.

A compreensão de por que os graus de liberdade são calculados desta forma está além do escopo de suas aulas de Estatística. Mas, se você realmente quiser saber, os graus de liberdade representam o número de células flexíveis, ou livres, da tabela, dados todos os totais marginais das linhas e colunas.

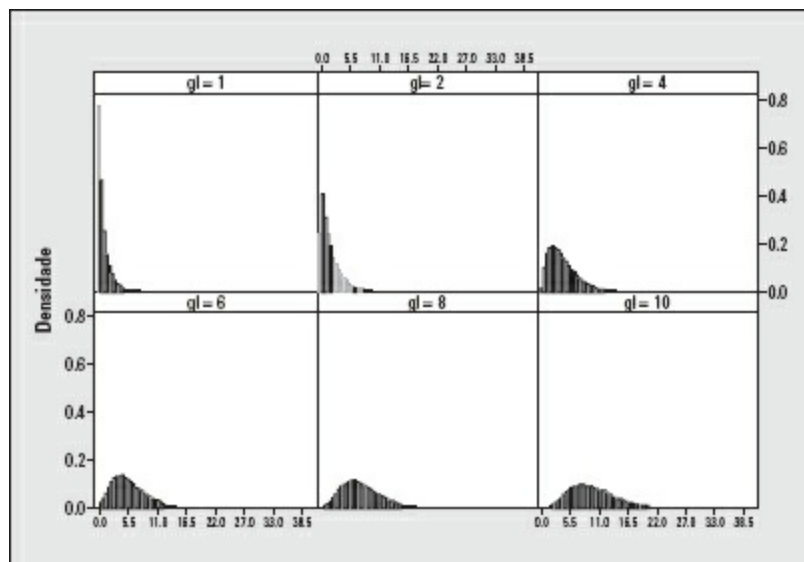
Por exemplo, suponha que todos os totais de linhas e colunas de uma tabela de dupla entrada sejam iguais a 100 e a célula superior esquerda seja 70. Assim, a célula superior direita deve ser igual a  $100 - 70 = 30$ . Como o total de uma coluna é 100 e a frequência da célula superior esquerda é 70, a frequência da célula inferior esquerda deve ser  $100 - 70 = 30$ . Da mesma forma, a frequência da célula inferior direita deve ser 70.

Portanto, depois que os totais marginais são estabelecidos, tem-se apenas uma célula livre nessa tabela de dupla entrada. E é por isso que o grau de liberdade para essa tabela de dupla entrada é 1. De modo geral, ao estabelecer os totais marginais, sempre perdemos uma linha e uma coluna, pois a última linha e a última coluna podem ser calculadas através da subtração, o que nos leva à fórmula  $(r - 1) * (c - 1)$ . (Isso foi mais do que você queria saber, não é?)

### ***Descobrimo o comportamento das distribuições do Qui-quadrado***

A Figura 14-2 mostra imagens de distribuições do Qui-quadrado com 1, 2, 4, 6, 8 e 10 graus de liberdade, respectivamente. Aqui, estão alguns dos pontos importantes para se ter em mente quando o assunto é a distribuição do Qui-quadrado:

- ✓ Com um grau de liberdade, a distribuição parece uma hipérbole (veja a Figura 14-2, no canto superior esquerdo); com mais de um grau de liberdade, ela se parece com uma montanha que tem uma longa cauda direita (veja a Figura 14-2, canto inferior direito).
- ✓ Todos os valores são maiores ou iguais a zero.
- ✓ Sua forma é sempre assimétrica à direita (com a cauda indo para a direita).
- ✓ Conforme o número de graus de liberdade aumenta, a média (a média global) aumenta (movendo-se para a direita), e a variância também aumenta (mais dispersão).
- ✓ Independentemente do grau de liberdade, os valores da distribuição do Qui-quadrado (conhecidos como densidade) se aproximam de zero à medida que os valores do Qui-quadrado ficam cada vez maiores. Isso significa que, quanto maiores forem os valores do Qui-quadrado, menor é a possibilidade de que



**Figura 14-2:** Distribuições do Qui-quadrado com 1, 2, 4, 6, 8 e 10 graus de liberdade (do canto superior esquerdo para o canto inferior direito).

### *Usando a tabela da distribuição Qui-quadrada*

Depois de encontrar a estatística de teste do Qui-quadrado e seus graus de liberdade, determine o tamanho de sua estatística, em relação à distribuição correspondente. (Você agora está entrando no sétimo passo para a realização do teste do Qui-quadrado.)

Graficamente falando, o que você deseja encontrar é a probabilidade de estar além (obter um número maior do que) de sua estatística de teste. Se essa probabilidade for pequena, sua estatística de teste do Qui-quadrado é incomum — ela está fora — e você pode rejeitar a  $H_0$ , concluindo, assim, que suas duas variáveis não são independentes (elas, de alguma forma, estão relacionadas).

Caso você esteja me acompanhando, a estatística de teste do Qui-quadrado para os dados independentes da Tabela 14-2 é zero, pois as frequências observadas são iguais às esperadas para cada célula e, portanto, a diferença entre elas é sempre igual a zero. (Porém, este resultado nunca acontece na vida real!) Este cenário representa uma situação *perfeitamente independente* e os menores valores possíveis para uma estatística de teste do Qui-quadrado.

Se a probabilidade de estar à direita de sua estatística de teste do Quiquadrado (em um gráfico) não for pequena o bastante, você não terá provas suficientes para rejeitar a  $H_0$  e, então, deve ficar com ela. Concluindo, assim, que suas duas variáveis são independentes (ou seja, não se relacionam).

O quão pequena essa probabilidade precisa ser para que você possa rejeitar  $H_0$ ? Para a maioria dos testes de hipótese, os estatísticos geralmente usam 0,05 como um limite de corte. (Para obter mais informações sobre os valores de corte, também conhecidos como



níveis  $\alpha$ , volte ao Capítulo 3, ou consulte meu outro livro, *Estatística Para Leigos*, da Alta Books.)

Sua missão agora é encontrar a probabilidade de estar além de sua estatística de teste do Qui-quadrado na distribuição Qui-quadrado correspondente com  $(r - 1) * (c - 1)$  graus de liberdade. Cada distribuição do Qui-quadrado é diferente e, já que o número dos possíveis graus de liberdade é infinito, é impossível mostrar todos os valores de cada distribuição Qui-quadrado.

Na tabela do Qui-quadrado (Tabela A-3 no apêndice), você pode ver alguns dos valores mais importantes em cada distribuição Qui-quadrado com graus de liberdade de 1 a 50.

Para usá-la, encontre a linha que representa os graus de liberdade (abreviado como gl), desloque-se ao longo dessa linha até chegar ao valor mais próximo de sua estatística de teste Qui-quadrado, mas sem ultrapassá-lo. (É como naqueles programas de televisão onde você ganha o prêmio se adivinhar seu preço.)

Então, vá ao topo da coluna em que está. Esse número representa a área à direita (acima) da estatística de teste do Qui-quadrado que você viu na tabela. A área acima de sua estatística de teste em particular é igual ou inferior a esse número. Tal resultado representa o valor- $p$  aproximado para seu teste do Qui-quadrado.

No exemplo da cor de preferência para a pintura da casa (veja a Figura 14-1), a estatística de teste do Qui-quadrado é 14,27, e você tem  $(2 - 1) * (2 - 1) = 1$  grau de liberdade. Na tabela do Qui-quadrado, vá à linha  $gl = 1$ , e procure o número mais próximo de 14,27 (mas não o ultrapasse), que é 7,88.

## ***Tirando conclusões***

Duas são as alternativas que você pode usar para tirar conclusões a partir da estatística de teste do Qui-quadrado. Primeiro, você deve procurar a estatística de teste na tabela do Qui-quadrado e ver a probabilidade de ser maior do que ela, método conhecido como *aproximação do valor- $p$* . (O valor- $p$  de uma estatística de teste é a probabilidade de ser igual ou maior do que sua estatística de teste na distribuição com a qual a estatística de teste estiver sendo comparada — neste caso, com distribuição do Qui-quadrado.) Ou você pode fazer com que o computador calcule o valor- $p$  exato para o teste. (Para uma rápida revisão sobre valores- $p$  e os níveis  $\alpha$ , volte ao Capítulo 3. Para uma revisão mais completa sobre esses tópicos, consulte meu outro livro, *Estatística Para Leigos*, da Alta Books.)

Antes de fazer qualquer coisa, porém, defina  $\alpha$ , a probabilidade de corte para seu valor- $p$ , com antecedência. Se o valor- $p$  for menor do que o nível predeterminado para  $\alpha$ , rejeite a  $H_0$ . Porém, se for maior, não a rejeite.

### ***Aproximando o valor- $p$ da tabela***


No exemplo da cor de preferência para a pintura da casa (veja a Figura 14-1), a estatística

de teste do Qui-quadrado é 14,27 com  $(2 - 1) * (2 - 1) = 1$  gl (grau de liberdade). Na linha um da tabela do Qui-quadrado (ver Tabela A-3 em anexo), o número que mais se aproxima é 7,88 (na última coluna).


O número que aparece na parte superior dessa coluna é 0,005, número que, por sua vez, é menor do que o nível  $\alpha$  normalmente usado (0,05) e, por isso, você deve rejeitar a  $H_0$ . Você sabe que seu valor- $p$  é menor do que 0,005, pois sua estatística de teste foi maior do que 7,88. Em outras palavras, se 7,88 é o mínimo que você precisa para rejeitar a  $H_0$ , o valor de 14,28 lhe fornece provas mais do que suficientes. Isso quer dizer que, quanto menor o valor- $p$ , mais provas você tem contra a  $H_0$ .

No entanto, já que as tabelas do Qui-quadrado, em geral, fornecem apenas alguns valores para cada distribuição do Qui-quadrado, o melhor que você pode dizer usando esta tabela é que o valor- $p$  para este teste é inferior a 0,005.

Mas eis a grande notícia: uma vez que seu valor- $p$  é menor do que 0,05, você pode concluir, com base nesses dados, que sexo e a cor de preferência para a pintura da casa provavelmente se relacionam na população (são dependentes), como afirmou a pesquisa da American Demographics (citada no início deste capítulo). Só agora você tem uma análise estatística formal que diz que o resultado encontrado na amostra também é passível de ocorrer em toda a população. Afirmação muito mais forte!

 Se seus dados demonstrarem que é possível rejeitar a  $H_0$ , você a esse ponto só terá certeza de que as duas variáveis têm alguma relação, pois a estatística de teste do Qui-quadrado não informa qual é essa relação. Para explorar a relação entre as duas variáveis, encontre as probabilidades condicionais de sua tabela de dupla entrada (veja o Capítulo 13). Assim, você poderá usar os resultados para ter uma ideia do que pode estar ocorrendo na população.

Para o exemplo em questão, uma vez que a cor de preferência para a pintura da casa se relaciona ao sexo, você pode examinar ainda mais essa relação, comparar as cores preferidas por homens versus mulheres e descrever como eles se diferenciam. Comece encontrando o percentual de homens que preferem o branco,  $180 \div 500 = 0,36$ , ou 36%, calculado a partir das informações da Tabela 14-1. Agora, compare este resultado ao percentual de mulheres que preferem a cor branca:  $125 \div 500 = 0,25$ , ou 25%. Assim, você pode concluir que nesta população (e não apenas na amostra) a quantidade de homens que preferem cor branca é maior do que a de mulheres. Portanto, as variáveis sexo e cor preferida para a pintura da casa são dependentes.

 As variáveis dependentes afetam os resultados, ou frequências, uma da outra. Se as frequências das células observadas a partir dos dados amostrais não coincidirem com as frequências esperadas, assumindo a  $H_0$ : as variáveis são independentes, deve-se concluir que a relação de dependência encontrada nos dados amostrais se expande à população. Em outras palavras, a observação de grandes diferenças entre as frequências observadas e esperadas significa que as variáveis são dependentes.



## ***Obtendo o valor- $p$ com a ajuda do computador***

Depois que o Minitab calcula a estatística de teste para você, ele lhe fornece o valor- $p$  exato para o teste de hipótese. O valor- $p$  mede a probabilidade de que os resultados encontrados se devam simplesmente ao acaso, enquanto  $H_0$  ainda é considerada uma verdade. Ele informa sua força contra a  $H_0$ . Se o valor- $p$  for 0,001, por exemplo, você teria muito mais força contra a  $H_0$  do que se o valor- $p$  fosse, digamos, 0,10.

Na saída do Minitab na Figura 14-1, o valor- $p$  é dado como 0,000, o que significa que ele é menor do que 0,001; por exemplo, pode ser 0,0009. Isso é o que chamo de valor- $p$  pequeno! (Os resultados do Minitab são dados em apenas três casas decimais, comum em muitos programas estatísticos.)



Já presenciei situações em que pessoas obtiveram resultados que não lhes convinham (como um valor- $p$  de 0,068) e, portanto, fizeram alguns ajustes para conseguir o que queriam. Elas mudaram o nível  $\alpha$  0,05 para 0,10, depois de terem encontrado o resultado. Tal alteração faz com que o valor- $p$  fique menor que o nível  $\alpha$  e eles possam rejeitar a  $H_0$  para dizer que existe uma relação.

Mas o que há de errado nisso? O  $\alpha$  foi alterado depois que eles viram as informações, o que não é permitido. É como mudar a sua aposta no blackjack depois de descobrir as cartas do carteador. (Tentador, mas uma falta gravíssima.) Sempre desconfie de níveis  $\alpha$  elevados, certificando-se de sempre escolher seu  $\alpha$  antes de coletar os dados — e se mantenha fiel a ele.

A boa notícia é que quando valores- $p$  são relatados, qualquer um que os lê poderá tirar sua própria conclusão; nenhuma rejeição ou aceitação é estabelecida como lei. Mas estabelecer um nível  $\alpha$  e, então, alterá-lo depois da coleta dos dados a fim de obter uma conclusão melhor não é nada legal!

## ***Colocando o Qui-quadrado à prova***

Se duas variáveis acabam se mostrando realmente dependentes, você pode descrever a relação entre elas. Mas, se elas forem independentes, os resultados serão os mesmos para os grupos que estão sendo comparados. O exemplo a seguir ilustra essa ideia.

Atualmente nos Estados Unidos há muita especulação e debate sobre a proibição ou não do uso do celular enquanto se dirige. O que lhe interessa aqui são as opiniões dos americanos sobre esse assunto, mas você também suspeita que os resultados podem se diferenciar dependendo do sexo do respondente (masculino ou feminino). Sendo assim, você decide fazer um teste do Qui-quadrado de independência para ver se sua teoria está correta. A Tabela 14-3 mostra uma tabela de dupla entrada para os dados observados a partir de uma amostra composta por 60 homens e 60 mulheres sobre se concordam ou não com tal política (a proibição do uso do celular na direção). Na Tabela 14-3, vemos que  $12 \div 60 = 20\%$  dos homens concordam com a política de proibição dos celulares na direção,



comparados a  $9 \div 60 = 15\%$  das mulheres. Vemos que esses percentuais são diferentes, mas será que isso é suficiente para dizer que o sexo e a opinião sobre este assunto são dependentes? Só um teste do Qui-quadrado para independência poderá ajudá-lo a decidir.

**Tabela 14-3** **Sexo e Opinião sobre a Proibição do Celular: Frequências Observadas**

	<i>Concorda com a Proibição</i>	<i>Discorda da Proibição</i>	<i>Total Marginal da Linha</i>
Homens	12	48	60
Mulheres	9	51	60
<b>Total Marginal das Colunas</b>	21	99	120 ( <i>Total Geral</i> )

Assumindo a hipótese nula como verdadeira, a Tabela 14-4 mostra as frequências esperadas para as células com seu cálculo.

**Tabela 14-4** **Sexo e Opinião sobre a Proibição do Celular: Frequências Esperadas**

	<i>Concorda com a Proibição</i>	<i>Discorda da Proibição</i>	<i>Total Marginal da Linha</i>
Homens	$(60 * 21) \div 120 = 10,5$	$(60 * 99) \div 120 = 49,5$	60
Mulheres	$(60 * 21) \div 120 = 10,5$	$(60 * 99) \div 120 = 49,5$	60
<b>Total Marginal das Colunas</b>	21	99	120 ( <i>Total Geral</i> )

Se executarmos um teste do Qui-quadrado no Minitab usando estes dados, os graus de liberdade serão iguais a  $(2 - 1) * (2 - 1) = 1$ , a estatística de teste aparece como sendo 0,519 e o valor- $p$  é 0,471. Como o valor- $p$  é maior do que 0,05 (o valor de corte mais comum), você não pode rejeitar a  $H_0$ , portanto, deve concluir que a variável sexo e a opinião sobre a proibição de celulares na direção são independentes e, portanto, não relacionadas. Sua teoria de que o sexo poderia influenciar a opinião não foi para frente, afinal, não há provas suficientes que mostrem o contrário.

Em geral, *independência* significa que você pode não encontrar nenhuma diferença importante na aparência das linhas à medida que se move de cima para baixo na coluna. Dito de outra forma, a proporção dos dados classificados em cada coluna ao longo de uma linha é quase igual para cada linha. Uma vez que a Tabela 14-4 tem o mesmo número de homens e mulheres, o total das duas linhas é igual e, portanto, você obtém a mesma frequência esperada para homens e mulheres, tanto na coluna Concorda (10,5) quanto na coluna Discorda (49,5).



# ***Comparando Dois Testes para Comparar Duas Proporções***

Você pode usar o teste do Qui-quadrado para verificar se duas proporções populacionais são iguais. Por exemplo, a proporção de usuários de celulares do sexo feminino é igual à proporção de usuários de celular do sexo masculino?

Talvez, você esteja pensando: "Mas, espere um minuto, os estatísticos já não têm um teste para duas proporções? Acho que me lembro de ter visto isso nas aulas de Estatística I... Deixe-me lembrar... Sim, é o teste-Z para duas proporções. Mas o que aquele teste tem a ver com o teste do Qui-quadrado?" Nesta seção, você vai obter a resposta a essa pergunta, além de praticar utilizando os dois métodos para investigar uma possível diferença quanto ao uso do celular relacionado ao sexo.

## ***Refamiliarizando-se com o teste-Z para duas proporções populacionais***

A maneira que a maioria das pessoas usa para testar a igualdade entre duas proporções da população é a utilização de um *teste-Z para duas proporções populacionais*. Com este teste você coleta uma amostra aleatória de cada uma das duas populações, encontra, subtrai as duas proporções amostrais e divide a diferença pelo erro padrão (consulte um livro de Estatística I para mais detalhes sobre este teste).

É possível realizar esse teste desde que o tamanho das amostras das duas populações sejam grandes — pelo menos cinco sucessos e cinco falhas em cada amostra.

A hipótese nula do teste-Z para duas proporções de população é  $H_0: p_1 = p_2$ , onde  $p_1$  é a proporção da primeira população que se enquadra na categoria de interesse e  $p_2$  é a proporção da segunda população enquadrada na categoria de interesse. E, como sempre, a hipótese alternativa é uma das opções a seguir,  $H_a$ : não é igual, maior ou menor.

Suponha que você queira comparar a proporção de usuários de celulares do sexo masculino (M) versus feminino (F), onde  $p_1$  é a proporção de homens que possuem um celular e  $p_2$  é a proporção de todas as mulheres que possuem um celular. Sendo assim, você coleta os dados, encontra as proporções da amostra para cada grupo, subtrai as duas e calcula a estatística-Z usando a fórmula

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ onde } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Aqui,  $x_1$  e  $x_2$  são o número de indivíduos das amostras um e dois, respectivamente, com a característica desejada;  $n_1$  e  $n_2$  são os dois tamanhos amostrais.

Suponha que ao coletar os dados de 100 homens e 100 mulheres, você descobre que entre os homens, 45 possuem celular, enquanto entre as mulheres, 55 possuem celular. Isso

significa que  $\hat{p}_1$  equivale a  $45 \div 100 = 0,45$  e  $\hat{p}_2$  a  $55 \div 100 = 0,55$ . Suas amostras têm pelo menos cinco sucessos (cinco indivíduos com a característica desejada, que, neste caso, é a posse de celular) e cinco falhas (cinco indivíduos que não têm a característica desejada). Portanto, calcule a estatística-Z comparando as duas proporções da população (homens versus mulheres) com base nesses dados, cujo valor é  $-1,41$ , como mostrado na última linha da saída do Minitab na Figura 14-3.

Test Cellphone for Two Proportions			
Sample	X	N	Sample p
M	45	100	0.450000
F	55	100	0.550000
Difference = p (1) - p (2)			
Estimate for difference: -0.1			
95% CI for difference: (-0.237896, 0.0378957)			
Test for difference = 0 (vs not = 0): Z = -1.41 P-Value = 0.157			

**Figura 14-3:** Saída do Minitab para comparação entre a proporção de homens e mulheres que possuem celular.

O valor- $p$  para a estatística de teste- $Z = -1,41$  é 0,157 (calculado pelo Minitab, ou observado na área abaixo do valor- $Z$  de  $-1,41$  em uma tabela- $Z$ , que você deve ter em seu livro de Estatística I). Esse valor- $p$  (0,157) é maior do que o nível  $\alpha$  (valor de corte predeterminado) de 0,05, portanto, não é possível rejeitar a  $H_0$ , ou seja, não se pode afirmar que as duas proporções populacionais não são iguais. Logo, deve-se concluir que a proporção de proprietários de celulares do sexo masculino não é diferente da proporção feminina.

Ainda que a amostra parecesse ter evidências de uma diferença (afinal, 45% não é igual a 55%), tais evidências não foram suficientes para dizer que esta mesma diferença se transfere à população. Sendo assim, você não pode afirmar que existe uma diferença entre os sexos com relação ao uso do celular, pelo menos não com esta amostra.

## ***Igualando os testes do Qui-quadrado e testes-Z para uma tabela dois por dois***

Aqui está o segredo de como relacionar o teste- $Z$  com um teste do Quiquadrado para independência. O teste- $Z$  para duas proporções e o teste do Qui-quadrado para independência em uma tabela dois por dois (duas linhas e duas colunas) são equivalentes se os tamanhos amostrais das duas populações forem grandes o suficiente — isto é, quando o número de sucessos e o número de falhas em cada célula das duas amostras for pelo menos igual a cinco.

Se você usar o teste- $Z$  para ver se a proporção de proprietários de celulares do sexo masculino é igual à proporção de proprietários de telefones celulares do sexo feminino,

estará, na verdade, analisando para ver se pode esperar a mesma proporção de proprietários de celular para ambos os sexos (depois de levar em conta os tamanhos amostrais), o que de fato significa que você está testando se a variável sexo (masculino ou feminino) é independente da variável posse de celular (sim ou não).

Se a proporção de proprietários de celulares do sexo feminino é igual à proporção de proprietários de celulares do sexo masculino, a proporção de proprietários de celulares é a mesma independentemente do sexo e, assim, a variável sexo e posse de celular são independentes. Se, por outro lado, a proporção de proprietários de celulares do sexo feminino for diferente da proporção de proprietários de celulares do sexo masculino, pode-se dizer que o uso de celulares se diferencia entre os sexos e, portanto, o sexo e a posse de celular são dependentes.

Nos dados coletados, vimos que 45 homens usam celular (de um total de 100 homens) e 55 mulheres usam celular (de um total de 100 mulheres). A saída do Minitab para o teste do Qui-quadrado para independência (completado as frequências observadas e esperadas, os graus de liberdade, a estatística de teste e valor- $p$ ) está na Figura 14-4. O valor- $p$  para esse teste é 0,157, valor maior do que o nível  $\alpha$  (geralmente, igual a 0,05) e, portanto, não é possível rejeitar a  $H_0$ .

Já que o teste do Qui-quadrado para independência e o teste- $Z$  são equivalentes quando se tem uma tabela de dupla entrada, o valor- $p$  do teste do Qui-quadrado para independência é idêntico ao valor- $p$  do teste- $Z$  para duas proporções. Se você comparar o valores- $p$  das Figuras 14-3 e 14-4, poderá comprovar isso por si mesmo.

Chi-Square Test Gender, Cellphone			
Expected counts are printed below observed counts			
Chi-Square contributions are printed below expected counts			
	Y	N	Total
M	45	55	100
	50.00	50.00	
	0.500	0.500	
F	55	45	100
	50.00	50.00	
	0.500	0.500	
Total	100	100	200
Chi-Sq = 2.000, DF = 1, P-Value = 0.157			

**Figura 14-4:** Saída do Minitab testando a independência entre as variáveis sexo e posse de celular.

Os pesquisadores têm realizado grandes estudos sobre os efeitos do uso do celular enquanto se dirige. Um estudo publicado no *New England Journal of Medicine* observou e registrou dados em 1997 sobre 699 motoristas que tinham celulares e se envolveram em colisões no trânsito, resultando em consideráveis danos materiais, mas nenhuma lesão corporal. As ligações dos celulares de cada pessoa realizadas no dia da colisão e durante a semana anterior foram analisadas através da utilização de registros de cobrança detalhados. Um total de 26.798 ligações foram feitas durante os 14 meses, o período do estudo.

Uma das conclusões a qual os pesquisadores chegaram foi que “... o risco de bater o carro quando se está usando o celular é quatro vezes maior do que quando o celular não está sendo usado”. Basicamente, o que eles fizeram foi realizar um teste do Qui-quadrado para verificar se o uso do celular e o fato de se envolver em uma colisão eram independentes e quando descobriram que não eram, foram capazes de examinar mais a fundo essa relação usando as proporções adequadas. Em particular, descobriram que o risco de uma colisão é quatro vezes maior para os motoristas que usam os celulares do que para aqueles que não o fazem.

Além disso, também descobriram que o risco relativo foi semelhante para os motoristas que se diferenciavam em características pessoais, como idade e experiência. (Essa descoberta significa que eles realizaram testes semelhantes para ver se os resultados eram os mesmos para motoristas de diferentes faixas etárias e com diferentes níveis de experiência, e os resultados foram sempre os mesmos. Portanto, a idade e a experiência do motorista não se relacionavam ao resultado da colisão.)

A pesquisa também mostra que “Chamadas feitas próximas da hora da colisão foram consideradas particularmente perigosas ( $p < 0,001$ ). O uso de acessórios *hands-free* não ofereceu nenhuma vantagem de segurança (valor- $p$  não significativo)”. **Observação:** os itens entre parênteses indicam a maneira normalmente usada pelos pesquisadores para relatar seus resultados: usar os valores- $p$ . O  $p$  nos dois parênteses representa o valor  $p$  de cada teste.

No primeiro caso, o valor- $p$  é muito pequeno, menor do que 0,001, indicando uma forte evidência de uma relação entre as colisões e o uso do celular naquele momento. O segundo valor- $p$  entre parênteses foi indicado por ser insignificante, o que significa que foi substancialmente maior do que 0,05, o nível  $\alpha$  mais utilizado. Esse segundo resultado indica que o uso de acessórios *hands-free* não afetou as chances de que uma colisão ocorresse, tanto a proporção de colisões ocorridas quando o motorista usava esses acessórios quanto a proporção ocorrida quando o motorista não os usava foram consideradas estatisticamente iguais (ou seja, poderiam ter ocorrido por acaso sob a hipótese de independência). Independentemente de usar um celular comum ou um com acessórios *hands-free*, espero que este estudo lhe sirva de lição!

Além disso, observe que se você pegar a estatística do teste- $Z$  para este exemplo (da Figura 14-3), cujo valor é  $-1,41$ , e a elevar ao quadrado, o resultado obtido será 2,00, que é igual à estatística de teste do Quiquadrado para os mesmos dados (última linha da Figura 14-4). O caso é que o quadrado da estatística do teste- $Z$  (quando se tratar do teste para a igualdade de duas proporções) é igual à estatística de teste do Qui-quadrado para independência.

O teste do Qui-quadrado e o teste- $Z$  são equivalentes apenas quando se trata de uma tabela dois por dois (duas linhas e duas colunas) e quando o teste- $Z$  é bicaudal (a hipótese alternativa é a de que as duas proporções não são iguais, em vez de usar a  $H_a$ : uma



proporção é maior ou menor do que a outra). Se o teste- $Z$  não for bicaudal, o uso do teste do Qui-quadrado passa a não ser uma técnica adequada. Se a tabela de dupla entrada tiver mais do que duas linhas e duas colunas, utilize o teste do Qui-quadrado para independência (pois uma grande quantidade de categorias significa que não se tratam mais de apenas duas proporções e, assim, o teste- $Z$  não pode ser aplicado).

---

<sup>1</sup> N.E.: O Minitab é um programa em inglês, com as entradas em inglês e, para fins de aprendizagem, colocamos os equivalentes em inglês dos termos abordados nas primeiras ocorrências.

## Capítulo 15

# Usando os Testes do Qui-quadrado para Qualidade de Ajuste (dos Dados, e Não de Seu Jeans)

---

### *Neste Capítulo*

- ▶ Entendendo o que realmente significa a qualidade de ajuste
  - ▶ Usando o modelo do Qui-quadrado para testar a qualidade de ajuste
  - ▶ Verificando as condições para os testes de qualidade de ajuste
- 

**M**uitos fenômenos da vida podem parecer casuais em um primeiro momento, mas, a longo prazo, percebemos que eles realmente ocorrem de acordo com algum modelo preconcebido, pré-selecionado ou preestabelecido. Por exemplo, mesmo que você não saiba se vai chover amanhã, o meteorologista local pode lhe dar um modelo para a porcentagem de dias em que chove, neva, faz sol ou fica nublado, com base nos últimos cinco anos. Se este modelo continua sendo relevante para esse ano, ninguém sabe, no entanto, ainda é um modelo. Para citar outro exemplo, um biólogo pode produzir um modelo para prever o número de filhotes que um casal de gansos pode ter por ano, mesmo que você não tenha ideia de quantos filhotes o casal que vive em seu quintal vai ter. Será que o modelo dele está correto? Aqui está a sua chance de descobrir.

Neste capítulo, você vai construir modelos para a proporção de resultados que se enquadram em cada categoria de uma variável categórica. Você, então, vai testar esses modelos, coletando dados e comparando o que observa em seus dados com o que espera do modelo. Para fazer essa avaliação, vai usar um teste de qualidade de ajuste baseado na distribuição do Qui-quadrado. De certa forma, um teste de qualidade de ajuste é comparado a uma verificação da realidade de um modelo construído para dados categóricos.

# Encontrando a Estatística de Qualidade de Ajuste

A ideia geral de um procedimento de *qualidade de ajuste* consiste em determinar o que se espera encontrar e compará-lo com o que realmente se pode observar na amostra através da utilização de uma estatística de teste. Esta estatística de teste é chamada de *estatística de teste de qualidade de ajuste*, pois mede o quanto seu modelo (o que você esperava) se ajusta aos dados reais (o que você observou).

Nesta seção, você vai ver como calcular os números esperados para cada categoria, segundo o modelo proposto, e também como juntá-los aos valores observados para dar forma à estatística de teste para a qualidade de ajuste.

## O observado versus o esperado

Para conseguir um exemplo do que pode ser observado versus o que pode ser esperado, você não precisa ir tão longe, basta examinar um pacote dos deliciosos M&M's de chocolate ao leite. Existe uma infinidade de diferentes tipos de M&M's, e cada um tem sua própria variação de cores e sabores. Neste estudo, no entanto, vou me referir ao M&M's de chocolate ao leite — o meu favorito.

A porcentagem de cada cor do M&M's que deve aparecer em um pacote é algo muito bem pensado pela Mars (a empresa que fabrica o M&M's). A Mars requer porcentagens específicas de cada cor nos pacotes de M&M's, as quais são determinadas por meio de grandes pesquisas de mercado baseadas no que as pessoas gostam e desejam ver. Assim, a Mars sempre posta em seu site as atuais porcentagens para cada cor do M&M's. A Tabela 15-1 mostra as porcentagens de cada cor do M&M's em 2006.

Tabela 15-1 Porcentagem Esperada para Cada Cor do M&M's de Chocolate ao Leite (2006)	
Cor	Porcentagem
Marrom	13%
Amarelo	14%
Vermelho	13%
Azul	24%
Laranja	20%
Verde	16%

Agora que você já sabe o que esperar em um pacote de M&M's, a próxima pergunta é: como a Mars consegue fazer isso? Se você abrisse um pacote de M&M's agora, será que encontraria aquelas porcentagens para cada cor? Por causa de seus estudos anteriores em Estatística, você sabe que os resultados amostrais variam (para uma rápida revisão sobre essa ideia, consulte o Capítulo 3). Por isso, não pode esperar que cada pacote de M&M's



tenha o número exato de cada cor dos M&M's, segundo a informação contida na Tabela 15-1. No entanto, a fim de manter seus clientes felizes, a Mars deve chegar o mais próximo possível daquelas expectativas. Mas como determinar o quão próximo a empresa conseguiu chegar?

A Tabela 15-1 lhe diz que as porcentagens devem se enquadrar em cada categoria de toda a população dos M&M's (ou seja, cada um dos M&M's de chocolate ao leite que está sendo fabricado). Esse conjunto de porcentagens é chamado de *modelo esperado* para os dados. O que você deseja ver é se os percentuais no modelo esperado estão realmente ocorrendo nos pacotes que você compra. Para iniciar este processo, colete uma amostra de M&M's (afinal, não é possível verificar um por um de toda a população) e faça uma tabela mostrando as porcentagens de cada cor observada. Depois, compare a tabela das porcentagens observadas com o modelo esperado.

Algumas porcentagens esperadas são conhecidas, como no caso do M&M's, mas você também pode calculá-las através de técnicas matemáticas. Por exemplo, se estiver examinando um único dado para determinar se ele está ou não viciado, saiba que, se ele não estiver viciado, você deve esperar que  $\frac{1}{6}$  dos resultados se enquadrem em cada uma das categorias de 1, 2, 3, 4, 5 e 6.

Como exemplo, examinei um pacote de 48 gramas de M&M's de chocolate ao leite (foi um trabalho difícil, mas alguém tinha que fazê-lo) e coloquei meus resultados na Tabela 15-2, coluna dois. (Pense neste pacote como uma amostra aleatória de 56 M&M's, embora, tecnicamente, não seja o mesmo que chegar a um silo cheio de M&M's e retirar uma amostra realmente aleatória de 48 gramas. Mas, para nosso exemplo, um pacote é suficiente.)

Tabela 15-2 Porcentagem de M&M's Observada em um Pacote (48 gramas) versus Porcentagem Esperada		
Cor	Porcentagem Observada	Porcentagem Esperada
Marrom	$\frac{4}{56} = 7,14$	13,00
Amarelo	$\frac{10}{56} = 17,86$	14,00
Vermelho	$\frac{4}{56} = 7,14$	13,00
Azul	$\frac{10}{56} = 17,86$	24,00
Laranja	$\frac{15}{56} = 26,79$	20,00
Verde	$\frac{13}{56} = 23,21$	16,00
TOTAL	100,00	100,00

Compare o que observei na amostra (coluna 2 da Tabela 15-2) e o que esperava obter (coluna 3 da Tabela 15-2). Repare que observei uma porcentagem menor de M&M's marrons, vermelhos e azuis do que a esperada. Mas também observei uma porcentagem maior de M&M's laranjas, amarelos e verdes do que a esperada. Porém, como os resultados amostrais variam ao acaso, de amostra para amostra, talvez a diferença que

observei possa ter sido causada por essa variação ao acaso. Mas será que essas diferenças poderiam indicar que a porcentagem esperada relatada pela Mars não está sendo cumprida?

É lógico que, se as diferenças entre o observado e o esperado forem pequenas, você deve atribuí-las ao acaso e manter o modelo esperado. Por outro lado, se as diferenças entre o observado e o esperado forem grandes o suficiente, talvez você tenha provas o bastante para indicar que o modelo esperado tem problemas. Mas por qual das conclusões optar? A frase de ordem é: “se as diferenças forem grandes o suficiente”. Portanto, é preciso quantificar o termo “*grandes o suficiente*”, e, para fazer isso, vamos precisar de um pouco mais de maquinário, que abordo na próxima seção.

## ***Calculando a estatística de qualidade de ajuste***

A estatística de qualidade de ajuste é um número que reúne o montante total da diferença entre o esperado para cada célula em relação ao observado. O termo *célula* é usado para expressar cada uma das categorias dentro de uma tabela. Por exemplo, no caso dos M&M’s, as primeiras colunas das Tabelas 15-1 e 15-2 contêm seis células, uma para cada cor dos M&M’s. Para qualquer célula, o número de itens observados para determinada célula é chamado de *frequência observada da célula*. Já o número de itens esperados para determinada célula (segundo o modelo) recebe o nome de *frequência esperada da célula*. A frequência esperada da célula é obtida através da multiplicação desta pelo tamanho da amostra.

A frequência esperada da célula é apenas uma proporção do todo e, por isso, não precisa ser um número inteiro. Por exemplo, se você lançar um dado não viciado por 200 vezes, deverá esperar que ele caia com a face 1 voltada para cima em  $\frac{1}{6}$ , ou 16,67% das vezes. Assim, o número de vezes que se espera obter a face 1 deve ser  $0,1667 * 200 = 33,33$ . Use 33,33 em seus cálculos para a qualidade de ajuste; não o arredonde para um número inteiro, pois, dessa forma, sua resposta final será mais precisa.

A razão pela qual a estatística de qualidade de ajuste se baseia no *número* que está em cada célula, em vez de usar a *porcentagem* presente em cada célula, se dá pelo fato de que as porcentagens podem ser um pouco ilusórias. Se você sabe que 8 em cada 10 pessoas apoiam determinada opinião, isso equivale a 80%. Mas 80 de 100 também equivale a 80%. Qual dos exemplos acima você acha que é uma estatística mais precisa? O segundo, pois ele utiliza mais informações. Se você usar apenas as porcentagens, não vai levar em conta o tamanho da amostra, porém, se usar a frequência (o número em cada grupo), vai poder controlar a quantidade de precisão que possui.

Por exemplo, no lançamento de um dado não viciado, espera-se que a porcentagem de que ele caia com a face 1 voltada para cima seja de  $\frac{1}{6}$ . Se esse dado for lançado 600 vezes, o número esperado para 1 será  $\frac{1}{6} * 600 = 100$ . Esse número (100) é a frequência esperada para a célula que representa o resultado para a categoria 1. Se este mesmo dado for



lançado 600 vezes e cair 95 vezes com a face 1 para cima, então, 95 é a frequência observada para essa célula.

A fórmula para a estatística de qualidade de ajuste é dada da seguinte forma:  $\sum_{\text{todas as células}} \frac{(O-E)^2}{E}$ , onde  $E$  é o número esperado, e  $O$  é o número observado em uma célula. Abaixo, veja os passos para esse cálculo:

1. Encontre o número esperado para a primeira célula ( $E$ ), multiplicando a porcentagem esperada para essa célula pelo tamanho amostral.
2. Faça a diferença entre o valor observado na primeira célula ( $O$ ) e o número de itens esperados para ela ( $E$ ).
3. Eleve essa diferença ao quadrado.
4. Divida a resposta pelo número esperado para a célula ( $E$ ).
5. Repita do primeiro ao quarto passos para cada célula.
6. Some os resultados para obter a estatística de qualidade de ajuste.



A razão para dividir pela frequência esperada na estatística de qualidade de ajuste (quarto passo) é levar em conta a grandeza das diferenças encontradas. Por exemplo, se você esperava que 100 itens se enquadrassem a uma determinada célula, mas obteve 95, a diferença é 5. No entanto, em termos de uma porcentagem, essa diferença é de apenas  $\frac{5}{100} = 5\%$ . Entretanto, se você esperava que 10 itens se enquadrassem nessa célula, mas observou 5 itens, a diferença ainda é 5, mas, em termos de porcentagem, ela é igual a  $\frac{5}{10} = 50\%$ . Assim, em se tratando do impacto, tal diferença é muito maior. A estatística da qualidade de ajuste funciona como a diferença percentual, o único elemento a mais é elevar a diferença ao quadrado a fim de torná-la positiva. (Isso se faz necessário, pois, se você espera 10 e obtém 15 ou espera 10 e obtém 5, não faz diferença nenhuma para os outros; você ainda estará 50% enganado.)

A Tabela 15-3 mostra o passo a passo do cálculo da estatística da qualidade de ajuste para o exemplo dos M&M's, onde  $O$  indica a frequência observada e  $E$  indica a esperada. Para obter as frequências esperadas, multiplique as porcentagens esperadas, mostradas na Tabela 15-1, por 56, pois 56 é o número de M&M's que havia em minha amostra. As frequências observadas são as encontradas na minha amostra e estão na Tabela 15-2.

**Tabela 15-3 Estatística da Qualidade de Ajuste para Exemplo dos M&M's**

<i>Cor</i>	<i>O</i>	<i>E</i>	<i>O - E</i>	<i>(O - E)<sup>2</sup></i>	$\frac{(O-E)^2}{E}$
Marrom	4	0,13 * 56 = 7,28	4 - 7,28 = -3,28	10,76	1,48
Amarelo	10	0,14 * 56 = 7,84	10 - 7,84 = 2,16	4,67	0,60
Vermelho	4	0,13 * 56 = 7,28	4 - 7,28 = -3,28	10,76	1,48
Azul	10	0,24 * 56 = 13,44	10 - 13,44 = -3,44	11,83	0,88

Laranja	15	$0,20 * 56 = 11,20$	$15 - 11,20 = 3,80$	14,44	1,29
Verde	13	$0,16 * 56 = 8,96$	$13 - 8,96 = 4,04$	16,32	1,82
TOTAL	56	56			7,55

A estatística da qualidade de ajuste para o exemplo dos M&M's acaba sendo 7,55, o número em negrito no canto inferior direito da Tabela 15-3. Este número representa o quadrado da diferença total que eu esperava e observei ajustado para a grandeza de cada frequência esperada. A próxima questão é como interpretar esse valor de 7,55. Será que ele é grande o suficiente para indicar que as cores dos M&M's do pacote não estão seguindo as porcentagens enviadas pela Mars? A próxima seção aborda como fazer estes resultados terem algum sentido.

# *Interpretando a Estatística da Qualidade de Ajuste Através do Qui-quadrado*

Depois de obter a estatística da qualidade de ajuste, sua próxima tarefa é interpretá-la. Para isso, é preciso descobrir os valores que você poderia ter obtido e o lugar ocupado por sua estatística entre eles. Para realizar essa tarefa, você pode usar o teste Qui-quadrado de qualidade de ajuste.

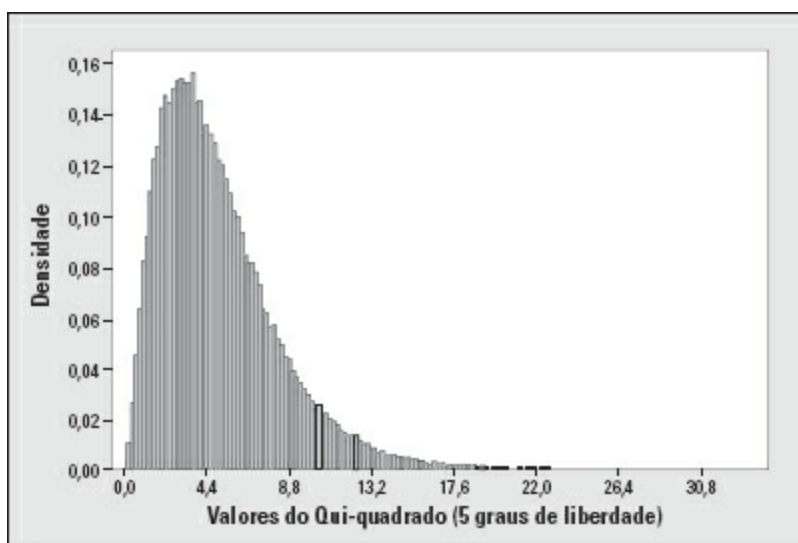
Os valores de uma estatística da qualidade de ajuste realmente seguem uma distribuição Qui-quadrado com  $k - 1$  graus de liberdade, onde  $k$  é o número de categorias em sua população. (Veja o Capítulo 14 para os detalhes completos sobre o Qui-quadrado.) Você também deve usar a tabela do Quiquadrado (Tabela A-3 em anexo) para encontrar o valor- $p$  de sua estatística do Qui-quadrado.

Se a estatística de qualidade de ajuste do Qui-quadrado for grande o suficiente, conclui-se que o modelo original não se ajusta e você deve jogá-lo fora; há muita diferença entre o que você observou e o que era esperado segundo o modelo. No entanto, se a estatística da qualidade de ajuste for relativamente pequena, você não deve rejeitar o modelo (um valor grande ou pequeno para uma estatística de teste do Qui-quadrado depende dos graus de liberdade).



A estatística da qualidade de ajuste segue as principais características da distribuição do Qui-quadrado. O menor valor possível da estatística de qualidade de ajuste é zero. Continuando o exemplo da seção anterior, se os M&M's em minha amostra seguissem as porcentagens exatas, encontradas na Tabela 15-1, a estatística da qualidade de ajuste seria zero. Isso porque as frequências observadas e as esperadas seriam iguais e, assim, a diferença entre os valores observados e os esperados seria zero.

O maior valor possível para o Qui-quadrado não é especificado, embora a ocorrência de alguns valores seja mais provável do que a de outros. Cada distribuição do Qui-quadrado tem seu próprio conjunto de possíveis valores, como você pode ver na Figura 15-1. Esta figura mostra uma distribuição simulada do Qui-quadrado com  $6 - 1 = 5$  graus de liberdade (relevantes ao exemplo dos M&M's). Ela basicamente decompõe todos os possíveis valores para a estatística de qualidade de ajuste nesta situação e mostra a frequência com que eles ocorrem. Você pode ver na Figura 15-1 que uma estatística de teste do Qui-quadrado igual a 7,55 não é tão difícil de ocorrer, indicando que o modelo para as cores dos M&M's provavelmente não pode ser rejeitado. No entanto, são necessárias mais informações antes de formalizarmos essa conclusão.



**Figura 15-1:** Distribuição do Qui-quadrado com 5 graus de liberdade.

## *Verificando as condições antes de começar*

Toda técnica estatística parece ter uma pegadinha, e esta não foge à regra. Para utilizar a distribuição do Qui-quadrado na interpretação da estatística da qualidade de ajuste, você precisa ter certeza de que tem informação suficiente para trabalhar com cada célula. Os gurus da Estatística costumam recomendar que a frequência esperada para cada célula seja maior ou igual a cinco. Se não, uma opção é combinar as categorias a fim de aumentar esses números.

No exemplo dos M&M's, todas as frequências esperadas são maiores do que sete (consulte a Tabela 15-3), sendo assim, essa condição está satisfeita. Mas, se este não fosse o caso, isso indica que você deveria ter coletado uma amostra maior, pois a frequência esperada é calculada pela multiplicação desta pelo tamanho da amostra. Se o tamanho da amostra aumenta, a frequência esperada também aumenta. Além disso, uma amostra maior também aumenta suas chances de detectar um desvio real do modelo. Tal ideia está relacionada ao conceito do poder do teste (veja o Capítulo 3 para obter mais informações sobre o assunto).



Depois de ter coletado os dados, não é permitido voltar e coletar uma nova amostra. Por isso, é melhor estabelecer o tamanho amostral correto antes de começar e, para fazer isso, você deve determinar o tamanho amostral necessário para obter uma frequência esperada igual a pelo menos cinco. Por exemplo, no lançamento de um dado não viciado, espera-se que a porcentagem de que ele caia com a face 1 voltada para cima seja de  $\frac{1}{6}$ . Se você coletar uma amostra de apenas seis lançamentos, a frequência esperada será  $\frac{1}{6} * 6 = 1$ , o que não é suficiente. No entanto, se o dado for lançado 30 vezes, a frequência esperada será  $\frac{1}{6} * 30 = 5$ , justamente o suficiente para satisfazer a condição.

## *Os passos para o teste Qui-quadrado de qualidade de ajuste*

Assumindo que a condição necessária seja atendida (consulte a seção anterior), você pode começar a, de fato, realizar um teste de qualidade de ajuste formal.

A versão geral da hipótese nula para o teste de qualidade de ajuste é  $H_0$ : o modelo é válido para todas as categorias; versus a hipótese alternativa  $H_a$ : o modelo não é válido para, pelo menos, uma categoria. Cada situação vai ditar as proporções que devem ser listadas na  $H_0$  para cada categoria. Por exemplo, se você está lançando um dado não viciado, tem-se a  $H_0$ : Proporção de  $1 = \frac{1}{6}$ ; proporção de  $2 = \frac{1}{6}$ ; ... ; proporção de  $6 = \frac{1}{6}$ .

A seguir, veja os passos para o teste Qui-quadrado de qualidade de ajuste, sendo que o exemplo dos M&M's vai ilustrar como realizar cada passo:

**1. Escreva a  $H_0$  usando as porcentagens esperadas em seu modelo para cada categoria.**

Usando um subscrito para indicar a proporção ( $p$ ) de M&M's que você espera se enquadrar em cada categoria (veja a Tabela 15-1), a sua hipótese nula é  $H_0: p_{\text{marrom}} = 0,13, p_{\text{amarelo}} = 0,14, p_{\text{vermelho}} = 0,13, p_{\text{azul}} = 0,24, p_{\text{laranja}} = 0,20, \text{ e } p_{\text{verde}} = 0,16$ . Todas essas proporções devem permanecer em ordem para que o modelo seja válido.

**2. Escreva sua  $H_a$ : este modelo não é válido para pelo menos uma categoria.**

Sua hipótese alternativa,  $H_a$ , neste caso, seria: uma (ou mais) das probabilidades dadas em  $H_0$  não está(ão) correta(s). Em outras palavras, conclui-se que pelo menos uma das cores dos M&M's possui uma proporção diferente do que a declarada no modelo.

**3. Calcule a estatística da qualidade de ajuste usando o passo a passo da seção anterior, “Calculando a estatística da qualidade de ajuste”.**

A estatística de qualidade de ajuste para os M&M's, da seção anterior, é 7,55. Só para lembrar, faça a diferença entre a frequência observada e a esperada da célula em questão, eleve-a ao quadrado e a divida pela frequência esperada para a célula em questão. Repita esse procedimento para cada célula da tabela e, depois, some os resultados. No exemplo dos M&M's, esse total é igual a 7,55, a estatística da qualidade de ajuste.

**4. Consulte a distribuição do Qui-quadrado com  $k - 1$  graus de liberdade, onde  $k$  é o número de categorias que você tem.**

Compare essa estatística (7,55) com a distribuição do Qui-quadrado com  $6 - 1 = 5$  graus de liberdade (pois  $k = 6$  possíveis cores para os M&M's). (Consulte a Tabela A-3 no apêndice.)

Observe a Figura 15-1. Você pode notar que o valor 7,55 está bem afastado da extremidade da distribuição. Sendo assim, é provável que você não tenha evidências suficientes para rejeitar o modelo fornecido pela Mars para as cores dos M&M's.

**5. Encontre o valor- $p$  de sua estatística de qualidade de ajuste.**

Utilize uma tabela do Qui-quadrado para encontrar o valor- $p$  de sua estatística de teste



(consulte a Tabela A-3 no apêndice). (Para mais informações sobre a distribuição do Qui-quadrado, consulte o Capítulo 14.)

Uma vez que a tabela do Qui-quadrado fornece apenas um determinado número de resultados para cada grau de liberdade, o valor- $p$  exato para sua estatística de teste pode se enquadrar entre dois valores- $p$  mostrados na tabela.

Para encontrar o valor- $p$  da estatística de teste no exemplo dos M&M's (7,55), localize a linha referente a 5 graus de liberdade na tabela do Qui-quadrado (Tabela A-3 do apêndice) e analise os números (o cálculo para os graus de liberdade é  $k - 1 = 6 - 1 = 5$ , onde  $k$  é o número de categorias). Você vai notar que o número 7,55 é menor do que o primeiro valor da linha (9,24), cujo valor  $p$  é 0,10. (O valor- $p$  é o número no título da coluna.) Assim, o valor- $p$  7,55, que é a área à direita de 7,55 na Figura 15-1, deve ser superior a 0,10, pois 7,55 está à esquerda de 9,24 nessa distribuição do Qui-quadrado.

Existem muitos programas de computador (online ou em calculadoras gráficas) que encontram os valores- $p$  exatos para o teste do Quiquadrado, poupando tempo e evitando dores de cabeça. Usando uma calculadora online para o valor- $p$ , descobri que o valor- $p$  exato para o teste de qualidade de ajuste para o exemplo dos M&M's (cuja estatística de teste é 7,55 com 5 graus de liberdade para o Qui-quadrado) é 0,1828. Para encontrar uma calculadora online para o valor- $p$ , basta digitar o nome da distribuição e a palavra “valor- $p$ ” em um site de busca na Internet. Para este exemplo, pesquise “valor- $p$  do Qui-quadrado”.

- 6. Se o valor- $p$  for menor do que o nível predeterminado para  $\alpha$ , normalmente, 0,05, rejeite  $H_0$ ; o modelo não se sustenta. Se o valor- $p$  for maior do que  $\alpha$ , então você não pode rejeitar o modelo.**

Normalmente, o valor de  $\alpha$  é 0,05, mas alguns analistas de dados podem utilizar um valor maior (até 0,10), e outros podem utilizar um valor menor (por exemplo, 0,010). Consulte o Capítulo 3 para obter mais informações sobre como escolher  $\alpha$  e comparar seu valor- $p$  a ele.

Voltando ao exemplo dos M&M's, o valor- $p$ , 0,18, é maior que 0,05, sendo assim, você não deve rejeitar  $H_0$ , ou seja, não pode dizer que o modelo está errado. Assim, a Mars parece realmente cumprir os percentuais de cada cor dos M&M's, como anunciado. Pelo menos, não se pode dizer que não. (E tenho certeza de que a Mars já sabia disso.)

Embora alguns testes de hipóteses sejam bicaudais, o teste de qualidade de ajuste é sempre *unicaudal à direita*. Quando você estiver fazendo um teste de qualidade de ajuste, sempre vai examinar a cauda direita da distribuição do Qui-quadrado. Isso porque uma estatística da qualidade de ajuste com valor pequeno indica que os dados observados e o modelo esperado não se diferenciam muito e, portanto, você deve ficar com o modelo. No entanto, se o valor dessa estatística estiver fora da cauda direita da distribuição do Qui-quadrado,



essa é outra história. Essa situação indica que a diferença entre o que você observou e o que você esperava é maior do que a que você obteria devido ao acaso e, portanto, você tem provas suficientes para dizer que o modelo esperado está errado.



Use o teste Qui-quadrado de qualidade de ajuste para verificar se um modelo especificado se ajusta. O *modelo especificado* é um modelo no qual cada possível valor da variável  $x$  é listado com sua respectiva probabilidade, dada pelo modelo. Por exemplo, se você quiser testar se três hospitais locais atendem à mesma porcentagem de pacientes na sala de emergência, teste  $H_0: p_1 = p_2 = p_3$ , onde cada  $p$  representa a porcentagem de pacientes de emergência que se direcionam a cada hospital, respectivamente. Neste caso, cada  $p$  deve ser igual a 0,30, se os hospitais dividirem igualmente a carga de emergência.

## Parte V

# Estatística Não Paramétrica: Rebeldes sem Distribuição



**“Ted e eu passamos 120 horas juntos analisando os dados da pesquisa e veja o que descobrimos: Ted pede as canetas emprestadas e nunca as devolve, além disso, ele intencionalmente arrasta a cadeira só para me irritar e, é evidente, eu falo enquanto durmo.”**



## *Nesta parte...*

**S**uponha que você esteja dirigindo de volta para casa, e uma das ruas está bloqueada. O que você faz? Dá ré e encontra outro caminho para chegar em casa. A estatística não paramétrica é esse caminho alternativo que você usa quando os métodos estatísticos paramétricos não forem permitidos. Além do mais, essa rota alternativa, na verdade, acaba sendo melhor. Nesta parte, você vai descobrir o quão melhor a estatística não paramétrica é utilizando o teste do sinal (sign), o teste de postos sinalizados de Wilcoxon (Wilcoxon signed rank) e muitos mais.

# Capítulo 16

## Ficando Não Paramétrico

---

### *Neste Capítulo*

- ▶ Compreendendo a necessidade de técnicas não paramétricas
  - ▶ Distinguindo os métodos regulares dos métodos não paramétricos
  - ▶ Erguendo os alicerces: o básico sobre a estatística não paramétrica
- 

**M**uitos pesquisadores fazem análises envolvendo testes de hipóteses, intervalos de confiança, testes de Qui-quadrado, regressão e ANOVA. No entanto, a estatística não paramétrica não parece ganhar a mesma popularidade que os outros métodos. Ela vive no submundo — é uma heroína anônima, se você preferir. Contudo, a estatística não paramétrica é, de fato, uma área muito importante e muito útil dentro da Estatística, pois fornece resultados precisos quando outros métodos, os mais comuns, falham.

Neste capítulo, você vai entender a importância das técnicas não paramétricas e por que elas deveriam ocupar um lugar de destaque em sua caixa de ferramentas para a análise de dados. Também vai descobrir alguns dos termos e das técnicas que fazem parte da estatística não paramétrica.

# Em Favor da Estatística Não Paramétrica

A estatística não paramétrica desempenha um papel importante no mundo da análise de dados, uma vez que ela pode salvar o dia quando você não puder usar outros métodos. O problema é que os pesquisadores muitas vezes desconsideram, ou nem sequer conhecem, essas técnicas e, portanto, não as utilizam quando deviam. Nesse caso, você nunca sabe que tipo de resultados obteve, o que sabe é que eles podem muito bem estar errados.

Nas seções a seguir, você vai ver as vantagens e a flexibilidade do uso de um procedimento não paramétrico, além de descobrir o quão insignificante é sua desvantagem, o que a torna uma opção de sucesso na maioria das vezes.

## *Não precisa se preocupar se as condições não forem atendidas*

Muitas das técnicas comumente usadas para analisar os dados, incluindo muitas das apresentadas neste livro, impõem condições muito rígidas em relação aos dados que devem ser atendidos caso queira usá-las: as populações das quais seus dados são coletados geralmente devem seguir uma distribuição normal. Os métodos que requerem um certo tipo de distribuição (como a distribuição normal) para serem usados são chamados de *métodos paramétricos*.

A seguir, veja as formas para decidir se uma população tem ou não uma distribuição normal, com base em sua amostra:

- ✓ Você pode representar graficamente os dados utilizando um histograma e ver se ele parece ter a forma de um sino.

Para realizar um histograma no Minitab, insira seus dados em uma coluna. Clique em Graph>Histogram e em OK. Clique na variável no campo à esquerda para que ela apareça no campo Graph Variables. Clique em OK e cheque seu histograma.

- ✓ Você ainda pode fazer um gráfico de probabilidade normal, que compara seus dados à distribuição normal, usando um gráfico  $x$ - $y$  (semelhante aos utilizados quando você faz o gráfico de uma reta). Se os dados seguirem mesmo uma distribuição normal, o gráfico de probabilidade normal mostra uma reta. Porém, se os dados não seguirem uma distribuição normal, o gráfico de probabilidade normal não irá mostrar uma reta; talvez ele mostre uma curva assimétrica para um ou outro lado.

Para realizar um gráfico de probabilidade normal no Minitab, insira seus dados em uma coluna. Clique em Graph>Probability Plot e OK. Clique na variável no campo à esquerda para que ela apareça na coluna Graph Variables, clique em OK, e você tem o seu gráfico de probabilidade normal.

Os métodos não paramétricos entram em cena quando você achar que a condição de distribuição normal não tiver sido satisfeita. Os *métodos não paramétricos* são técnicas de análise de dados que não exigem que os dados tenham uma distribuição específica. Os



procedimentos não paramétricos podem exigir uma das duas seguintes condições (mas isso em apenas algumas situações):

- ✓ Que os dados venham de uma distribuição simétrica (cujos lados são idênticos se você cortá-la ao meio).
- ✓ Que os dados das duas populações venham do mesmo tipo de distribuição (uma que tenha o mesmo formato).

Note também que a distribuição normal tem apenas a média como sua estatística principal (por exemplo, o valor-Z do teste de hipótese para uma média populacional é a diferença entre os valores dos dados e a média, dividida pelo desvio padrão). Assim, a condição de que a população tenha uma distribuição normal automaticamente lhe indica que você está trabalhando com a média. No entanto, muitos procedimentos não paramétricos trabalham com a *mediana*, que é uma estatística muito mais flexível, pois não é influenciada pelos *valores discrepantes* (valores extremos acima ou abaixo da média) ou pela *assimetria* (o pico de um lado e uma longa cauda do outro), como a média.

## ***Uma chance para a mediana mostrar seu potencial***

Muitas vezes, uma questão estatística gira em torno do centro de uma população, ou seja, do número que representa um valor típico, ou um valor central, na população. Uma dessas medidas do centro é a *média*. A *média populacional* é o valor médio para toda população e que, normalmente, não é conhecido (e é por isso que você coleta uma amostra). Muitos analistas de dados focam muito a média populacional; eles querem estimá-la, testá-la, comparar as médias de duas ou mais populações ou prever o valor médio da variável  $y$  dada uma variável  $x$ . No entanto, a média não é a única medida do centro de uma população, existe também a boa e velha mediana.

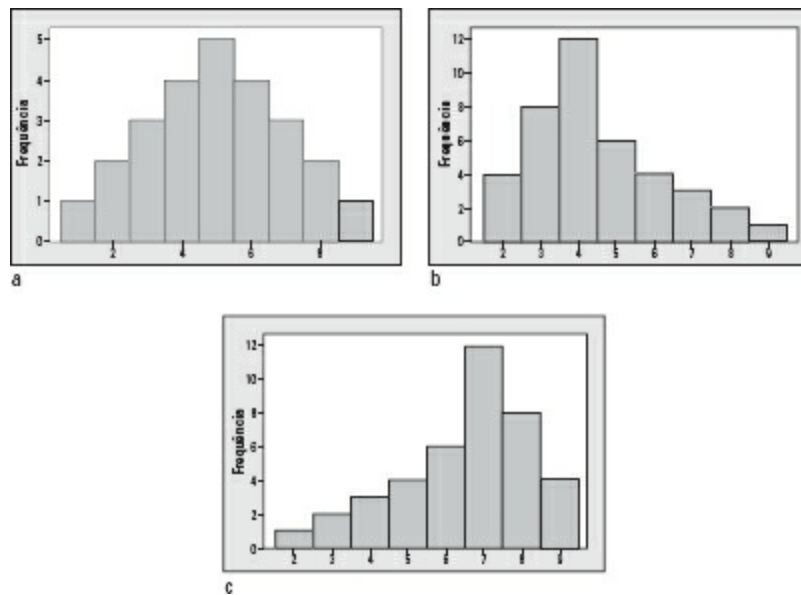
Você deve se lembrar que a mediana de um conjunto de dados é o valor que representa o meio exato quando os dados são ordenados do menor para o maior. Por exemplo, no conjunto de dados 1, 5, 4, 2, 3, você deve ordenar os dados de forma a obter: 1, 2, 3, 4, 5 e, então, descobrir que o número no meio, 3, é a mediana. Se o conjunto de dados tem um número par de valores, por exemplo, 2, 4, 6, 8, você deve encontrar a média dos dois números do meio para obter a mediana —  $(4 + 6) \div 2 = 5$ , neste caso.

Talvez você se lembre das aulas de Estatística I que é possível encontrar a média e a mediana de um conjunto de dados e compará-las. Para isso, primeiro, organize os dados em um histograma e examine sua forma.

- ✓ **Se o conjunto de dados for simétrico**, ou seja, os dois lados forem iguais quando você traçar uma linha dividindo-o ao meio, a média e a mediana serão iguais (ou próximas). A Figura 16-1a mostra um exemplo para essa situação. Nesse caso, a média e a mediana são iguais a 5.
- ✓ **Se o histograma for assimétrico à direita**, ou seja, você tem muitos valores

pequenos e poucos valores grandes, a média aumentará devido aos valores grandes, mas a mediana não será afetada. Neste caso, a média é maior do que a mediana. A Figura 16-1b mostra um exemplo desta situação, na qual a média é 4,5 e a mediana é 4,0.

- ✓ **Se o histograma for assimétrico à esquerda**, onde você tem muitos valores grandes, mas apenas alguns valores pequenos, a média diminuirá por causa dos valores pequenos, mas, ainda assim, a mediana não será afetada. Neste caso, a média é menor do que a mediana. A Figura 16-1c ilustra essa situação com uma média de 6,5 e uma mediana de 7,0.



**Figura 16-1:** Histogramas simétricos e assimétricos

Meu ponto aqui é mostrar a importância da mediana, uma medida do centro de uma população ou de um conjunto de dados amostrais. A mediana concorre com a média e, muitas vezes, ganha. Os pesquisadores usam procedimentos não paramétricos quando querem fazer uma estimativa, um teste, ou comparar a(s) mediana(s) de uma ou mais populações. Também usam a mediana nos casos em que seus dados são simétricos, mas não necessariamente seguem uma distribuição normal, ou quando desejam focar uma medida de centro que não seja influenciada por valores discrepantes ou pela assimetria.

Por exemplo, se você analisar os preços das casas em seu bairro, talvez, encontre um grande número de casas dentro de uma determinada faixa de preço relativamente pequena, como também pode encontrar algumas casas com preços muito mais elevados. Se uma corretora de imóveis que estiver vendendo uma casa em seu bairro quiser justificar o alto preço cobrado, ela poderá informar o preço médio das casas na vizinhança, pois a média é afetada pelos valores discrepantes. Nesse caso, a média é maior do que a mediana. Mas, se a corretora quiser ajudar alguém a comprar uma casa em seu bairro, ela vai examinar a mediana dos preços das casas, pois a mediana não é influenciada pelas casas com preços mais elevados e, portanto, será menor do que a média.





Agora, suponha que você queira chegar a um número que descreva o preço de casa em sua cidade. Você deveria usar a média ou a mediana? Nas aulas de Estatística I, você aprendeu as técnicas para estimar a média de uma população (consulte o Capítulo 3 para uma rápida revisão), mas, provavelmente, não aprendeu a calcular um intervalo de confiança para a mediana de uma população. Ah, sim, você pode coletar uma amostra aleatória e calcular a mediana a partir dela. Mas vai precisar de uma margem de erro para acompanhá-lo. E vou lhe dizer uma coisa: a fórmula para a margem de erro da média não funciona para a margem de erro da mediana (mas, espere, este livro vai abordar esse assunto).

## *Então, qual é a pegadinha?*

Você pode estar se perguntando: “Onde está a pegadinha em usar uma técnica não paramétrica? Deve haver uma câmera escondida em algum lugar por aqui”. Bom, muitos pesquisadores acreditam que as técnicas não paramétricas jogam por água abaixo os resultados estatísticos; por exemplo, suponha que você encontre uma diferença real entre duas médias populacionais, e as populações realmente tenham uma distribuição normal. Na técnica paramétrica, o teste de hipótese para duas médias provavelmente detectaria essa diferença (se o tamanho da amostra fosse grande o suficiente).

A questão é: se você usar uma técnica não paramétrica (que não requer que as populações sigam uma distribuição normal), vai correr o risco de não encontrar essa diferença? Talvez. Mas o risco não é tão grande quanto você pensa. Na maioria das vezes, os procedimentos não paramétricos são só um pouco menos eficientes do que os procedimentos paramétricos (o que significa que eles não funcionam tão bem quanto os procedimentos paramétricos na hora de detectar um resultado significativo ou estimar um valor) quando a condição de normalidade é satisfeita, mas essa diferença com relação à eficiência é pequena.

Mas a grande recompensa acontece quando as condições de distribuição normal não são atendidas, situação em que as técnicas paramétricas podem nos fazer chegar à conclusão errada, enquanto as técnicas não paramétricas correspondentes podem nos levar à resposta correta. Muitos pesquisadores não sabem disso, então, vamos espalhar a boa-nova!

Conclusão: antes de mais nada, sempre verifique se há normalidade. Se você estiver certo de que a condição de normalidade foi satisfeita, vá em frente e use os procedimentos paramétricos, pois, nesse caso, eles são mais precisos. Porém, se você tiver qualquer dúvida sobre a condição de normalidade, use os procedimentos não paramétricos. Mesmo quando a condição de normalidade for satisfeita, os procedimentos não paramétricos serão só um pouco menos precisos do que os procedimentos paramétricos. No entanto, se a condição de normalidade não for atendida, os não paramétricos fornecerão resultados adequados e justificáveis para as situações em que os procedimentos paramétricos talvez não funcionem.



# ***Dominando o Básico das Estatísticas Não Paramétricas***

Uma vez que você possa não ter visto nada sobre estatística não paramétrica em suas aulas de Estatística I, seu primeiro passo rumo à utilização dessas técnicas é entender alguns dos princípios básicos. Nesta seção, você vai conhecer alguns dos principais conceitos e a terminologia envolvidos na estatística não paramétrica. Esses termos e conceitos serão muito usados do Capítulo 17 ao 20 deste livro.

## ***Sinal***

O *sinal* é o valor 0 ou 1 atribuído a cada número do conjunto de dados. O sinal de um valor no conjunto de dados representa se o valor do dado é maior ou menor do que um número especificado. O valor +1 é dado quando o valor dos dados é maior do que o número especificado, e o valor 0 é dado quando o valor dos dados é menor ou igual ao número especificado. Por exemplo, suponha que o conjunto de dados seja 10, 12, 13, 15, 20, e o número especificado para a comparação seja 16. Uma vez que 10, 12, 13 e 15 são menores do que 16, cada um recebe um sinal 0. Como 20 é maior do que 16, ele recebe um sinal +1.

Várias utilizações da estatística de sinal aparecem nas estatísticas não paramétricas. Você pode usar os sinais para ver se a mediana de uma população é igual a um valor especificado ou usá-los para analisar os dados de um experimento de pares combinados (onde os indivíduos são combinados de acordo com uma variável e um tratamento é aplicado e comparado). Você também pode usar os sinais combinados a outras estatísticas não paramétricas. Por exemplo, você pode combinar sinais a ranques a fim de elaborar estatísticas para comparar a mediana de duas populações. (Na próxima seção, discuto os ranques e, no Capítulo 18, utilizo-os em um teste de hipótese para duas medianas populacionais.)

Nas seções a seguir, você vai ver exatamente como usar a estatística do sinal para testar a mediana de uma população e analisar os dados de um experimento de pares combinados.

## ***Testando a mediana***

Você pode usar os sinais para testar se a mediana de uma população é igual a algum valor  $m$ . Para fazer isso, realize um teste de hipóteses baseado nos sinais. Sendo assim, tem-se  $H_0$ : Mediana =  $m$  versus  $H_a$ : mediana  $\neq m$  (mas você também pode usar um sinal de  $>$  ou  $<$  na  $H_a$ ). Sua estatística de teste é a soma dos sinais de todos os dados. Se essa soma for significativamente maior ou menor do que o esperado se a  $H_0$  fosse verdade, você deve rejeitar a  $H_0$ . O quão grande ou pequena deve ser a soma dos sinais para que você rejeite a  $H_0$  é definido pelo teste do sinal (consulte o Capítulo 17).

Suponha que você esteja testando se a mediana de uma população é igual a 5. Ou seja, está testando a  $H_0$ : Mediana = 5 versus a  $H_a$ : Mediana  $\neq 5$ . Então, você coleta os seguintes

dados: 4, 4, 3, 3, 2, 6, 4, 3, 3, 5, 7, 5. Ordenando os dados, tem-se 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 7. Agora, encontre o sinal para cada valor no conjunto de dados, determinado pelo fato de o valor ser maior ou menor do que 5. O sinal do primeiro valor dos dados, 2, é 0, pois ele é inferior a 5. Cada um dos 3 recebe um sinal 0, como os 4 e os 5, pela mesma razão. Somente os números 6 e 7 recebem um sinal 1, já que são os únicos valores no conjunto de dados maiores do que 5 (o número de interesse para a mediana).

Ao somar os sinais, você está, em essência, contando o número de valores no conjunto de dados que são maiores do que a quantidade indicada na  $H_0$ . Por exemplo, o total de todos os sinais dos valores ordenados é:

$$0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 1 + 1 = 2$$

Assim, conclui-se que o número total de valores de dados acima de 5 (o número de interesse para a mediana) é 2. O fato de que o total dos sinais (2) é muito menor do que a metade do tamanho amostral lhe dá uma evidência de que a mediana provavelmente não seja 5, pois ela representa o centro da população. Se a mediana fosse de fato 5 para a população, a amostra deveria ter cerca de seis valores abaixo e seis acima dela.

### ***Realizando um experimento de pares combinados***

Você pode usar os sinais em um *experimento de pares combinados* em que o mesmo indivíduo é usado duas vezes ou os indivíduos são arranjados em pares para algumas variáveis importantes. Por exemplo, você pode usar os sinais para testar se um determinado tratamento resultou ou não na melhora dos pacientes, quando comparado a um controle. Nos casos em que a estatística do sinal é utilizada, a melhora não é medida pela média das diferenças nas respostas ao tratamento em relação ao controle (como em um teste- $t$  pareado), mas, sim, pela mediana dessas diferenças.

Suponha que você esteja testando um novo anti-histamínico para pacientes alérgicos. A amostra coletada tem 100 pacientes e cada paciente avalia a gravidade dos sintomas de sua alergia antes e depois de tomar a medicação usando uma escala de 1 (melhor) a 10 (pior). (É claro que você precisa fazer um experimento controlado em que alguns dos pacientes recebem um placebo para considerar o fato de que algumas pessoas podem sentir o alívio dos sintomas só por terem tomado alguma coisa.)

Neste estudo, você não está interessado no nível em que estão os sintomas dos pacientes, mas em quantos pacientes tiveram o menor nível de sintomas depois de tomar o medicamento. Sendo assim, encontre a diferença entre o nível dos sintomas antes do experimento e o nível dos sintomas após o experimento para cada indivíduo.

- ✓ Se essa diferença for positiva, o medicamento parece ter ajudado e, por isso, você atribui um sinal +1 a essa pessoa (em outras palavras, conte-o como um sucesso).
- ✓ Se a diferença for zero, o remédio não teve efeito e, portanto, dê a essa pessoa um sinal 0.

Lembre-se, porém, que a diferença pode ser negativa, indicando que os sintomas antes eram mais baixos do que os sintomas depois, ou seja, o remédio piorou os sintomas. Assim, essa situação também resulta em um sinal 0.

Depois de ter encontrado o sinal para cada valor ou par do conjunto de dados, você está pronto para analisá-lo usando o teste do sinal ou teste de postos sinalizados (consulte o Capítulo 17).

## Postos

Os *postos* configuram uma forma interessante de usar a informação importante do conjunto de dados sem usar seus reais valores. O posto entra em jogo na estatística não paramétrica quando você não está interessado nos valores dos dados, mas, sim, na posição que ocupam em comparação a determinado valor para a mediana ou quando comparados aos postos de valores em outro conjunto de dados de outra população. (Para ver os postos em ação, consulte o Capítulo 18).

O *posto* de um valor dentro de um conjunto de dados é o número que representa seu lugar em uma ordem que vai do menor ao maior. Por exemplo, se o conjunto de dados é 1, 10, 4, 2, 1.000, você pode atribuir os postos da seguinte maneira: 1 ocupa o posto 1 (porque é o menor), 2 ocupa o posto 2, 4 ocupa o posto 3 (sendo o terceiro menor número no conjunto de dados ordenado), 10 ocupa o posto 4 e 1.000 ocupa o posto 5 (por ser o maior valor).

Agora, suponha que seu conjunto de dados seja 1, 2, 20, 20, 1.000. Como atribuir os postos nessa situação? Você sabe que o 1 ocupa o posto 1 (por ser o menor valor), o 2 ocupa o posto 2 e 1.000 ocupa o posto 5 (por ser o maior). Mas, e os dois 20 deste conjunto de dados? Será que o primeiro 20 deve ocupar o posto 3 e o segundo, o posto 4? Isso não parece fazer sentido, pois você não pode diferenciar os dois 20.

A situação em que dois valores em um conjunto de dados são iguais é chamada de empate. Para atribuir os postos quando existem empates, calcule a média dos dois postos que os valores precisariam ocupar e atribua a cada valor empatado a média do posto. Se você tiver um empate entre três números, terá três postos sobrando e, portanto, some-os e divida o resultado por três.

Neste caso, uma vez que os 20 estão disputando os postos 3 e 4, atribua a cada um deles o posto 3,5, a média dos dois postos que devem ser compartilhados. Mostro a classificação final para o conjunto de dados 1, 2, 20, 20, 1.000 na Tabela 16-1.

**Tabela 16-1** Postos dos Valores do Conjunto de Dados 1, 2, 20, 20, 1.000

<i>Valor do Dado</i>	<i>Posto Atribuído</i>
1	1
2	2
20	3,5



O menor posto pode ser 1 e o maior,  $n$ , onde  $n$  é o número de valores do conjunto de dados. Se você tiver um valor negativo em um conjunto de dados, por exemplo, se o conjunto de dados for  $-1, -2, -3$ , você ainda atribui postos de 1 a 3 para esses valores. Nunca atribua postos negativos a dados negativos. (A propósito, quando você ordenar o conjunto de dados  $-1, -2, -3$ , a classificação será  $-3, -2, -1$ ; sendo assim,  $-3$  ocupa o posto 1,  $-2$ , o posto 2, e  $-1$ , o posto 3.)

## Postos com sinais

A técnica dos *postos com sinais* combina os conceitos do sinal e dos postos ocupados por um valor em um conjunto de dados, com uma pequena diferença. O sinal indica se o número é maior, menor ou igual a um valor especificado. Um posto indica o local que esse número ocupa na classificação, do menor para o maior, dos valores do conjunto de dados.

Para calcular os postos com sinais para cada valor do conjunto de dados, siga estes passos:

- 1. Atribua um sinal +1 ou 0 para cada valor do conjunto de dados, levando em consideração se o valor do conjunto de dados é maior ou menor do que um valor especificado.**

Se for maior que o valor especificado, o sinal atribuído é +1, se for menor ou igual ao valor especificado, o sinal atribuído é 0.

- 2. Classifique os dados originais do menor para o maior, de acordo com seus valores absolutos.**

Os estatísticos denominam estes valores de *postos absolutos*. O valor absoluto de qualquer número é a versão positiva desse número. O símbolo para o valor absoluto é  $||$ , com o número entre essas linhas. Por exemplo,  $|-2| = 2$ ,  $|+2| = 2$  e  $|0| = 0$ .

- 3. Multiplique o sinal pelo posto absoluto para obter o posto com sinal para cada valor do conjunto de dados.**

Uma situação em que você pode usar o teste de postos com sinais é o experimento em que se compara uma variável de resposta para um grupo tratamento versus um grupo controle. Você pode testar a diferença causada por um tratamento através da coleta de dados em pares, podendo ser os dois provenientes da mesma pessoa (pré-teste versus pós-teste) ou provenientes de dois indivíduos que aparentam ser muito semelhantes.

Por exemplo, suponha que você compare quatro pacientes em relação à sua perda de peso através de um programa de dieta. O que realmente lhe interessa é saber se a alteração global no peso é menor do que zero para a população. Os dois fatores a seguir são



importantes:

- ✓ Saber se a pessoa perdeu ou não peso
- ✓ Quanto o peso de uma pessoa se alterou em comparação ao peso de todas as outras pessoas do conjunto de dados

Você mede o peso da pessoa antes do programa (pré-teste), bem como o seu peso após o programa (pós-teste). Essa alteração é a faceta importante dos dados em que você está interessado, logo, é nela que os sinais vão ser aplicados, atribuindo, então, um sinal +1 para quem perdeu peso (o que constitui um sucesso para o programa) e um sinal 0 para quem continuou com o mesmo peso ou ganhou mais (não contribuindo para o sucesso do programa). Depois, converta todas as alterações no peso em valores absolutos e, então, classifique-os em postos (em outras palavras, você vai encontrar os postos absolutos para as alterações no peso). O posto com sinais é o produto entre o sinal e o posto absoluto. Depois de sinalizar os postos é que você realmente poderá comparar a eficiência do programa: postos sinalizados altos indicam uma grande perda de peso.

Por exemplo, as diferenças de peso  $-20$ ,  $-10$ ,  $+1$  e  $5$  apresentam os sinais  $+1$ ,  $+1$ ,  $0$  e  $0$ . Os valores absolutos das diferenças de peso são  $20$ ,  $10$ ,  $1$  e  $5$ . Seus postos absolutos são, respectivamente,  $4$ ,  $3$ ,  $1$  e  $2$ . Os postos sinalizados são  $4 * 1 = 4$ ,  $3 * 1 = 3$ ,  $1 * 0 = 0$  e  $2 * 0 = 0$ .

## *Soma de postos*

A *soma de postos* é simplesmente a soma de todos os postos. A soma de postos geralmente é usada em situações em que se está comparando duas ou mais populações para ver se uma delas possui uma localização central maior que a outra. (Ou seja, se analisarmos as populações por meio de seus histogramas, uma estaria deslocada à direita da outra.)

Veja aqui um exemplo de como os pesquisadores utilizam a soma de postos: suponha que você esteja comparando as notas de prova de duas classes e elas não têm uma distribuição normal, portanto, você pretende utilizar técnicas não paramétricas para compará-las. A nota máxima para essa prova é  $30$ . Você, então, coleta amostras aleatórias com cinco notas de cada uma das classes. Suponha que os dados coletados sejam os seguintes:

<i>Classe Número Um</i>	<i>Classe Número Dois</i>
22	23
23	30
20	27
25	28
26	25

O truque aqui é combinar todos os dados em um grande conjunto de dados, classificando todos os valores e somando os postos para a primeira amostra e, em seguida, para a

segunda. Depois, compare as duas somas. Se uma delas for maior do que a outra, esse resultado pode indicar que uma das classes se saiu melhor na prova.

Para o exemplo em questão, os dados ordenados para as classes combinadas aparecem na primeira linha, com seus respectivos postos na segunda linha. Os dados circulados são provenientes da primeira classe.

Dados Ordenados	20	22	23	23	25	25	26	27	28	30
Respectivos Postos	1	2	3.5	3.5	5.5	5.5	7	8	9	10

A soma de postos para a primeira classe é  $1 + 2 + 3,5 + 5,5 + 5,5 + 7 = 19$ , bem menor do que a soma de postos da segunda classe,  $3,5 + 5,5 + 5,5 + 8 + 9 + 10 = 36$ . Esse resultado indica que, nesta amostra, a segunda turma se saiu melhor na prova do que a primeira.

O Capítulo 18 mostra como usar um teste da soma de postos para ver se as formas de duas distribuições populacionais são iguais, ou seja, os valores que elas assumem e a frequência com que esses valores ocorrem em cada população. No Capítulo 19, você pode encontrar mais usos para a soma de postos, incluindo o teste de Kruskal-Wallis.

Observe que comparar a média de cada conjunto de dados usando um teste- $t$  para duas amostras seria errado para o exemplo da prova, pois as notas não possuem uma distribuição normal. De fato, se a prova fosse fácil, você observaria muitas notas altas e poucas notas baixas e a população seria assimétrica à esquerda. Por outro lado, se a prova fosse difícil, você observaria muitas notas baixas e poucas notas altas, e a população seria assimétrica à direita (mas não se preocupe muito com essa situação). Em ambos os casos, você precisaria de um procedimento não paramétrico. Consulte o Capítulo 18 para mais informações sobre o equivalente não paramétrico para o teste- $t$ .



# Capítulo 17

## Todos os Sinais Apontam para o Teste dos Sinais e o Teste de Postos Sinalizados

### *Neste Capítulo*

- ▶ Testando e estimando a mediana: o teste dos sinais
- ▶ Descobrindo quando e como usar um teste de postos sinalizados

O teste de hipótese que você aprendeu em Estatística I usa distribuições bem conhecidas, como a distribuição normal ou a distribuição- $t$  (veja o Capítulo 3). O uso destes testes exige a satisfação de certas condições, tais como o tipo de dado que está sendo usado, a distribuição da população de origem dos dados e o tamanho do conjunto de dados. Os procedimentos que envolvem tais condições são chamados de *procedimentos paramétricos*. Em geral, os procedimentos paramétricos são muito poderosos e precisos, e os estatísticos os usam o quanto podem.

Porém, há situações em que os dados não satisfazem as condições exigidas para a realização de um procedimento paramétrico. Talvez, você simplesmente não tenha dados o suficiente (o maior obstáculo é saber se os dados são provenientes de uma população com distribuição normal) ou seus dados não sejam quantitativos, como, por exemplo, os postos (quando você não coleta dados numéricos, mas apenas ordena os dados do menor para o maior ou vice-versa).

Nessas situações, a sua melhor aposta é um *procedimento não paramétrico* (veja o Capítulo 16 para obter mais informações). Em geral, os procedimentos não paramétricos não são tão poderosos quanto os procedimentos paramétricos, mas têm muito poucas condições vinculadas a eles. Além disso, são fáceis de realizar e suas fórmulas fazem sentido. Mas o mais importante: os procedimentos não paramétricos fornecem resultados precisos, em comparação aos procedimentos paramétricos, quando as condições para os procedimentos paramétricos não são satisfeitas ou não estão adequadas.

Neste capítulo, você vai usar o teste do sinal (Sigh Test) e o teste de postos sinalizados de Wilcoxon para testar ou estimar a mediana (Median) de uma população. Estes procedimentos não paramétricos são as contrapartidas para os testes  $t$  para uma amostra ou para pares combinados, que requerem dados provenientes de uma população com distribuição normal.



# *Interpretando os Sinais: O Teste dos Sinais*

O teste- $t$  para uma amostra é usado em Estatística I para testar se a média da população é igual a um determinado valor. Ele, no entanto, exige que os dados tenham uma distribuição normal, mas, quando esta condição não é satisfeita, o *teste do sinal* é uma alternativa não paramétrica ao teste- $t$  para uma amostra. Ele testa se a mediana da população é ou não igual a um determinado valor.

O que faz com que o teste do sinal seja tão prático é que ele se baseia em uma distribuição muito simples, a distribuição binominal. Você pode usar a distribuição binomial quando tiver uma sequência de  $n$  tentativas de um experimento, com apenas dois possíveis resultados para cada momento (sucesso ou fracasso). A probabilidade de sucesso é denotada por  $p$ , e é a mesma para cada tentativa. A variável é  $x$ , o número de sucessos nas  $n$  tentativas. (Para mais informações sobre a distribuição binomial, consulte um livro de Estatística I.)

A única condição para o teste do sinal é que os dados sejam ordinais ou quantitativos — não categóricos. No entanto, isso não é uma grande coisa, porque, se você está interessado na mediana, não iria mesmo coletar dados categóricos.

Aqui estão os passos para a realização do teste do sinal. Note que o Minitab pode fazer do quarto ao oitavo passos para você, no entanto, como sempre, é importante compreender o que o Minitab faz nos bastidores.

## **1. Estabeleça sua hipótese nula: $H_0: m = m_o$**

O verdadeiro valor da mediana é  $m$ , e  $m_o$  é o valor atribuído a ela (o valor que está sendo testado).

## **2. Estabeleça sua hipótese alternativa. Suas opções são: $H_a: m \neq m_o$ ; ou $H_a: m > m_o$ ; ou $H_a: m < m_o$ .**

A escolha da  $H_a$  depende da conclusão que você quer tirar caso a  $H_0$  seja rejeitada. Por exemplo, se só quiser saber quando a mediana é maior do que um número  $m$ , utilize  $H_a: m > m_o$ . O Capítulo 3 fala mais sobre a formulação de hipóteses alternativas.

## **3. Colete uma amostra aleatória de dados (ordinais ou quantitativos) da população.**

## **4. Atribua um sinal de mais ou de menos a cada valor do conjunto de dados.**

Se uma observação é menor do que  $m_o$ , atribua um sinal de menos (-). Se a observação é maior do que  $m_o$ , atribua um sinal de mais (+). Se a observação é igual a  $m_o$ , ignore-a e diminua o tamanho amostral em um.

Na distribuição binomial, tem-se  $n$  valores no conjunto de dados, e cada um tem dois possíveis resultados: ele pode ficar abaixo ou acima de  $m_o$ . (Semelhante a sucesso e

fracasso.)

**5. Conte todos os sinais positivos. Essa soma será sua estatística de teste, indicada por  $k$ .**

Na distribuição binomial, esta soma representa o número total de sucessos, onde um sinal de mais (+) representa o sucesso.

**6. Localize a estatística de teste  $k$  (do quinto passo) na distribuição binomial (utilizando a Tabela A-2 no apêndice).**

Você deve determinar a posição ocupada pela estatística de teste na distribuição binomial, procurando-a em uma tabela de distribuição binomial (consulte um livro didático). Para fazer isso, você precisa saber de  $n$ ,  $k$  e  $p$ .

O tamanho amostral é  $n$ , a estatística de teste é  $k$  (do quinto passo), mas qual é o valor de  $p$ , a probabilidade de sucesso? Se a hipótese nula  $H_0$  for verdadeira, 50% dos dados devem ficar abaixo de  $m_0$  e 50% devem ficar acima. Isto quer dizer que, na distribuição binomial, um sucesso (+) tem probabilidade de  $p = 0,50$ .

**7. Encontre o valor- $p$  de sua estatística:**

- Se  $H_a$  tem um sinal  $<$ , some todas as probabilidades na tabela binomial para  $x \leq k$ .
- Se  $H_a$  tem um sinal  $>$ , some todas as probabilidades na tabela binomial para  $x \geq k$ .
- Se  $H_a$  tem um sinal  $\neq$ , some as probabilidades na tabela binomial para que  $x$  seja maior ou igual a  $k$  e dobre este valor. Isto lhe dá o valor- $p$  do teste.

**8. Tire suas conclusões:**

Se o valor- $p$  (do sexto passo) for menor do que o valor predeterminado para  $\alpha$  (normalmente, 0,05), rejeite a  $H_0$  e conclua que a mediana é maior, menor ou  $\neq m_0$ , dependendo da  $H_a$  escolhida. Caso contrário, você não pode rejeitar a  $H_0$ .

Para realizar um teste do sinal no Minitab, insira seus dados em uma coluna. Clique em Stat>Nonparametric>One-sample Sign. Clique na variável na caixa à esquerda e, depois, em Select. Essa variável vai aparecer na caixa Variables. Em seguida, clique em OK e os resultados do teste estão prontos.

Nas seções a seguir, vou mostrar duas maneiras diferentes de usar o teste do sinal:

- ✓ Para testar ou estimar a mediana de uma população
- ✓ Para testar ou estimar a diferença mediana de dados em que as observações vêm em pares, seja do mesmo indivíduo (pré-teste versus pós-teste) ou de indivíduos pareados de acordo com características relevantes

## ***Testando a mediana***

Surgem situações em que você não está interessado na média, mas na mediana de uma



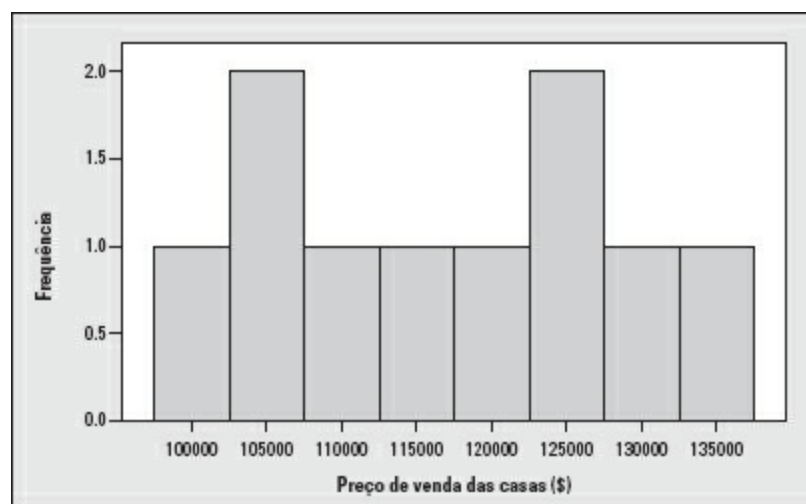
população (consulte o Capítulo 16 para mais informações sobre mediana). Por exemplo, talvez, os dados não tenham uma distribuição normal, ou mesmo simétrica. Quando se deseja estimar ou testar a mediana de uma população (vamos chamá-la de  $m$ ), o teste do sinal é uma ótima opção.

Suponha que você seja um corretor de imóveis e esteja vendendo casas em um determinado bairro. Então, você ouve de outros corretores que o preço de venda mediano (Selling Price) das casas naquele bairro é \$110.000, mas você acha que a mediana é, na verdade, mais alta. Uma vez que está interessado no preço mediano de uma casa, em vez do preço médio, decide testar a afirmação através de um teste do sinal. Siga os passos para este teste:

1. Estabeleça sua hipótese nula: já que a afirmação original é que o preço mediano de uma casa é \$110.000, tem-se a  $H_0:m = \$110.000$ .
2. Estabeleça sua hipótese alternativa: uma vez que você acha que a mediana é mais alta do que \$110.000, sua hipótese alternativa é  $H_a:m > \$110.000$ .
3. Colete uma amostra aleatória de dez casas no bairro. Você pode ver os dados na Tabela 17-1; o histograma está na Figura 17-1.

Agora, a pergunta é: o preço mediano das casas no bairro é igual a \$110.000, ou é maior (como você suspeita)?

Tabela 17-1		Amostra de preços de casas em um bairro	
<i>Casa</i>	<i>Preço</i>	<i>Sinal (Comparado a \$110.000)</i>	
1	\$132.000	+	
2	\$107.000	–	
3	\$111.000	+	
4	\$105.000	–	
5	\$100.000	–	
6	\$113.000	+	
7	\$135.000	+	
8	\$120.000	+	
9	\$125.000	+	
10	\$126.000	+	



**Figura 17-1:** Histograma do preço de venda de dez casas. 2.0 1.5

4. Atribua um sinal positivo a qualquer preço mais alto do que \$110.000 e um sinal negativo para todo preço menor que \$110.000 (veja a coluna três Tabela 17-1).
5. Calcule a estatística de teste. Sua estatística de teste é 7, o número de sinais “+” em seu conjunto de dados (veja a Tabela 17-1), que representa o número de casas em sua amostra cujos preços estão acima de \$110.000.
6. Compare sua estatística de teste com a distribuição binomial (consulte uma tabela de distribuição binomial) para encontrar o valor- $p$ .

Para este caso, verifique a linha na tabela binomial onde  $n = 10$  (o tamanho da amostra) e  $k = 7$  (a estatística de teste) e a coluna onde  $p = 0,50$  (porque, se a mediana da população for igual a  $m_o$ , 50% dos valores da população devem estar acima dela e 50%, abaixo). De acordo com a tabela, a probabilidade de que  $x$  seja igual a 7 é 0,117.

Uma vez que se trata de um teste unicaudal à direita (ou seja,  $H_a$  tem um sinal  $>$ ), deve-se somar as probabilidades de se estar em ou além de 7 para chegar ao valor- $p$ . O valor- $p$  neste caso é  $0,117 + 0,044 + 0,010 + 0,001 = 0,172$ .

7. Para concluir, compare o valor- $p$  (0,172) ao nível  $\alpha$  predeterminado (eu sempre uso 0,05). Já que o valor- $p$  é maior que 0,05, você não pode rejeitar a  $H_o$ . Não há provas suficientes nos dados para dizer que o preço mediano de venda das casas naquele bairro seja maior do que \$110.000. A Figura 17-2 mostra esses resultados calculados pelo Minitab.

Sign Test for Median: Selling Price						
Sign test of median = 110000 versus > 110000						
	N	Below	Equal	Above	P	Median
Selling Price	10	3	0	7	0.1719	116500

**Figura 17-2:** Teste do sinal para preço de casas realizado pelo Minitab.



Se os dados se aproximam de uma distribuição normal e a média for a medida mais adequada para a sua situação, não use o teste do sinal, mas, sim, o teste- $t$  para uma amostra (ou teste- $Z$ ). O teste do sinal não é tão potente (capaz de rejeitar a  $H_0$  quando deveria) quanto o teste- $t$  nas situações em que as condições para ele são atendidas. Mais importante, porém, não recorra ao teste- $t$  para reanalisar seus dados caso o teste do sinal não rejeite a  $H_0$ . Isso seria impróprio e antiético. Em geral, os estatísticos consideram a ideia de realizar um procedimento paramétrico depois de um não paramétrico na esperança de obter resultados mais significativos como *data fishing*, o conceito de ficar analisando os dados de maneiras diferentes até que um resultado estatisticamente significativo seja obtido.

## ***Estimando a mediana***

Você também pode usar o teste do sinal para encontrar o intervalo de confiança para uma mediana populacional. Isto vem a calhar quando você está interessado em estimar o valor mediano de uma população, como a renda mediana de uma família nos Estados Unidos ou o salário mediano de pessoas que acabaram de concluir um MBA.

A seguir, estão os passos para o cálculo de um intervalo de confiança para a mediana usando a estatística de teste do teste do sinal, assumindo que sua amostra aleatória de dados já tenha sido coletada. Note que o Minitab pode calcular o intervalo de confiança para você (do segundo ao quinto passos), mas é importante saber como o Minitab executa esses passos:

- 1. Determine o seu nível de confiança,  $1 - \alpha$  (isto é, o grau de confiança que você quer ter de que este processo irá estimar  $m$  corretamente a longo prazo).**

O nível de confiança que os analistas de dados normalmente usam é de 95% (consulte o Capítulo 3 para mais informações).

- 2. Na tabela binomial (Tabela A-2 no apêndice) encontre a seção para  $n$  igual ao tamanho da amostra e a coluna onde  $p = 0,50$  (pois a mediana é o ponto onde 50% dos dados se encontram acima e 50% abaixo).**

Nesta seção, você vai encontrar as probabilidades para os valores de  $x$  que vão de 0 a  $n$ .

- 3. Começando pelas extremidades ( $x = 0$  e  $x = n$ ) e movendo, passo a passo, em direção ao meio dos valores de  $x$ , some as probabilidades para esses valores de  $x$  até ultrapassar o total do  $\alpha$  (um menos seu nível de confiança).**
- 4. Anote o número de passos que você deu antes de ultrapassar o valor de  $1 - \alpha$ . Chame esse número de  $c$ .**
- 5. Ordene o conjunto de dados do menor para o maior. Começando pelas extremidades, vá em direção ao meio até chegar ao  $c$ -ésimo número de baixo e o  $c$ -ésimo número de cima.**

- Utilize estes números como as pontas baixa e alta do intervalo. Este resultado é o
6. intervalo de confiança para a mediana.

Você pode usar esses passos para encontrar o intervalo de confiança para a mediana do exemplo da seção anterior. Veja como:

1. Deixe que o nível de confiança seja  $1 - \alpha = 0,95$ .
2. Na tabela binomial (Tabela A-2 no apêndice), encontre a seção onde  $n = 10$  (tamanho da amostra) e  $p = 0,50$ . Esses valores se encontram na Tabela 17-2.

Tabela 17-2	Probabilidades Binomiais para Cálculo do Intervalo de Confiança para Mediana ( $n = 10, p = 0,50$ )
$x$	$p(x)$
0	0,001
1	0,010
2	0,044
3	0,117
4	0,205
5	0,246
6	0,205
7	0,117
8	0,044
9	0,010
10	0,001

3. Comece com os valores periféricos de  $x$  ( $x = 0$  e  $x = 10$ ) e some essas probabilidades para obter  $0,001 + 0,001 = 0,002$ . Uma vez que você ainda não ultrapassou 0,05 (o valor de  $\alpha$ ), vá para os valores mais internos de  $x$  ( $x = 1$  e  $x = 9$ ). Some as probabilidades deles ao que tem até agora para obter 0,002 (o primeiro total) + 0,010 + 0,010 = 0,022. Você ainda não ultrapassou 0,05 ( $\alpha$ ), então, dê mais um passo. Some as probabilidades do próximo valor interno,  $x = 2$  e  $x = 8$ , ao total geral para obter 0,022 (total anterior) + 0,044 + 0,044 = 0,110. Agora, você ultrapassou o valor de  $\alpha = 0,05$ . O valor de  $c$  é 2, pois você ultrapassou 0,05 no terceiro par de valores internos de  $x$ , portanto, deve recuar um passo para obter o valor de  $c$ .
4. Ordene o conjunto de dados (Tabela 17-1), do menor para o maior, o que lhe dá (em dólares) 100.000, 105.000, 107.000, 111.000, 113.000, 120.000, 125.000, 126.000, 132.000 e 135.000
5. Uma vez que  $c = 2$ , dê dois passos das extremidades ao centro do conjunto de dados para encontrar os dois valores internos, ou seja, \$105.000 e \$132.000. Junte esses dois números para formar um intervalo e concluir que um intervalo de confiança de 95% para o preço mediano de venda de uma casa no bairro em questão fica entre \$105.000 e



Para encontrar um intervalo de confiança  $1 - \alpha$  por cento para a mediana, usando o Minitab com base no teste do sinal, insira seus dados em uma única coluna. Clique em Stat>Nonparametric>One-sample Sign. Clique na variável na coluna à esquerda, para a qual deseja calcular o intervalo de confiança, e ela aparecerá na coluna Variables. Clique no círculo que diz: Confidence Interval e digite em  $1 - \alpha$  o valor que deseja para o nível de confiança (o padrão é 95%, escrito como 95). Clique em OK para obter o intervalo de confiança.

## *Testando os pares combinados*

A aplicação mais útil do teste dos sinais está em testar os pares combinados — ou seja, dados que vêm em pares e representam duas observações da mesma pessoa (pré-teste versus pós-teste, por exemplo) ou um conjunto de dados em que cada par é formado por dados de pessoas que são combinadas de acordo com características relevantes. Nesta seção, você vai ver como comparar os dados de um estudo de pares combinados para analisar o efeito de um tratamento, utilizando o teste dos sinais para a mediana.



A ideia de usar o teste dos sinais para a diferença mediana com dados em pares combinados é semelhante a usar o teste- $t$  para as diferenças médias com dados em pares combinados. (Para mais informações sobre dados em pares combinados e teste- $t$ , consulte um livro de Estatística I.) O teste da mediana (e não da média) é usado quando os dados não têm necessariamente uma distribuição normal, ou quando se está interessado apenas na diferença mediana, e não na diferença média.

Primeiro, estabeleça sua hipótese nula,  $H_0$ : a mediana é zero (indicando que não há diferença entre os pares). A hipótese alternativa, então, é  $H_a$ : a mediana é  $\neq 0$ ,  $> 0$  ou  $< 0$ , se você quiser saber se o tratamento fez alguma diferença, uma diferença positiva ou uma diferença negativa em relação ao controle, respectivamente. Depois de coletar os dados (duas observações por pessoa, ou um par de observações provenientes de duas pessoas combinadas). Depois disso, use o Minitab para realizar os passos de quatro a sete do teste dos sinais.

Por exemplo, suponha que você queira saber se mascar chiclete diminui a ansiedade na hora da prova. Para isso, 20 alunos são combinados em pares de acordo com fatores relevantes, tais como a média de notas, pontuação em provas anteriores e outros. Apenas um membro de cada par é selecionado aleatoriamente para mascar um chiclete durante a prova. A ansiedade de cada pessoa é medida através de uma breve enquête depois da entrega das provas. Os resultados são medidos em uma escala de 1 (nível mais baixo de ansiedade) a 10 (nível mais alto de ansiedade). A Tabela 17-3 mostra os dados baseados em uma amostra com 10 pares.

## Ansiedade na Hora da Prova

<i>Par</i>	<i>Nível de Ansiedade — Chiclete</i>	<i>Nível de Ansiedade — sem Chiclete</i>	<i>Diferença (Chiclete/sem Chiclete)</i>	<i>Sinal</i>
1	9	10	-1	-
2	6	8	-2	-
3	3	1	+2	+
4	3	5	-2	-
5	4	4	0	nenhum
6	2	7	-5	-
7	2	6	-4	-
8	8	10	-2	-
9	6	8	-2	-
10	1	3	-2	-



Os reais níveis de ansiedade não são importantes aqui, o que importa é a diferença entre os níveis de ansiedade dentro de cada par. Assim, em vez de analisar todos os níveis de ansiedade individual, basta analisar a diferença nos níveis de ansiedade de cada par. Este método faz com que você tenha apenas um conjunto de dados no lugar de dois. (Neste caso, para calcular as diferenças em cada par, você pode usar a fórmula: ansiedade sem chiclete menos ansiedade com chiclete e procurar uma diferença geral que seja positiva.)

Normalmente, no caso de dados em pares combinados, testa-se se a diferença mediana é igual a zero. Em outras palavras,  $H_0: m = 0$ ; o mesmo se mantém para o exemplo da ansiedade na hora da prova.

Agora, as diferenças nos níveis de ansiedade de cada par do conjunto de dados passam a ser um único conjunto de dados (veja a coluna quatro da Tabela 17-3). Assim, você pode usar o método do teste dos sinais para analisar esses dados, sendo que  $H_0: m = 0$  (nenhuma diferença mediana entre ansiedade com chiclete versus ansiedade sem chiclete) versus  $H_a: m < 0$  (mascar chiclete diminui a ansiedade na hora da prova).

Atribua a cada diferença um sinal positivo ou negativo; se for maior do que zero (sinal de mais), se for menor do que zero (sinal de menos). Sua estatística de teste é o número total de sinais positivos, 1, e o tamanho amostral em questão é  $10 - 1 = 9$ . (Não conte os dados cuja mediana for zero logo de cara).

Agora, compare esta estatística de teste à distribuição binomial com  $p = 0,50$  e  $n = 9$ , usando a tabela binomial (Tabela A-2 no apêndice). Você tem uma estatística de teste  $k = 1$  e quer encontrar a probabilidade de que  $x \leq 1$  (já que tem um teste unicaudal à esquerda, consulte o sexto passo para o teste dos sinais na seção anterior “Interpretando os Sinais: O Teste dos Sinais”). Embaixo da coluna  $p = 0,50$ , na seção para  $n = 9$ , encontra-se a probabilidade de 0,018 para  $x = 1$  e 0,002 para  $x = 0$ . Some esses valores para obter 0,020, o valor- $p$ . Esse resultado significa que você rejeita a  $H_0$  ao nível  $\alpha$  de 0,05, o que,



por sua vez, quer dizer que os níveis de ansiedade com chiclete versus sem chiclete são diferentes. Mas o quanto são diferentes? Com base nesses dados, conclui-se que mascar chiclete durante a prova parece diminuir a ansiedade, pois há mais diferenças negativas do que positivas.

# ***Um Passo Adiante com o Teste de Postos Sinalizados***

O teste de postos sinalizados é mais poderoso do que o teste dos sinais para detectar reais diferenças na mediana. O uso mais comum do teste de postos sinalizados está em testar dados em pares combinados em busca da diferença mediana causada por um tratamento (como mascar chiclete durante a prova e seu efeito sobre a ansiedade). Nesta seção, você vai descobrir o que é teste de postos sinalizados e como realizá-lo. Além disso, vou conduzi-lo através de uma aplicação que envolve o teste de um programa de emagrecimento.

## ***Uma limitação do teste dos sinais***

O teste de sinal tem a vantagem de ser muito simples e fácil de ser feito à mão. No entanto, uma vez que só analisa se um valor está acima ou abaixo da mediana, ele não leva em consideração a grandeza da diferença.

Analisando as Tabelas 17-1 e 17-3, você vê que para cada valor dos dados, a estatística de teste para o teste dos sinais só conta se cada valor dos dados for maior ou igual à mediana na hipótese nula,  $m_0$ . Ela não contabiliza o tamanho dessas diferenças. Por exemplo, na Tabela 17-3, você pode ver que o sexto par teve uma enorme redução na ansiedade quando mascou chiclete (de 7 para 2), mas, por outro lado, o primeiro par teve uma redução muito pequena na ansiedade (de 10 para 9). No entanto, ambas as diferenças receberam o mesmo resultado (um sinal de menos) na estatística de teste do teste dos sinais.

Uma vez que ela não leva em conta o quanto os valores nos dados se diferem da mediana, o teste dos sinais é menos potente (ou seja, menos capaz de detectar quando a  $H_0$  é falsa) do que poderia ser. Então, se quiser testar a mediana e quiser levar em conta a grandeza das diferenças (e estiver disposto a mergulhar de cabeça em alguns cálculos para chegar lá), pode realizar o *teste de postos sinalizados*, também conhecido como o *teste de postos sinalizados de Wilcoxon*. A próxima seção vai guiá-lo por ele.

## ***Seguindo os passos para realizar um teste de postos sinalizados***

Assim como o teste dos sinais, a única condição para um teste de postos sinalizados é a de que os dados sejam ordinais ou quantitativos.

A seguir, estão os passos para a realização de um teste de postos sinalizados para dados pareados:

### **1. Estabeleça sua hipótese nula:**

A hipótese nula é  $H_0: m = 0$ . Suas opções para a hipótese alternativa são  $H_a: m \neq 0$ ;  $H_a: m > 0$ ; ou  $H_a: m < 0$ , se você quiser detectar apenas uma diferença, uma diferença positiva ou uma negativa, respectivamente.

### **2. Colete uma amostra aleatória de dados pareados.**

3. Calcule a diferença para cada par de observação.
4. Calcule o valor absoluto de cada diferença.
5. Ordene os valores absolutos do menor para o maior.

Se dois dos valores absolutos estiverem empatados, atribua a cada um o posto médio dos dois valores. Por exemplo, se o quarto e o quinto números na ordem estiverem empatados, atribua a cada um o posto 4,5.

6. Some os postos que correspondem às diferenças originais positivas.

A soma das diferenças positivas é sua estatística de teste para os postos sinalizados, designada por T.


7. Encontre o valor- $p$ .

Busque todas as possíveis formas sob as quais as diferenças absolutas poderiam ter aparecido em uma amostra, com um sinal de mais ou menos, assumindo que a  $H_0$  seja verdadeira. Encontre todas as suas estatísticas de teste (valores-T) a partir de todas essas possíveis classificações usando do quarto ao sexto passos e compare o valor-T a elas. A porcentagem de valores-T que estão dentro ou fora de sua estatística de teste é seu valor- $p$ .

O Minitab pode fazer esse passo por você.

8. Tire suas conclusões:

Se o valor- $p$  for menor do que o nível predeterminado para  $\alpha$ , normalmente, 0,05, rejeite  $H_0$  e conclua que a diferença mediana não é zero. Caso contrário, você não pode rejeitar a  $H_0$ .



Para realizar o teste de postos sinalizados no Minitab, insira as diferenças (às quais se refere o terceiro passo) em uma única coluna. Clique em Stat>Nonparametric>One-sample Sign. Clique no nome da variável para suas diferenças no campo à esquerda e ela aparecerá no campo Variables à direita. Clique no círculo que diz Test Median e indique a  $H_a$  desejada ( $> 0$ ,  $< 0$  ou  $\neq$ ). Clique em OK, e seu teste está pronto. (Observe que, embora o Minitab calcule a estatística de teste para o teste de postos sinalizados de forma um pouco diferente da qual você usaria no cálculo manual, os resultados são próximos. A razão para essa sutil diferença no cálculo está além do escopo deste livro.)

O que fazer quando uma parte dos dados é exatamente igual à mediana? Na maioria das vezes (incluindo todos os conjuntos de dados que irá encontrar), esta ocorrência é rara, e o que pode ser feito é ignorar os valores de dados e reduzir o tamanho amostral em um a cada vez que isso ocorrer.

## *Emagrecendo com os postos sinalizados*

Esta seção mostra os postos sinalizados em ação. Primeiro, mostro cada passo como se

estivesse fazendo o processo à mão. Depois, veremos os resultados no Minitab.

Suponha que você queira testar se um programa de emagrecimento é eficaz ou não. Para isso, vai analisar o emagrecimento mediano dos participantes do programa por meio de um experimento de pares combinados. Você também quer que a grandeza do emagrecimento esteja na análise, o que significa que vai usar um teste de postos sinalizados para analisar os dados. Aqui estão os passos para a realização do teste para este exemplo:

1. Estabeleça sua hipótese nula:  $H_0:m = 0$ , onde  $m$  representa o emagrecimento mediano (antes do programa versus depois do programa). A hipótese alternativa, então, é  $H_a:m > 0$ , indicando que a diferença mediana do emagrecimento é positiva.
2. Colete uma amostra aleatória de, digamos, três pessoas e pese-as antes e depois de um programa de emagrecimento de oito semanas. Pode-se calcular a diferença do peso de cada pessoa (peso antes do programa menos peso depois do programa). Uma diferença positiva significa que a pessoa emagreceu, enquanto uma diferença negativa significa que ela engordou.

A Tabela 17-4 mostra os dados e as estatísticas relevantes para o teste de postos sinalizados. (Note que só tenho três pessoas neste estudo, pois o exemplo tem apenas uma finalidade ilustrativa.) As diferenças entre os pesos (antes – depois) estão na coluna quatro.

<b>Tabela 17-4      Dados Referentes a Emagrecimento Antes e Depois do Programa</b>					
<i>Pessoa</i>	<i>Antes</i>	<i>Depois</i>	<i>Diferença</i>	<i> Diferença </i>	<i>Posto</i>
1	200	205	–5	5	1
2	180	160	+20	20	2*
3	134	110	+24	24	3*

\* Representa os postos associados à diferença positiva.

3. Calcule o valor absoluto de cada diferença. Você pode vê-los na coluna cinco da Tabela 17-4.
4. Ordene as diferenças absolutas. A coluna seis reflete os postos ocupados pelos valores absolutos, de 1 a 3.
5. Encontre a estatística de teste, que é a soma dos postos correspondentes às diferenças positivas. (Em outras palavras, conte apenas os postos das pessoas que emagreceram.) Por este conjunto de dados, os postos que você deve contar estão indicados com um \* na Tabela 17-4. Essa soma, então, é  $2 + 3 = 5$ . Esse número, 5, é sua estatística de teste, e você pode chamá-lo de  $T$ .
6. Encontre o valor-p. Agora, você precisa comparar esta estatística de teste a algumas distribuições para ver onde ela se encontra. Para fazer isso, determine todas as possíveis formas sob as quais as três diferenças absolutas (coluna cinco da Tabela

17-4), 5, 20 e 24, poderiam ter aparecido em uma amostra, com suas diferenças reais, levando um sinal positivo ou negativo. (Assuma que a  $H_0$  é verdadeira e que as reais diferenças têm uma chance de 50% de serem positivas ou negativas, como o lançamento de uma moeda.)

Então, encontre todas as estatísticas de teste (valores- $T$ ) a partir de todos esses possíveis arranjos e compare o valor- $T$ , 5, a elas. A porcentagem dos outros valores- $T$  que estão dentro ou além de sua estatística de teste é o valor- $p$ .

Para o exemplo do programa de emagrecimento, existem oito possíveis maneiras de obter as diferenças absolutas 5, 20 e 24, incluindo tanto os sinais positivos quanto os negativos (dois possíveis sinais para cada diferença são iguais a  $2 * 2 * 2 = 8$ ). Essas oito possibilidades estão em colunas separadas na Tabela 17-5.  $T$  denota a soma dos postos positivos em cada caso (estas são as estatísticas de teste para cada possível arranjo).

<b>Tabela 17-5 Possíveis Amostras com as Diferenças Absolutas, 5, 20 e 24</b>								
<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>Posto da  Dif </i>
5	-5*	5	5	-5*	-5*	5	-5*	1
20	20	-20*	20	-20*	20	-20*	-20*	2
24	24	24	-24*	24	-24*	-24*	-24*	3
$T = 6$	$T = 5$	$T = 4$	$T = 3$	$T = 3$	$T = 2$	$T = 1$	$T = 0$	—

\* Denota as diferenças negativas.

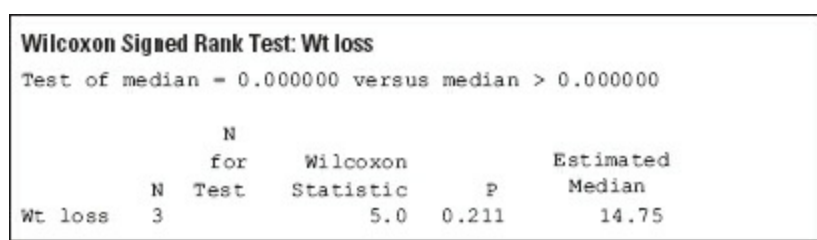
Para dar sentido à Tabela 17-5, considere o seguinte: as três diferenças absolutas que você tem em seu conjunto de dados são 5, 20 e 24, cujos postos são 1, 2 e 3, respectivamente (os quais podem ser vistos na Tabela 17-4). Você pode encontrar as oito diferentes combinações para 5, 20 e 24, em que pode colocar tanto um sinal de menos ou mais em qualquer um desses valores. Para cada cenário, encontre a estatística de teste pela soma dos postos das diferenças positivas (as pessoas que emagreceram). Tais postos estão na coluna nove da Tabela 17-5, representados pelos valores de dados sem asterisco (\*).

Por exemplo, a coluna sete tem duas diferenças negativas, -20 e -24, e uma diferença positiva, 5 (cujo posto entre as diferenças absolutas é 1, uma vez que é o menor valor; veja a coluna nove). Somando os postos positivos na coluna sete, obtemos uma estatística de teste ( $T$ ) igual a 1, pois 5 é o único número positivo. (Você pode ver na coluna dois os dados realmente observados na amostra.)

Agora, compare a estatística de teste, 5 (do quinto passo), a todos os valores de  $T$  na última linha da Tabela 17-5. Uma vez que a  $H_a$  é  $m > 0$ , você pode encontrar a porcentagem de postos sinalizados ( $T$ ) que são iguais ou maiores que 5. De oito você tem dois, assim, seu valor- $p$  (a porcentagem de possíveis estatísticas de teste iguais ou maiores do que sua  $H_0$ , caso ela seja verdadeira) é  $\frac{2}{8} = 0,25$ , ou 25%.

- Já que o valor- $p$  (0,25) é maior do que o nível predeterminado para  $\alpha$ , normalmente, 7. 0,05, você não pode rejeitar a  $H_0$  e, portanto, não pode dizer que há um emagrecimento positivo por meio do programa. (**Observação:** com uma amostra de apenas três, é difícil encontrar qualquer diferença real e, assim, esse programa de emagrecimento pode estar funcionando de fato, embora este pequeno conjunto de dados simplesmente não tenha sido capaz de determinar isso. E, além do mais, uma pessoa realmente engordou, o que não ajuda muito.)

A Figura 17-3 mostra a saída do Minitab para este teste, usando os dados da Tabela 17-4. O valor- $p$  acaba sendo 0,211, devido a uma ligeira diferença na maneira como o Minitab calcula a estatística de teste. Note que a mediana estimada, encontrada na Figura 17-3, refere-se a um cálculo feito sobre todas as possíveis amostras e as medianas obtidas a partir delas.



Wilcoxon Signed Rank Test: Wt loss

Test of median = 0.000000 versus median > 0.000000

	N	for	Wilcoxon		Estimated
	N	Test	Statistic	P	Median
Wt loss	3		5.0	0.211	14.75

**Figura 17-3:** Saída do Minitab para teste de postos sinalizados referente aos dados de emagrecimento.

Você também pode usar a estatística- $T$  para estimar a mediana de uma população (ou a mediana da diferença em um experimento de pares combinados). Para encontrar um intervalo de confiança de  $1 - \alpha$  por cento para a mediana, usando o Minitab com base no teste de postos sinalizados, insira os dados em uma única coluna. (Se os dados representarem as diferenças em um conjunto de dados em pares combinados, insira-as como uma coluna.) Clique em Stat>Nonparametrics>1-Sample Wilcoxon. Clique no nome da variável na coluna esquerda e ela aparecerá na coluna Variables no lado direito. Clique no círculo que diz: Confidence Interval e digite em  $1 - \alpha$  o valor que deseja para o nível de confiança. Clique em OK.



# Capítulo 18

## Subindo de Posto com o Teste das Somas dos Postos

---

### *Neste Capítulo*

- ▶ Comparando duas populações por meio das medianas, não das médias
  - ▶ Realizando o teste da soma dos postos
- 

***E***m Estatística I, quando se deseja comparar duas populações, realiza-se um teste de hipótese para duas médias populacionais. A ferramenta mais comum para a comparação de médias populacionais é o teste- $t$  (veja o Capítulo 3). No entanto, o teste- $t$  exige que os dados venham de uma distribuição normal. Quando as condições para os procedimentos paramétricos (aqueles que envolvem distribuições normais) não forem satisfeitas, sempre haverá uma alternativa não paramétrica para salvar o dia.

Neste capítulo, você vai trabalhar com um teste não paramétrico que compara os centros de duas populações — *o teste da soma dos postos*. Este teste tem como foco a *mediana*, a medida do centro mais adequada às situações em que os dados não são simétricos.

# *Realizando o Teste da Soma dos Postos*

Esta seção aborda as condições para o teste da soma de postos e percorre as etapas para sua realização. Além disso, você ainda poderá testar (literalmente) seu entendimento e sua habilidade na seção “Executando um teste das somas dos postos: qual corretor de imóveis vende casas mais rápido?” mais adiante neste capítulo.

## *Verificando as condições*

Antes de pensar em realizar um teste da soma dos postos para comparar as medianas de duas populações, você precisa se certificar de que os conjuntos de dados atendem às condições para o teste, que são as seguintes:

- ✓ **As duas amostras aleatórias, uma retirada de cada população, são independentes uma da outra.**

A primeira condição deve ser levada em conta no momento de coletar os dados. Certifique-se apenas de não utilizar pares combinados, por exemplo, usando os dados de uma mesma pessoa como em um pré versus pós-teste, pois, neste caso, os dois conjuntos de dados são dependentes.

- ✓ **As duas populações têm a mesma distribuição — isto é, seus histogramas têm a mesma forma.**

Você pode verificar essa condição fazendo histogramas para comparar as formas da amostra de dados das duas populações. (Consulte um livro de Estatística I ou meu livro *Estatística Para Leigos*, da Alta Books, para obter ajuda na construção de histogramas.)

- ✓ **As duas populações têm a mesma variância, ou seja, a dispersão dos valores é a mesma.**

Você pode verificar essa condição encontrando as variâncias ou os desvios padrão das duas amostras, os quais devem ser próximos. (Existe um teste de hipótese para duas variâncias, mas ele está fora do escopo deste livro.)

Observe que os centros das duas populações não precisam ser iguais; afinal, é isso o que o teste vai decidir.

Os métodos mais sofisticados para a verificação da segunda e terceira condições listadas aqui estão além do escopo deste livro. No entanto, os métodos de verificação dessas condições que descrevo aqui lhe permitem ficar livre de quaisquer maiores problemas.

## *Seguindo os passos para a realização de um teste*

O teste da soma dos postos é um teste para a igualdade entre duas medianas populacionais — as quais vamos chamar de  $\eta_1$  e  $\eta_2$ . Depois de ter verificado as condições de utilização





do teste da soma dos postos (veja a seção anterior), realize o teste seguindo estes passos. (**Observação:** o Minitab pode executar este teste para você, mas, ainda assim, você deve saber o que está se passando nos bastidores.)

1. Estabeleça a  $H_0: \eta_1 = \eta_2$  versus  $H_a: \eta_1 > \eta_2$  (um teste unilateral);  $H_a: \eta_1 < \eta_2$  (um teste unilateral), ou  $H_a: \eta_1 \neq \eta_2$  (um teste bilateral), se estiver procurando uma diferença positiva, uma diferença negativa ou qualquer diferença entre as duas medianas populacionais, respectivamente.
2. Veja os dados como um grupo combinado e ordene os valores dos mais baixos (posto = 1) para os mais altos.

No caso de empate, atribua a ambos os valores da média dos postos que teriam recebido. Por exemplo, suponha que o terceiro e o quarto valores na ordem sejam iguais. Se fossem diferentes, receberiam os postos 3 e 4, respectivamente. Mas, uma vez que são iguais, atribua a eles o mesmo posto, 3,5, que é a média entre 3 e 4. Observe que o número seguinte (na ordem) é o quinto número, o que vai receber o posto 5.

3. Some os postos atribuídos à amostra que tiver o menor tamanho; chame essa estatística de  $T$ .

Nesse passo, a menor amostra é usada por uma questão de convenção — os estatísticos gostam de ser coerentes. Se os tamanhos amostrais forem iguais, some os postos da primeira amostra para obter  $T$ . Se o valor de  $T$  for pequeno (em relação à soma total de todos os postos dos dois conjuntos de dados), isso significa que os números da primeira amostra tendem a ser menores do que os da segunda amostra, portanto, a mediana da primeira população pode ser menor do que a mediana da segunda.

4. Consulte as tabelas da soma de postos (Tabelas A-4 (a) e (b) no apêndice). Na tabela escolhida, encontre a coluna e a linha para o tamanho amostral do grupo um e dois, respectivamente.

Aqui, você vai ver dois valores críticos,  $VCI$  (o menor valor crítico) e  $VCS$  (o maior valor crítico). Estes valores críticos fazem as fronteiras entre rejeitar a  $H_0$  e não rejeitá-la.

5. Compare sua estatística de teste,  $T$ , para os valores críticos na Tabela A-4 no apêndice para concluir se pode rejeitar a  $H_0$  — a hipótese de que as medianas populacionais não são diferentes.

O método utilizado para comparar estes valores depende do tipo de teste que está sendo realizado:

- **Teste unilateral ( $H_a$  tem um sinal de  $>$  ou  $<$ ):** A Tabela A-4 no apêndice mostra os valores críticos para o nível  $\alpha$  de 0,05. Para um teste unilateral à direita (o que significa que a  $H_a: \eta_1 > \eta_2$ ), rejeite a  $H_0$  se  $T \geq VCS$ . Para um teste



unilateral à esquerda (o que significa que a  $H_a: \eta_1 < \eta_2$ ), rejeite  $H_0$  se  $T \leq VCI$ .

Se você rejeitar  $H_0$ , conclua que as medianas populacionais são diferentes e que uma delas é maior que a outra, dependendo da  $H_a$ . (Caso contrário, você não poderá concluir que há diferença entre as medianas.)

- **Teste bilateral:** A Tabela A-4 no apêndice mostra os valores críticos para o nível  $\alpha$  de 0,025. Rejeite a  $H_0$  se  $T$  estiver fora do intervalo ( $VCI, VCS$ ); ou seja, rejeite a  $H_0$  quando  $T \leq VCI$  ou  $T \geq VCS$ . Conclua, então, que as medianas populacionais não são iguais. (Caso contrário, você não poderá concluir que há diferença entre elas.)

Para realizar um teste da soma dos postos no Minitab, insira os dados da primeira amostra em Column 1 e os dados da segunda amostra em Column 2. Clique em Stat>Nonparametrics>Mann-Whitney. Clique no nome da variável da coluna 1 e ela aparecerá no campo First Sample. Clique no nome da variável na coluna 2 e ela aparecerá no campo Second Sample. Ao clicar em Alternative, você vai abrir um menu para selecionar se sua  $H_a$  é igual, maior ou menor (o que vai depender do problema em questão). Clique em OK e o teste está pronto.

## *Aumentando o tamanho da amostra*

Depois que os tamanhos amostrais chegam a um certo ponto, os valores da tabela se esgotam. A Tabela A-4 no apêndice (que mostra os valores críticos para a rejeição de  $H_0$  para o teste da soma de postos) só mostra os valores críticos para tamanhos amostrais entre três e dez. Se as amostras forem maiores do que dez, use o teste- $Z$  para duas amostras para obter uma aproximação para sua resposta. Isso porque, para os tamanhos amostrais grandes, a estatística de teste- $T$  para o teste da soma dos postos se assemelha a uma distribuição normal. (Então, por que não usá-la? Já que pode deixar o ônus da prova para os profissionais!) Quanto maior forem os dois tamanhos amostrais, mais precisa será a aproximação.

Portanto, se todos os tamanhos amostrais forem maiores do que dez, conduza os passos de um a três do teste da soma de postos normalmente. Porém, em vez de consultar o valor de  $T$  na Tabela A-4 no apêndice, como pede o quarto passo do teste da soma de postos, transforme-o em um valor- $Z$  (um valor na distribuição normal padrão), subtraindo-o de sua média e dividindo o resultado pelo erro padrão.

$$Z = \frac{T - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

A fórmula usada para obter este valor- $Z$  para a estatística de teste é  $Z = \frac{T - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$ , onde  $T$  é dado no terceiro passo da seção anterior,  $n_1$  é o tamanho amostral do primeiro conjunto de dados (retirado da primeira população) e  $n_2$  é o tamanho amostral do segundo conjunto de dados (retirado da segunda população). Depois de conseguir o valor- $Z$ , siga os mesmos procedimentos para qualquer teste que envolva um valor- $Z$ , como o teste para duas médias



populacionais. Isto é, encontre o valor  $p$  consultando o valor- $Z$  em uma tabela- $Z$  (consulte um livro de Estatística I ou o meu livro *Estatística Para Leigos*, da editora Alta Books), ou procure na última linha da tabela- $t$  (que pode ser encontrada no apêndice), e encontre a área fora dele. (Se o teste for bilateral, dobre o valor- $p$ .) Se o valor- $p$  for menor do que o nível predeterminado para  $\alpha$ , rejeite a  $H_0$ . Caso contrário, você não pode rejeitá-la.



Mesmo quando  $n$  for grande e você usar um valor- $Z$  para a estatística de teste, ainda poderá usar o Minitab (na verdade, é o recomendado se quiser evitar o tédio de ter que manipular um exemplo grande). Consulte a seção anterior “Seguindo os passos para a realização de um teste” para encontrar as instruções do Minitab.

# Realizando um Teste da Soma dos Postos: Qual Corretor de Imóveis Vende Casas Mais Rápido?

Suponha que você precise de um corretor de imóveis para vender sua casa e conhece dois corretores que trabalham em seu bairro. Seu critério mais importante é o de vender a casa rapidamente, então, decide descobrir qual corretor vende casas mais rápido. Para isso, coleta uma amostra aleatória de oito casas que cada corretor vendeu no ano passado e registra o número de dias que cada casa ficou no mercado antes de ser vendida. Veja os dados que estão na Tabela 18-1.

**Tabela 18-1**                      **Tempo no Mercado das Casas Vendidas por dois Corretores de Imóveis**

	<i>Corretora Suzy “Venderápido”</i>	<i>Corretor Tommy “Nãoperdetempo”</i>
Casa 1	48 dias	109 dias
Casa 2	97 dias	145 dias
Casa 3	103 dias	160 dias
Casa 4	117 dias	165 dias
Casa 5	145 dias	185 dias
Casa 6	151 dias	250 dias
Casa 7	220 dias	251 dias
Casa 8	300 dias	350 dias

Confira os dados resumidos em *boxplots* (um gráfico que resume os dados, mostrando o seu valor mínimo, primeiro quartil, mediana, terceiro quartil e valores máximos) na figura 18-1a e as estatísticas descritivas na Figura 18-1b. Nas seções a seguir, você vai usar esses dados para ver o teste da soma de postos em ação. Prepare-se para se surpreender.

Para fazer dois boxplots paralelos usando o Minitab, clique em Graph>Boxplots>Simple Multiple Y’s. Clique em cada uma das duas variáveis no campo à esquerda e elas aparecerão no campo Variables à direita. Clique em OK.

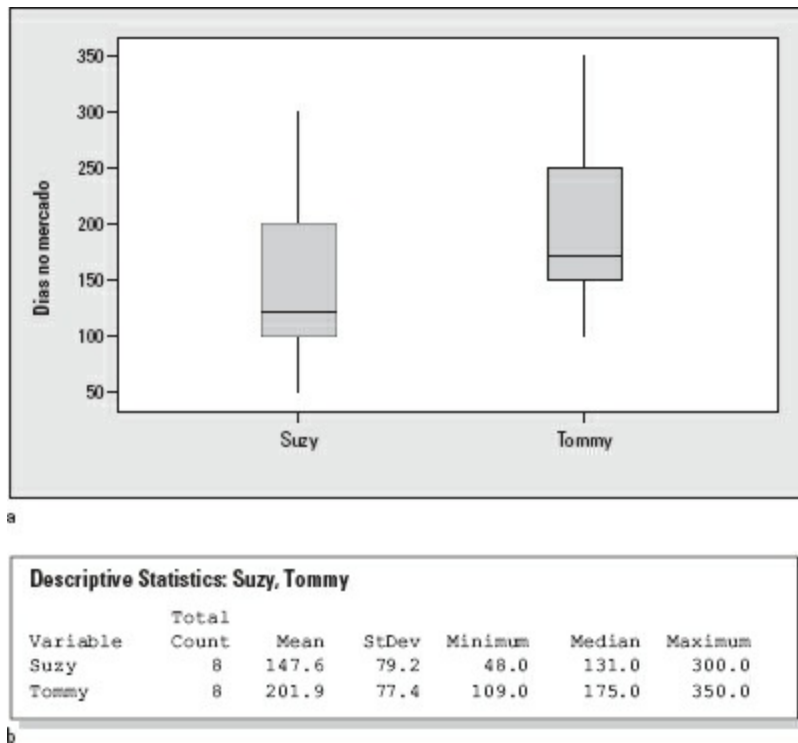


## Verificando as condições para este teste

Ao verificar as condições, você descobre que os dados das duas amostras são independentes, uma vez que Suzy e Tommy são concorrentes.

Os boxplots na Figura 18-1a mostram a mesma forma básica e variabilidade para cada conjunto de dados. (Não há dados suficientes para que você faça histogramas e vá mais a fundo na verificação.) Assim, com base nesses dados, é razoável supor que as distribuições para os dados referentes aos dias no mercado são iguais para os dois corretores. Na Figura 18-1b, os desvios padrão da amostra são próximos: 79,2 dias para

Suzy e 77,4 dias para Tommy. Como os dados satisfazem as condições para o teste da soma de postos, pode ir em frente e aplicá-los na análise de seus dados.



**Figura 18-1:** Boxplots e estatísticas descritivas para dados referentes aos corretores de imóveis.

Para encontrar as estatísticas descritivas (como o desvio padrão) no Minitab, clique em Stat>Basic Statistics>Display Descriptive Statistics. Clique em Options, depois no campo para cada estatística que deseja calcular. Caso algum campo esteja assinalado para uma estatística que você não deseja, clique nele novamente para que a marcação desapareça.

A Figura 18-1b mostra que a mediana para Suzy (131 dias no mercado) é menor do que a mediana para Tommy (175 dias). Então, provavelmente, você diria que a Suzy vende casas mais rápido e pronto. Mas a mediana não conta toda a história. Observando a Figura 18-1, é possível ver que uma parte dos dois boxplots se sobrepõe. Isso significa que alguns tempos de venda da Suzy e do Tommy foram próximos. Além disso, também há uma grande variabilidade nos tempos de venda de cada corretor, conforme você pode ver pelo conjunto de valores em cada boxplot. Isto lhe diz que as evidências não são totalmente claras. Embora Suzy possa realmente ser a corretora mais rápida, não é possível dizer isso com certeza se compararmos apenas os boxplots das duas amostras. É preciso um teste de hipótese para fazer a avaliação final.

## Testando a hipótese

A hipótese nula para o teste é  $H_0: \eta_1 = \eta_2$ , onde  $\eta_1$  = número mediano de dias no mercado para a população de todas as casas que Suzy vendeu no ano passado, e  $\eta_2$  = número

mediano de dias no mercado para a população de todas as casas que Tommy vendeu no ano passado. A hipótese alternativa é  $H_a: \eta_1 \neq \eta_2$ .



Suponha que você tenha examinado os dados e tido um palpite de que o corretor que vende casas mais rápido seja Suzy. No entanto, antes de ver os dados, você não tinha nenhuma ideia a respeito de quem era o mais rápido. Você deve basear suas  $H_0$  e  $H_a$  no que pensava *antes de* ver os dados, e não depois. Estabelecer as hipóteses depois de coletar os dados é injusto e antiético.

Depois de determinar as  $H_0$  e  $H_a$ , é chegada a hora de testar os dados.

### ***Combinando e classificando***

O primeiro passo na análise de dados é combinar todos os dados e classificar os dias no mercado do menor (posto = 1) para o maior. Veja todos os postos para os dados combinados na Tabela 18-2.

No caso de empate, atribua a ambos os valores da média dos postos que teriam ocupado. Na Tabela 18-2, é possível ver que há dois 145 no conjunto de dados. Uma vez que representam o sexto e o sétimo números no conjunto de dados ordenado, atribua a cada um deles o posto  $(6 + 7) \div 2 = 6,5$ .

**Tabela 18-2**                      **Postos dos Dados Combinados para o Exemplo do Corretor de Imóveis**

<i>Suzy</i> <i>“Venderápido”</i>	<i>Postos Tommy</i>	<i>“Naoperdetempo”</i>	<i>Postos</i>
48 dias	1	109 dias	4
97 dias	2	145 dias	6,5
103 dias	3	160 dias	9
117 dias	5	165 dias	10
145 dias	6,5	185 dias	11
151 dias	8	250 dias	13
220 dias	12	251 dias	14
300 dias	15	350 dias	16

### ***Calculando a estatística de teste***

Depois de ter classificado seus dados, determine qual dos grupos é o grupo um, para que, assim, você possa encontrar sua estatística de teste,  $T$ . Como os tamanhos amostrais são iguais, vamos dizer que o grupo um é o grupo de Suzy, pois seus dados aparecem primeiro. Agora, some os postos do conjunto de dados de Suzy. A soma dos postos de Suzy é  $1 + 2 + 3 + 5 + 6,5 + 8 + 12 + 15 = 52,5$ ; este valor de  $T$  é a sua estatística de teste.

## Rejeitar ou não a $H_0$

Suponha que você queira usar o nível  $\alpha$  de 0,05 para este teste; usar esse valor de corte pressupõe a utilização da Tabela A-4(a) no apêndice, pois se trata de um teste bilateral ao nível  $\alpha = 0,05$  com 0,025 para cada lado. Procure a coluna para  $n_1 = 8$  e a linha para  $n_2 = 8$ , onde vai encontrar  $VCI = 49$  e  $VCS = 87$ . Rejeite a  $H_0$  se  $T$  estiver fora deste intervalo; ou seja, rejeite a  $H_0$  se  $T \leq VCI = 49$  ou  $T \geq VCS = 87$ . A estatística  $T = 52,5$  não está fora desse intervalo; portanto, você não tem provas suficientes para rejeitar  $H_0$  ao nível  $\alpha = 0,05$  e, conseqüentemente, não pode dizer que há uma diferença significativa no número mediano de dias no mercado para Suzy e Tommy.

Estes resultados podem parecer muito estranhos, dado o fato de que as medianas para os dois conjuntos de dados foram bem diferentes: 131 dias no mercado para Suzy e 175 dias para Tommy. No entanto, duas coisas pesam contra você quando se trata da capacidade de encontrar uma diferença real para o exemplo em questão:

- ✓ **Os tamanhos amostrais são muito pequenos (apenas oito para cada grupo).**  
Quanto menor a amostra, mais difícil é obter provas suficientes para rejeitar a  $H_0$ .
- ✓ **Os desvios padrão dos dois grupos estão na casa dos 70, bastante elevados em comparação às medianas.**

Os dois problemas fazem com que o teste tenha dificuldade em encontrar alguma coisa através de toda a variabilidade demonstrada pelos dados. Em outras palavras, ele tem uma potência baixa (veja o Capítulo 3).

Para realizar um teste da soma dos postos usando o Minitab, clique em Stat>Nonparametric>Mann-Whitney. Selecione suas duas amostras e escolha a  $H_a$ :  $>$ ,  $<$  ou  $\neq$ . O nível de confiança é igual a um menos seu valor de  $\alpha$ . Depois de fazer todas essas configurações, clique em OK.

A Figura 18-2 mostra a saída do Minitab para o teste da soma dos postos ou para o teste de Mann-Whitney, para os dados do exemplo dos corretores de imóveis. Para interpretar os resultados na Figura 18-2, é preciso notar que o Minitab usa ETA no lugar de  $\eta$  para se referir às medianas. Os resultados na parte inferior da saída indicam que o teste para medianas iguais (versus medianas diferentes) é significativo ao nível 0,1149, quando os empates são ajustados. Este é seu valor- $p$  ajustado para empates. (Se não houver empate em seus dados, use os resultados que estiverem logo acima dessa linha. Isso lhe dará o valor- $p$  não ajustado a empates.) Para concluir, compare o valor- $p$  ao nível  $\alpha$  predeterminado (normalmente, 0,05). Se o valor- $p$  for igual ou menor do que 0,05, rejeite a  $H_0$ ; caso contrário, você não pode fazer isso. Neste caso, como 0,1149 é maior do que 0,05, não é possível rejeitar  $H_0$ . Isso significa que, com base nesses dados, você não tem provas suficientes para dizer que há diferença entre as medianas populacionais para os dias em que as casas vendidas por Suzy versus Tommy ficaram no mercado. Estes resultados confirmam as conclusões tiradas na seção anterior.



#### Mann-Whitney Test and CI: Suzy, Tommy

	N	Median
Suzy	8	131.0
Tommy	8	175.0

Point estimate for ETA1-ETA2 is -49.0  
95.9 Percent CI for ETA1-ETA2 is (-137.0, 36.0)  
W = 52.5  
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1152  
The test is significant at 0.1149 (adjusted for ties)

**Figura 18-2:** Teste da soma dos postos para descobrir quem vende casas mais rápido.

A saída do Minitab ilustrada na Figura 18-2 fornece também um intervalo de confiança para a diferença entre as medianas das duas populações, com base nos dados dessas duas amostras. A diferença entre as medianas da amostra (Suzy – Tommy) é  $131,0 - 175,0 = -44,0$ . Ao adicionar e subtrair a margem de erro (esses cálculos estão fora do escopo deste livro), o Minitab encontra o intervalo de confiança para a diferença entre as medianas (Suzy – Tommy) como sendo  $-137,0, +36,0$ , a diferença entre as medianas das populações pode ser qualquer valor entre  $-137,0$  e  $36,0$ . Como 0, o valor na  $H_0$ , também está neste intervalo, você não pode rejeitar a  $H_0$ . Então, novamente, você não pode dizer que as medianas são diferentes apenas com base nesse conjunto (limitado) de dados.

## Usando o teste da soma dos postos para comparar as notas dadas por juízes de competição

Você pode usar o teste da soma de postos para comparar dois grupos de juízes de uma competição para ver se há diferença em suas notas. Por exemplo, nas competições de patinação artística suspeita-se, às vezes, que o sexo dos juízes influencie a pontuação dada a determinados patinadores. Imagine uma competição masculina de patinação artística com dez juízes: cinco homens e cinco mulheres. Suponha que você queira saber se os juízes do sexo masculino e feminino usam os mesmos critérios para dar as notas aos competidores e, portanto, faz um teste da soma de postos para comparar as medianas das pontuações. Suas hipóteses são  $H_0$ : os juízes do sexo masculino e feminino têm a mesma pontuação mediana versus  $H_a$ : eles têm pontuações medianas diferentes. Para sua amostra, deixe que cada juiz dê as notas para todos os competidores. Então, ordene as notas da menor para a maior e as chame de M para as notas dos juízes do sexo masculino e F para as notas dos juízes do sexo feminino. Veja o resultado a seguir: F, M, M, M, M, F, F, F, F, M. O valor da estatística de teste T é a soma dos postos do grupo um (o dos homens), que é  $T = 2 + 3 + 4 + 5 + 10 = 24$ . Agora, compare-a aos valores críticos na Tabela A-4 no apêndice, onde os dois tamanhos amostrais são iguais a cinco,  $VCI = 18$  e  $VCS = 37$ . Como a estatística de teste,  $T = 24$ , ela está dentro deste intervalo, e você não deve rejeitar a  $H_0$ . Os critérios de julgamento são os mesmos para os juízes do sexo masculino e feminino. Nesta situação, você não tem provas suficientes para dizer que há uma diferença.





# Capítulo 19

## Faça o Kruskal-Wallis e Ordene as Somas com Wilcoxon

---

### *Neste Capítulo*

- ▶ Comparando mais de duas medianas populacionais com o teste de Kruskal-Wallis
  - ▶ Determinando quais populações são diferentes com o teste da soma de postos de Wilcoxon
- 

**O**s estatísticos da área não paramétrica transformaram seu trabalho na busca constante por procedimentos não paramétricos (aqueles que não dependem da distribuição normal) que sejam equivalentes aos paramétricos. E, no caso de ter que comparar mais do que duas populações, esses super-heróis da Estatística não nos decepcionaram. Neste capítulo, você vai ver como o teste de Kruskal-Wallis, um procedimento não paramétrico, compara mais de duas populações versus o seu correspondente paramétrico, a ANOVA (veja o Capítulo 9). Se o teste de Kruskal-Wallis lhe diz que pelo menos duas populações são diferentes, este capítulo também vai ajudá-lo a descobrir como usar o teste da soma dos postos de Wilcoxon para determinar a população que é diferente da mesma forma que os procedimentos de comparações múltiplas são usados na ANOVA (consulte o Capítulo 10).

# Fazendo o Teste de Kruskal-Wallis para Comparar Mais de Duas Populações

O teste de Kruskal-Wallis compara as medianas das várias (mais do que duas) populações para ver se elas são ou não diferentes. A ideia básica de Kruskal-Wallis é coletar uma amostra de cada população, classificar todos os dados combinados do menor ao maior e depois procurar um padrão na forma como os postos se distribuem entre as várias amostras. Por exemplo, se uma amostra obtém todos os postos mais baixos e a outra amostra fica com todos os mais altos, talvez as medianas de suas populações sejam diferentes. Ou se todas as amostras tiverem postos semelhantes, talvez as medianas das populações sejam consideradas iguais. Nesta seção, você vai ver como conduzir o teste de Kruskal-Wallis usando os postos, as somas e todas essas coisas interessantes, além de vê-lo aplicado a um exemplo que compara as avaliações de três companhias aéreas.

Suponha que seu chefe viaje muito e queira que você determine qual de três companhias aéreas é melhor classificada pelos clientes. Sabendo que as classificações envolvem dados não normais (trocadilho intencional), você pode optar pelo teste de Kruskal-Wallis. Colete três amostras aleatórias de nove pessoas que viajaram com as três companhias aéreas. Peça a cada pessoa que classifique sua satisfação com relação à companhia aérea. Cada pessoa deve usar uma escala de 1 (a pior) a 4 (a melhor). Veja os dados das amostras na Tabela 19-1.



Você pode estar pensando em usar a ANOVA, o teste que compara as médias de várias populações (veja o Capítulo 9), para analisar esses dados. Mas os dados de cada companhia consistem em classificações de 1 a 4, o que não satisfaz a mais forte das condições para a realização da ANOVA — os dados de cada população devem seguir uma distribuição normal. (A *distribuição normal* é contínua, o que significa que ela assume todos os números reais em um determinado intervalo. Dados que são números inteiros, como 1, 2, 3 e 4, não se enquadram nessa categoria.) Mas, não se preocupe: sempre existe uma alternativa não paramétrica para salvar o dia. O teste de Kruskal-Wallis compara as medianas das várias (mais do que duas) populações para ver se elas são ou não diferentes. Em outras palavras, é como a ANOVA, salvo pelo fato de que usa medianas, e não médias.

**Tabela 19-1**                      **Classificação dos Clientes com Relação a Três Companhias Aéreas**

<i>Companhia Aérea A</i>	<i>Companhia Aérea B</i>	<i>Companhia Aérea C</i>
4	2	2
3	3	3
4	3	3
4	3	2
3	4	2

3	4	1
2	3	3
3	4	2
4	3	2

Ao examinar os dados da Tabela 19-1, parece que as companhias A e B têm melhores classificações do que a companhia C. No entanto, os dados têm uma grande variabilidade e, por isso, é preciso realizar um teste de hipótese antes de tirar qualquer conclusão a partir deste conjunto de dados.

Nesta seção, você vai descobrir como verificar as condições para o teste de Kruskal-Wallis, além de configurá-lo e executá-lo passo a passo.

## Verificando as condições

Todas as condições a seguir precisam ser satisfeitas a fim de que o teste Kruskal-Wallis possa ser realizado.

- ✓ As duas amostras aleatórias, retiradas de cada população, devem ser independentes uma da outra. (Isso significa que os dados combinados, como os do Capítulo 17, estão fora de cogitação.)
- ✓ Todas as populações devem ter a mesma distribuição, ou seja, suas formas, vistas em um histograma, devem ser iguais. (Note que o tipo de distribuição não é especificado.)
- ✓ As variâncias das populações devem ser iguais. A dispersão dos valores da população deve ser igual de uma população para outra.

Note que tais condições mencionam a forma e a dispersão, mas não o centro da distribuição. Este teste tenta determinar se o centro das populações está ou não no mesmo lugar.

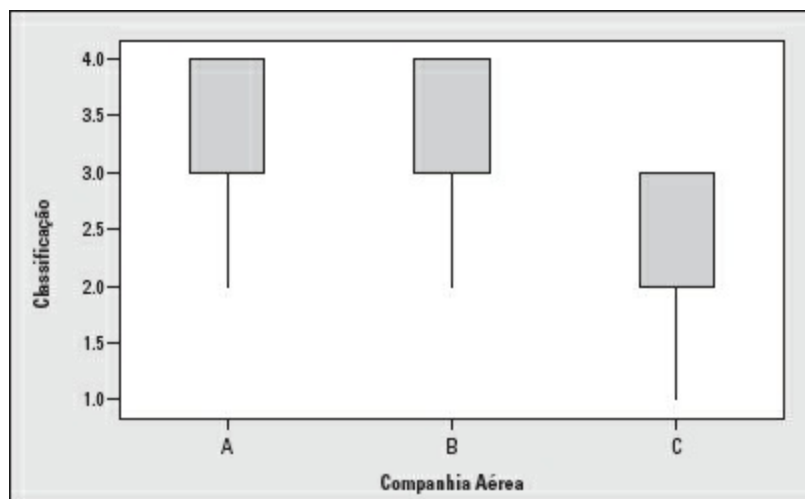
Na estatística não paramétrica, muitas vezes você vai ver a palavra *localização* sendo usada em referência a uma distribuição populacional em vez da palavra *centro*, embora as duas signifiquem a mesma coisa. A localização indica o local ocupado pela distribuição. Se você tem duas curvas em forma de sino com a mesma variância e uma delas tem a média igual a 10 e a outra tem média 15, a segunda distribuição está localizada a 5 unidades à direita da primeira. Ou seja, sua localização é 5 unidades à direita da primeira distribuição. Na estatística não paramétrica, onde não se tem uma distribuição em forma de sino, normalmente se usa a mediana como medida de localização de uma distribuição. Então, ao longo de toda esta discussão, você poderia usar a palavra mediana em vez de localização (embora localização seja mais ampla).

Em relação à pesquisa com as companhias aéreas, sabe-se que as amostras são independentes, pois você não usou a mesma pessoa para classificar mais de uma



companhia aérea. As outras duas condições têm a ver com as distribuições de onde as amostras vieram; cada população deve ter a mesma forma e a mesma dispersão. Você pode examinar ambas as condições através dos boxplots dos dados (veja a Figura 19-1) e das estatísticas descritivas, tais como a mediana, o desvio padrão e o resto das estatísticas de síntese que compõem os boxplots (veja a Figura 19 2).

Todos os boxplots na Figura 19-1 têm a mesma forma e seus desvios padrão, mostrados na Figura 19-2, são muito próximos. Todas estas evidências lhe permitem prosseguir com o teste de Kruskal-Wallis. (Ao examinar a sobreposição nos boxplots para as companhias A e B na Figura 19-1, você também pode fazer uma previsão inicial de que as companhias aéreas A e B foram classificadas de forma semelhante. No entanto, é impossível dizer se C é ou não suficientemente diferente de A e B sem um teste de hipótese.)



**Figura 19-1:** Boxplots comparando as classificações de três companhias aéreas.

Descriptive Statistics: Rating							
Variable	Airline	StDev	Minimum	Q1	Median	Q3	Maximum
Rating	A	0.707	2.000	3.000	3.000	4.000	4.000
	B	0.667	2.000	3.000	3.000	4.000	4.000
	C	0.667	1.000	2.000	2.000	3.000	3.000

**Figura 19-2:** Estatísticas descritivas comparando as classificações de três companhias aéreas.



Tanto um boxplot quanto um histograma podem revelar a forma e a dispersão de uma distribuição (bem como o centro). O *boxplot* é um tipo comum de gráfico usado para procedimentos não paramétricos, pois mostra a mediana (estatística não paramétrica escolhida) em vez da média. Na melhor das hipóteses, um *histograma* mostra o formato dos dados, mas não diz diretamente onde fica o centro — você só terá uma ideia de sua localização.



Observe que o boxplots na Figura 19-1 não são atravessados por linhas, como era de se esperar. Isso porque, em cada caso, a mediana pode ser igual a Q1, o primeiro quartil (veja

a Figura 19-2). Tal situação pode acontecer com os dados de classificação quando muitas classificações assumem o mesmo valor.

Para fazer com que os boxplots de cada amostra de dados apareçam lado a lado em um gráfico (sabidamente chamados *boxplots paralelos*) no Minitab, clique em Gráfico> Box Plots e selecione Multiple Y's Simple version. No lado esquerdo, clique em cada um dos nomes das colunas para seus conjuntos de dados. Cada uma delas vai aparecer na janela Graph Variables à direita. Clique em OK e você terá um conjunto de boxplots paralelos, todos no mesmo gráfico usando a mesma escala (interessante, não?).

## ***Estabelecendo o teste***

O teste de Kruskal-Wallis avalia a  $H_0$ : todas as  $k$  populações têm a mesma localização versus  $H_a$ : as localizações de pelo menos duas das  $k$  populações são diferentes. Aqui,  $k$  é o número de populações que estão sendo comparadas.

Em  $H_0$ , você vê que todas as populações têm a mesma localização (o que significa que todas estão uma em cima da outra na reta numérica e são, em essência, a mesma população). A  $H_a$ , neste caso, procura a situação oposta. No entanto, o oposto de “as localizações são todas iguais” não é “as localizações são todas diferentes”. O oposto é que pelo menos duas delas são diferentes. Ao não reconhecer essa diferença, você é levado a acreditar que todas as populações são diferentes quando, na realidade, apenas duas se diferem, enquanto as outras são todas iguais. É por isso que, no teste de Kruskal-Wallis, a  $H_a$  é colocada do jeito que você viu. (A mesma ideia vale para a comparação de médias através da ANOVA, veja o Capítulo 9.)

Para o exemplo da classificação das companhias aéreas (veja a Tabela 19-1), as hipóteses são as seguintes:  $H_0$  — as avaliações da satisfação de todas as três companhias aéreas têm a mesma mediana versus  $H_a$  — as medianas dos índices de satisfação de pelo menos duas companhias aéreas são diferentes.

## ***Realizando o teste passo a passo***

Depois de ter determinado suas hipóteses e verificado as condições, você pode realizar o teste. Aqui, estão os passos para a realização do teste de Kruskal-Wallis utilizando o exemplo da companhia aérea para mostrar o funcionamento de cada etapa:

- 1. Classifique todos os números do conjunto de dados do menor para o maior (utilizando todas as amostras combinadas); no caso de empate, utilize a média dos postos que os valores teriam ocupado caso não tivessem empatado.**

A Figura 19-3 mostra os resultados para a classificação e soma dos dados do exemplo das companhias aéreas; veja como classificar os empates. Por exemplo, você tem apenas um 1, que ocupa o posto número 1. Porém, você tem sete 2 que teriam ocupado os postos 2, 3, 4, 5, 6, 7 e 8. Uma vez que os 2 são todos iguais, dê a cada um deles a

média de todos esses postos, que é  $\frac{(2+3+4+5+6+7+8)}{7} = 5$ . Da 7 mesma forma, vemos doze 3, cujos postos seriam de 9 a 20. Uma vez que são todos iguais, dê a cada um deles um posto equivalente a  $\frac{(9+10+\dots+20)}{12} = 14,5$ . Finalmente, vemos sete 4, cada um ocupando um posto 24, que é a média do que seriam seus postos, de 21 a 27.

Companhia Aérea A		Companhia Aérea B		Companhia Aérea C	
Classificação	Posto	Classificação	Posto	Classificação	Posto
4	24	2	5	2	5
3	14,5	3	14,5	3	14,5
4	24	3	14,5	3	14,5
4	24	3	14,5	2	5
3	14,5	4	24	2	5
3	14,5	4	24	1	1
2	5	3	14,5	3	14,5
3	14,5	4	24	2	5
4	24	3	14,5	2	5
	$T_1 = 159$		$T_2 = 149,5$		$T_3 = 69,5$

**Figura 19-3:** Classificação e soma de postos para o exemplo das companhias aéreas.

2. Some os postos de cada uma das amostras; vamos chamar esses totais de  $T_1$ ,  $T_2$ , ...,  $T_k$ , onde  $k$  é o número de populações.

Os totais dos postos em cada coluna da Figura 19-3 são  $T_1 = 159$  (total dos postos para companhia A),  $T_2 = 149,5$  e  $T_3 = 69,5$ . Nas etapas que se seguem, você vai usar esses totais na estatística de teste para o teste de Kruskal-Wallis (indicada por KW). (Note que  $T_1$  e  $T_2$  estão quase iguais, mas  $T_3$  é muito menor, dando a impressão de que a companhia C pode ser carta fora do baralho).

3. Calcule a estatística de teste Kruskal-Wallis,  $KW = \frac{12}{n(n+1)} \sum \frac{T_j^2}{n_j} - 3(n+1)$ , onde  $n$  é o número total de observações (todos os tamanhos amostrais combinados).

Para o exemplo das companhias aéreas, a estatística de teste Kruskal-Wallis é  $KW = \frac{12}{27(27+1)} \left( \frac{159^2}{9} + \frac{149,5^2}{9} + \frac{69,5^2}{9} \right) - 3(27+1)$ , que equivale a  $0,0159 * 5.829,056 - 3(28) = 8,52$ .

4. Encontre o valor- $p$  para sua estatística de teste KW comparando-a à distribuição do Qui-quadrado com  $k - 1$  graus de liberdade (consulte a Tabela A-3 no apêndice).

Para o exemplo das companhias aéreas, na tabela do Qui-quadrado (Tabela A-3 do apêndice) você deve encontrar a linha com  $3 - 1 = 2$  graus de liberdade. Depois, veja onde sua estatística de teste (8,52) se encontra nessa linha. Como 8,52 se situa entre 7,38 e 9,21 (na linha dois da tabela), o valor- $p$  para 8,52 se encontra entre 0,025 e

0,010 (nos respectivos títulos de suas colunas).

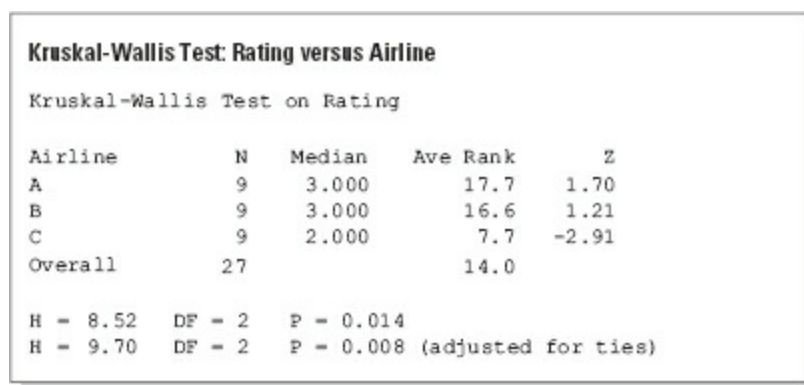
### 5. Ao examinar o valor- $p$ , veja se poderá ou não rejeitar a $H_0$ .

Você deve rejeitar a  $H_0$ : todas as populações têm a mesma localização, e favorecer a  $H_a$ : pelo menos duas populações têm localizações diferentes, se o valor- $p$  associado a KW for  $< \alpha$ , onde  $\alpha$  é 0,05 (ou seu nível  $\alpha$  predefinido). Caso contrário, você não pode rejeitar a  $H_0$ .

Seguindo o exemplo das companhias aéreas, já que o valor- $p$  está entre 0,010 e 0,025, ambos menores do que de  $\alpha = 0,05$ , você pode rejeitar a  $H_0$ . Conclui-se, então, que as classificações de pelo menos duas das três companhias aéreas são diferentes.

Para realizar o teste de Kruskal-Wallis usando o Minitab, insira seus dados em duas colunas, a primeira coluna deve representar os valores reais dos dados e a segunda coluna deve representar a população de origem dos dados (por exemplo, 1, 2, 3). Em seguida, clique em Stat>Nonparametrics>Kruskal-Wallis. No campo ao lado esquerdo, clique na coluna um e ela aparecerá no lado direito, como sua variável resposta (*response variable*). Em seguida, clique na coluna dois no campo à esquerda. Esta coluna vai aparecer ao lado direito como a variável fator (*factor variable*). Clique em OK e o teste KW estará pronto. Os principais resultados do teste KW são mostrados nas duas últimas linhas da saída do Minitab.

Os resultados da análise de dados do Minitab para os dados das companhias aéreas são mostrados na Figura 19-4. Na Figura 19-4, você pode ver que a estatística de teste KW para o exemplo das companhias aéreas é 8,52, que corresponde à encontrada manualmente (uau!). O valor- $p$  exato dado pelo Minitab é de 0,014.



Kruskal-Wallis Test: Rating versus Airline

Kruskal-Wallis Test on Rating

Airline	N	Median	Ave Rank	Z
A	9	3.000	17.7	1.70
B	9	3.000	16.6	1.21
C	9	2.000	7.7	-2.91
Overall	27		14.0	

H = 8.52	DF = 2	P = 0.014
H = 9.70	DF = 2	P = 0.008 (adjusted for ties)

**Figura 19-4:** Comparando as avaliações de três companhias aéreas usando o teste de Kruskal-Wallis.

No entanto, esse conjunto de dados tem alguns empates e as fórmulas devem se ajustar um pouco a esse fato (ajuste que sai do escopo deste livro). Levando em conta esses empates, o computador lhe dá KW = 9,70, com valor- $p$  de 0,008. Todas as evidências aqui pronunciam a mesma sentença, em alto e bom som, rejeite a  $H_0$ : as avaliações de todas as três companhias aéreas têm a mesma mediana (localização). Conclui-se, então, que as



classificações de pelo menos duas das três companhias aéreas são diferentes. (Mas quais são diferentes? Veja a resposta na próxima seção.)

# Localizando as Diferenças: O Teste da Soma dos Postos de Wilcoxon

Suponha que a  $H_0$  seja rejeitada no teste de Kruskal-Wallis, o que significa que você teve evidências suficientes para concluir que pelo menos duas das populações possuem medianas diferentes. Porém, você não sabe quais são. Quando se descobre que um conjunto de populações não compartilha a mesma mediana é muito provável que a próxima pergunta seja: “OK, então, quais são as diferentes?”. Para descobrir quais são as populações diferentes, depois que o teste de Kruskal-Wallis tenha rejeitado a  $H_0$ , você pode usar o *teste da soma de postos de Wilcoxon* (também conhecido como o *teste de Mann-Whitney*).



Você não pode sair procurando diferenças em pares específicos de populações até que tenha estabelecido que pelo menos duas populações sejam diferentes (ou seja, a  $H_0$  foi rejeitada no teste de Kruskal Wallis). Se você não fizer primeiro esta verificação, poderá se deparar com um monte de problemas, além de ter uma chance muito maior de tomar a decisão errada.

Nas seções seguintes, você verá como realizar comparações pareadas e interpretá-las a fim de descobrir onde estão as diferenças entre as  $k$  medianas populacionais que estiverem sendo estudadas.

## Comparações pareadas

O teste da soma de postos é um teste não paramétrico que compara duas localizações da população (por exemplo, as medianas). Quando você tiver mais do que duas populações, deverá conduzir o teste da soma de postos para cada par de populações a fim de verificar a existência de diferenças. Este procedimento é chamado de *comparações pareadas* (pairwise) ou *comparações múltiplas*. (Veja o Capítulo 10 para obter informações sobre a versão paramétrica das comparações múltiplas.) Por exemplo, uma vez que estamos comparando três companhias aéreas no exemplo de satisfação do cliente (veja a Tabela 19-1), temos que executar o teste da soma de postos três vezes para comparar as companhias aéreas A e B, A e C, B e C. Portanto, precisamos de três comparações pareadas para descobrir quais são as populações diferentes.



Para determinar quantos pares de comparações são precisos quando se tem  $k$  populações, use a fórmula  $\frac{k(k-1)}{2}$ . Você terá  $k$  populações para escolher e, depois,  $k-1$  populações para serem comparadas a elas. Por último, não se importe com a ordem entre as populações (contanto que não as perca de vista); sendo assim, divida por dois, pois há duas maneiras de ordenar qualquer par (por exemplo, a comparação de A e B lhe dá os mesmos resultados que a comparação entre B e A). No exemplo das companhias aéreas, temos  $k = 3$  populações, então, devemos ter  $\frac{k(k-1)}{2}$  pares de populações para comparar, o que corresponde ao que havia sido determinado anteriormente. (Para mais informações e

exemplos sobre como contar o número de maneiras de escolher ou ordenar um grupo de itens usando permutações e combinações, consulte meu outro livro publicado, *Probability For Dummies*, da Wiley).

## ***Realizando testes de comparação para ver quem é diferente***

O teste da soma de postos de Wilcoxon avalia a  $H_0$ : as duas populações têm a mesma localização versus a  $H_a$ : as duas populações têm diferentes localizações. Aqui, estão os passos de como usar o teste da soma de postos de Wilcoxon para fazer comparações:

- 1. Verifique as condições para o teste usando estatísticas descritivas e histogramas para as duas últimas e procedimentos de amostragem adequados para a primeira:**
  - As duas amostras devem provir de populações independentes.
  - As populações devem ter a mesma distribuição (forma).
  - As populações devem ter a mesma variância.
- 2. Estabeleça sua  $H_0$ : as duas medianas são iguais versus a  $H_a$ : as duas medianas são diferentes.**
- 3. Combine todos os dados e ordene os valores do menor para o maior.**
- 4. Some todos os postos da primeira amostra (ou da amostra menor, caso o tamanho das amostras não sejam iguais).**

Esse resultado será sua estatística de teste,  $T$ .

- 5. Compare  $T$  aos valores críticos da Tabela A-4 no apêndice, na linha e coluna correspondentes aos dois tamanhos amostrais (denotados por  $VCI$  e  $VCS$ ).**

Se  $T$  for igual ou maior do que os valores críticos (inferior ou igual ao valor inferior [ $VCI$ ] ou maior ou igual o superior [ $VCS$ ]), rejeite a  $H_0$  e conclua que as duas medianas populacionais são diferentes. Caso contrário, você não pode rejeitar a  $H_0$ .

- 6. Repita os passos de um a cinco para cada par de amostras do conjunto de dados e tire suas conclusões.**

Examine todos os resultados para ver quais pares populacionais têm a mesma mediana e quais não.

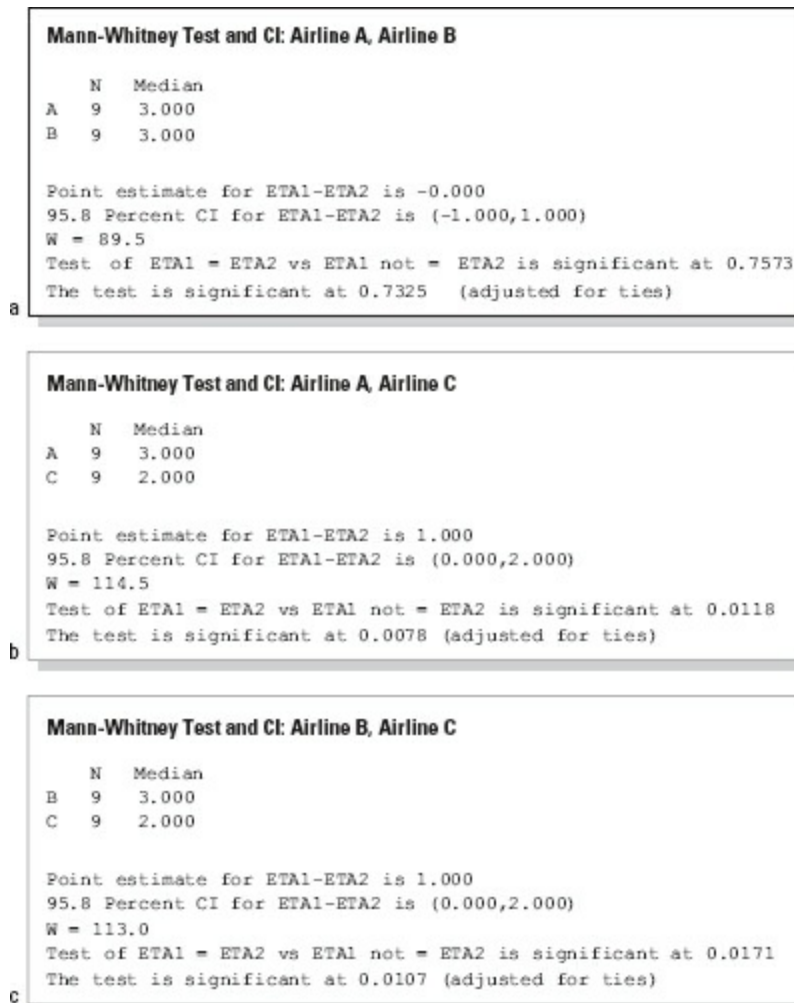
Para ver como realizar o teste da soma de postos de Wilcoxon para comparações pareadas, usando o Minitab, consulte o Capítulo 18. Note que o Minitab chama este teste por seu outro nome, teste de Mann-Whitney.

Veja a saída do Minitab para os três testes da soma de postos de Wilcoxon comparando as companhias A e B, A e C, B e C nas Figuras 19-5a, 19-5b e 19-5c, respectivamente.

A Figura 19-5a compara as avaliações das companhias aéreas A e B. O valor- $p$  (ajustado para empates) é 0,7325, muito maior do que os 0,05 que você precisa para rejeitar  $H_0$ .



Portanto, não se pode concluir que as companhias aéreas A e B têm índices de satisfação com medianas diferentes. A Figura 19-5b mostra que o valor- $p$  para a comparação entre as companhias aéreas A e C é 0,0078. Como esse valor- $p$  é muito menor do que o nível  $\alpha$  de 0,05, ele é uma evidência muito convincente de que as avaliações das companhias aéreas A e C não têm as mesmas medianas. A Figura 19-5c também mostra um valor- $p$  pequeno (0,0107), dando indícios de que as companhias aéreas B e C têm avaliações significativamente diferentes.



**Figura 19-5:** Testes da soma de postos de Wilcoxon comparando os índices de satisfação do cliente de duas companhias aéreas ao mesmo tempo.

## *Examinando as medianas para ver como elas se diferem*

A rejeição da  $H_0$  em uma comparação múltipla significa que você concluiu que as duas populações examinadas possuem medianas diferentes. Há duas maneiras de prosseguir daqui para ver como as medianas se diferem em todas as comparações pareadas:

- ✓ Você pode examinar boxplots paralelos de todas as amostras e comparar suas medianas (localizadas na linha do meio de cada caixa).
- ✓ Você pode calcular a mediana de cada amostra e ver quais são mais altas e quais são mais baixas do que as das populações que tenham sido concluídas como

estatisticamente diferentes.

Na seção anterior, você viu que as comparações pareadas para os dados das companhias aéreas realizadas pelos testes da soma de postos de Wilcoxon concluíram que as avaliações das companhias aéreas A e B não se diferenciaram, porém, ambas se diferenciaram da companhia aérea C.

Mas você ainda pode dizer muito mais; é possível dizer como a companhia diferente se compara em relação às outras. Voltando à Figura 19-2, vemos que as medianas das companhias A e B são 3,0, enquanto a mediana da companhia C é apenas 2,0. Essa diferença significa que as companhias aéreas A e B foram avaliadas de forma semelhante, mas a companhia aérea C teve avaliações inferiores em relação a A e B. Os boxplots na Figura 19-1 confirmam esses resultados.

## Apontando Correlações com o Posto de Spearman

---

### *Neste Capítulo*

- ▶ Entendendo a correlação a partir de um ponto de vista não paramétrico
  - ▶ Descobrindo e interpretando a correlação de posto de Spearman
- 

**O**s analistas de dados normalmente pesquisam e tentam quantificar as relações entre duas variáveis,  $x$  e  $y$ . Dependendo do tipo de dados com o qual você está lidando, é preciso utilizar procedimentos diferentes para quantificar a relação entre elas.

Quando  $x$  e  $y$  são variáveis *quantitativas* (isto é, seus possíveis resultados são medidas ou contagens), o coeficiente de correlação (também conhecido como *coeficiente de correlação de Pearson*) mede a força e o sentido de sua relação linear. (Consulte o Capítulo 4 para obter todas as informações sobre o coeficiente de correlação de Pearson, representado pela letra  $r$ .) Se  $x$  e  $y$  forem duas variáveis *categóricas* (ou seja, seus possíveis resultados são categorias que não têm qualquer significado numérico, tais como masculino e feminino), você pode usar procedimentos do Qui-quadrado e probabilidades condicionais para procurar e descrever as relações entre elas. (Descrevo todos esses processos nos Capítulos 13 e 14.)

Ainda há um terceiro tipo de variável, chamado de variável *ordinal*. Seus valores estão divididos em categorias, mas os possíveis valores podem ser ordenados e receberem um valor numérico que tenha significado, por exemplo, as notas em uma escala de  $A = 4$ ;  $B = 3$ ,  $C = 2$ ,  $D = 1$  e  $E = 0$  ou a avaliação de um aluno em relação ao professor em uma escala de melhor = 5 a pior = 1. Para procurar uma relação entre duas variáveis ordinais, como essas, os estatísticos usam a *correlação de posto de Spearman*, a contrapartida não paramétrica do coeficiente de correlação de Pearson (abordado no Capítulo 4). Neste capítulo, você vai ver por que as variáveis ordinais não atendem às condições de Pearson, além de descobrir como usar e interpretar a correlação de posto de Spearman para quantificar e interpretar corretamente as relações envolvendo variáveis ordinais.

# Pearson e Suas Preciosas Condições

O coeficiente de correlação de Pearson é a medida de correlação mais comum, e muitos analistas de dados acham que é a *única* existente. O problema é que a correlação de Pearson exige que certas condições sejam cumpridas antes de usá-la. Caso tais condições não sejam atendidas, a correlação de Spearman estará esperando-o de braços abertos.



O coeficiente de correlação de Pearson,  $r$ , (a correlação) é um número que mede a direção e a força da relação linear entre duas variáveis,  $x$  e  $y$ . (Para mais informações sobre correlação, consulte o Capítulo 4.)

Várias são as condições que devem ser cumpridas para o uso do coeficiente de Pearson:

- ✓ **As variáveis  $x$  e  $y$  devem ser numéricas (ou quantitativas).** Elas devem representar medições sem nenhuma restrição em seu nível de precisão. Por exemplo, os números com muitas casas decimais (por exemplo, 12,322 ou 0,219) devem ser preservados (e não arredondados).
- ✓ **As variáveis  $x$  e  $y$  devem ter uma relação linear (como mostrado em um diagrama de dispersão; consulte o Capítulo 4).**
- ✓ **Os valores de  $y$  devem ter uma distribuição normal para cada  $x$ , com a mesma variância em cada  $x$ .**



Normalmente, com variáveis ordinais, você não vai ver muitas categorias diferentes sendo oferecidas ou comparadas por razões de simplicidade. Isto significa que não haverá valores numéricos suficientes para tentar construir um modelo de regressão linear envolvendo variáveis ordinais, como é possível fazer com duas variáveis quantitativas.

Em outra situação, se você tiver uma variável de gênero com categorias masculino e feminino, poderá atribuir os números 1 e 2 para cada sexo, mas esses números não terão qualquer significado numérico. O sexo não é uma variável ordinal, mas, sim, uma *variável categórica* (uma variável que apenas separa os indivíduos em categorias). As variáveis categóricas também não se prestam a relações lineares e, portanto, também não atendem às condições de Pearson. (Para explorar as relações entre as variáveis categóricas, consulte o Capítulo 14.)

## Quem são esses caras? As pessoas por trás da estatística

Algumas pessoas têm o privilégio de que uma estatística receba seu nome. Normalmente, esse é o caso da pessoa que inventou a estatística, que, por reconhecer uma necessidade e apresentar uma solução, fica com a homenagem. Se a nova estatística for escolhida e utilizada por outras pessoas, ela acaba sendo chamada pelo nome de seu inventor.

A correlação de posto de Spearman recebe o nome de seu criador, Charles Edward Spearman (1863-1945). Ele foi um psicólogo inglês que estudou psicologia experimental, trabalhou na área da inteligência humana e foi professor por muitos anos na London University. Spearman acompanhou de perto o trabalho de Francis Galton, que, originalmente, desenvolveu o conceito de correlação. Spearman desenvolveu a correlação de posto em 1904.

O coeficiente de correlação de Pearson foi desenvolvido vários anos antes, em 1893, por Karl Pearson, um dos colegas de Spearman na London University e outro seguidor de Galton. Pearson e Spearman não se davam bem; Pearson tinha uma personalidade especialmente forte e volátil e, na verdade, teve problemas de convivência com várias pessoas.



# Correlação de Posto de Spearman

A correlação de posto de Spearman não exige que a relação entre as variáveis  $x$  e  $y$  seja linear, nem que as variáveis sejam numéricas. Em vez de examinar uma relação linear entre  $x$  e  $y$ , o teste de correlação de posto de Spearman verifica se duas variáveis ordinais e/ou quantitativas são independentes (isto é, não se relacionam entre si). **Observação:** o posto de Spearman se aplica apenas aos dados ordinais. Para verificar se duas variáveis categóricas (e não ordinais) são independentes, use o teste do Quiquadrado; consulte o Capítulo 14.



A correlação de posto de Spearman é igual à correlação de Pearson, exceto pelo fato de que é calculada com base nos *postos* das variáveis  $x$  e  $y$  (ou seja, o posto onde ocupam na ordem, veja o Capítulo 16) em vez de seus valores reais. Interprete o valor da correlação de Spearman,  $r_s$ , da mesma forma que interpretar a correlação de Pearson,  $r$  (veja o Capítulo 4). Os valores de  $r_s$  podem ir de  $-1$  a  $+1$ . Quanto maior a grandeza de  $r_s$  (nas direções positivas ou negativas), mais forte a relação entre  $x$  e  $y$ . Se  $r_s$  for zero,  $x$  e  $y$  são independentes. E, como o que ocorre em  $r$ , se a correlação entre  $x$  e  $y$  não for zero, não se pode dizer que as variáveis são ou não independentes.

Nesta seção, você vai ver como calcular e interpretar a correlação de posto de Spearman e como aplicá-la a um exemplo.

## Calculando a correlação de posto de Spearman

A notação para a correlação de Spearman é  $r_s$ , onde  $s$  significa Spearman. Para encontrar a  $r_s$ , realize os passos descritos nesta seção. Do segundo ao sexto passos, o Minitab faz o trabalho para você, embora alguns professores possam lhe pedir para fazer o trabalho à mão (não eu, é claro).

1. **Colete os dados na forma de pares de valores  $x$  e  $y$ .**
2. **Ordene os dados da variável  $x$ , onde 1 = menor até  $n$  = maior, onde  $n$  é o número de pares de dados do conjunto de dados.**

Este passo vai formar um novo conjunto de dados para a variável  $x$  chamada de postos dos valores de  $x$ . Se algum dos valores aparece mais de uma vez, o Minitab atribui a cada valor empatado a média dos postos que eles ocupariam se não estivessem empatados.

3. **Complete o segundo passo com os dados da variável  $y$ .**

Este passo vai formar um novo conjunto de dados chamado de *posto dos valores de  $y$* .

4. **Encontre o desvio padrão dos postos dos valores de  $x$  usando a fórmula habitual**

para o desvio padrão,  $\sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$ , aqui denotado por  $s_{xx}$ . De maneira similar,

encontre o desvio padrão dos postos dos valores de  $y$  usando  $\sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$ , aqui representado por  $s_{yy}$ .

Observe que  $n$  é o tamanho da amostra,  $\bar{x}$  é a média dos postos dos valores de  $x$ , e  $\bar{y}$  é a média dos postos dos valores de  $y$ .

**5. Encontre a covariância dos valores  $x$ - $y$ , utilizando a fórmula**

$$\text{Cov}(x, y) = \frac{\sum \sum (x - \bar{x})(y - \bar{y})}{n-1}, \text{ aqui representada por } s_{xy}.$$

A covariância de  $x$  e  $y$  é uma medida do desvio total dos valores  $x$  e  $y$  a partir do ponto  $(\bar{x}, \bar{y})$ .

**6. Calcule o valor da correlação de posto de Spearman, utilizando a fórmula  $r_s = \frac{s_{xy}}{s_{xx}s_{yy}}$ .**

A fórmula para a correlação de Spearman é exatamente igual à fórmula do coeficiente de correlação de Pearson, exceto pelo fato de que os dados que Spearman usa são os postos de  $x$  e  $y$  em vez dos valores originais de  $x$  e  $y$  usados por Pearson. Assim, Spearman apenas se preocupa com a ordem dos valores de  $x$  e  $y$ , e não com seus valores reais.

Para calcular a correlação de Spearman usando o Minitab, ordene os valores de  $x$  e os de  $y$  e, então, encontre a correlação de seus postos. Ou seja, vá ao menu Data>Rank e clique na variável  $x$  para obter os postos de  $x$ . Em seguida, faça a mesma coisa para obter os postos de  $y$ . Vá ao menu Stat>Basic Statistics>Correlation, clique nas duas colunas que representam os postos e clique em OK.

## ***Spearman em ação: Relacionando aptidão ao desempenho***

Saber calcular a correlação de Spearman é uma grande coisa, mas, se você conseguir aplicá-la às situações do mundo real, vai realmente se destacar no mundo da Estatística (ou, pelo menos, em suas aulas de Estatística). Então, tente se colocar no cenário desta seção para obter o efeito completo da correlação de Spearman.

Você é um professor de Estatística que dá provas de vez em quando (um trabalho sujo, mas alguém tem de fazê-lo). Depois de examinar as notas finais dos alunos ao longo dos anos (sim, você é um professor mais velho ou pelo menos por volta de seus 45 anos), você percebe que os alunos que se dão bem em suas aulas tendem a ter mais aptidão para Matemática e Estatística. Para então verificar esta teoria, você dá um teste de aptidão aos alunos dessas matérias no primeiro dia de aula, pois pretende comparar as notas dos alunos no teste de aptidão com suas notas finais no fim do curso.

Agora, os detalhes. Suas variáveis são  $x$  = nota no teste de aptidão (usando um pré-teste de 100 pontos no primeiro dia de aula) e  $y$  = nota final numa escala de 1 a 5, onde 1 = F (não aprovado), 2 = D (aprovado), 3 = C (na média), 4 = B (acima da média) e 5 = A (excelente). A variável  $y$ , a nota final, é uma variável ordinal e a variável  $x$ , a aptidão, é uma variável numérica. O que se deseja saber é se há ou não uma relação entre  $x$  e  $y$ . Para



isso, coletam-se dados usando uma amostra aleatória de 20 alunos; veja-os na Tabela 20-1. Este é o primeiro passo no processo de cálculo da correlação de Spearman. (Veja os passos descritos na seção anterior.)

<b>Tabela 20-1</b>		
<b>Nota do Teste de Aptidão e Nota Final em Estatística</b>		
<i>Aluno</i>	<i>Aptidão</i>	<i>Nota Final</i>
1	59	3
2	47	2
3	58	4
4	66	3
5	77	2
6	57	4
7	62	3
8	68	3
9	69	5
10	36	1
11	48	3
12	65	3
13	51	2
14	61	3
15	40	3
16	67	4
17	60	2
18	56	3
19	76	3
20	71	5

Usando o Minitab, a correlação de posto de Spearman é igual a 0,379. A discussão a seguir percorre do segundo ao sexto passos, caso você encontre essa correlação à mão, muito provável no caso de um exame.

O segundo e terceiro passos para se encontrar a correlação de posto de Spearman pedem que você ordene os resultados dos testes de aptidão ( $x$ ) do menor (1) ao maior; depois, ordene as notas finais ( $y$ ) da menor (1) para a maior. Observe que as notas do exame final possuem diversos empates, portanto, use a média dos postos. Por exemplo, na coluna três da Tabela 20-1, vemos um único 1 para o aluno 10, que vai ocupar o posto 1. Porém, vemos quatro 2 para os alunos 2, 5, 13 e 17. Esses valores, se não estivessem empatados, teriam ocupado os postos 2, 3, 4 e 5. A média desses quatro postos é  $\frac{2+3+4+5}{4} = \frac{14}{4} = 3,5$ . Cada um dos 2 na coluna três, portanto, ocupará o posto 3, 5.

A Tabela 20-2 mostra os dados originais, os postos para as notas de aptidão ( $x$ ) e os postos para as notas finais ( $y$ ), calculados pelo Minitab.

<b>Tabela 20-2</b>		<b>Nota do Teste de Aptidão, Nota final e Postos</b>		
<i>Aluno</i>	<i>Aptidão</i>	<i>Posto da Aptidão</i>	<i>Nota Final</i>	<i>Posto da Nota Final</i>
1	59	9	3	10,5
2	47	3	2	3,5
3	58	8	4	17,0
4	66	14	3	10,5
5	77	20	2	3,5
6	57	7	4	17,0
7	62	12	3	10,5
8	68	16	3	10,5
9	69	17	5	19,5
10	36	1	1	1,0
11	48	4	3	10,5
12	65	13	3	10,5
13	51	5	2	3,5
14	61	11	3	10,5
15	40	2	3	10,5
16	67	15	4	17,0
17	60	10	2	3,5
18	56	6	3	10,5
19	76	19	3	10,5
20	71	18	5	19,5

No quarto passo para o cálculo da correlação de Spearman, o Minitab calcula o desvio padrão dos postos para as notas no teste de aptidão (localizado na coluna dois da Tabela 20-2) e o desvio padrão das notas finais (localizado na coluna quatro da Tabela 20-2). No quinto passo, o Minitab calcula as covariâncias dos postos das notas no teste de aptidão e das notas finais. Estas estatísticas são mostradas na Figura 20-1.

Descriptive Statistics: Ranks of X, Ranks of Y	
Variable	StDev
Ranks of X	5.92
Ranks of Y	5.50

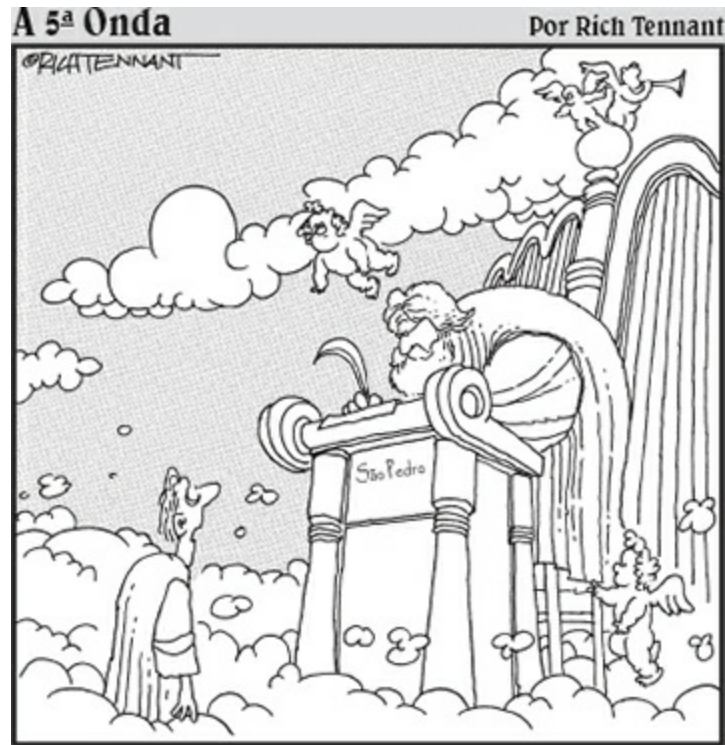
Covariances: Ranks of X, Ranks of Y		
	Ranks of X	Ranks of Y
Ranks of X	35.0000	
Ranks of Y	12.3421	30.2632

**Figura 20-1:** Desvios padrão e covariância dos postos das notas de aptidão (x) e das notas finais (y).

No sexto e último passo, para encontrar a correlação de posto de Spearman, calcule  $r_s$  dividindo a covariância dos postos de  $x$  e  $y$  (que se encontram no canto inferior esquerdo) pelo produto dos desvios padrão dos postos de  $x$  ( $s_{xx}$ ) e postos de  $y$  ( $s_{yy}$ ). E, assim, temos  $\frac{12,34}{5,92 * 5,50} = 0,379$ , que corresponde ao valor da correlação de Spearman encontrado pelo Minitab.

Esta correlação de posto de Spearman de 0,379 é bastante baixa, indicando uma fraca relação entre as notas no teste de aptidão realizado no início do curso e as notas finais do curso. Moral da história? Se um aluno não começar bem, ele ainda tem chances de melhorar, e se ele começar por cima, pode não sair da mesma forma que entrou. Embora haja ainda muito o que falar sobre o empenho de cada um durante o curso (comprar o *Estatística II Para Leigos* certamente só lhe fará bem!).

## **Parte VI:** **A Parte dos Dez**



**“Como assim eu não me enquadro em sua amostra populacional desejada?”**



## *Nesta parte...*

***E***sta parte é essencial nos livros da série *Para Leigos*, e por boas razões. Nesta parte, você vai obter informações úteis sobre as conclusões estatísticas com as quais deve ser cauteloso, além de ver as muitas maneiras com que a Estatística é usada no trabalho. Também vou lhe dizer por que você não deve tentar fugir das estatísticas, mas, sim, mergulhar nelas, uma vez que saber Estatística significa a segurança de seu emprego em praticamente qualquer área em que atuar.



# Capítulo 21

## Os Dez Erros Mais Comuns nas Conclusões Estatísticas

---

### *Neste Capítulo*

- ▶ Reconhecendo e evitando os erros ao interpretar os resultados estatísticos
  - ▶ Decidindo se deve ou não confiar nas conclusões dos outros
- 

**A** Estatística II trata da construção de modelos e da realização da análise de dados. Seu foco é a análise dos dados com a finalidade de descobrir o que está por trás deles. Trata-se de assegurar que a história seja contada de forma correta, justa e abrangente. Neste capítulo, discuto alguns dos erros mais comuns que já vi, como professora e consultora de Estatística, ao longo desses anos de trabalho. Você pode usar esta lista do que não deve ser feito na hora de fazer sua lição de casa e seus relatórios ou como uma forma rápida de revisão antes de uma prova ou exame. Acredite em mim — seu professor vai adorá-lo se você fizer isso!

## *Dizer o que as Estatísticas Provam...*

Não acredite em pessoas que usam os termos *essas estatísticas* e *provar* na mesma frase. A palavra *provar* expressa um tipo de conceito definitivo, final, encerrado, concluído, e a Estatística, por natureza, não é definitiva. Em vez disso, a Estatística lhe fornece evidências a favor ou contra a teoria, modelo ou argumento de alguém com base nos dados coletados e, portanto, deixa que você tire suas próprias conclusões. Uma vez que as evidências se baseiam em dados que variam de amostra para amostra, os resultados também podem variar — e esse é o desafio, a beleza e, às vezes, a frustração da Estatística. O melhor que você pode dizer é que a Estatística sugere, leva você a acreditar, ou lhe fornece elementos suficientes para concluir — mas nunca vá tão longe a ponto de dizer que a Estatística prova alguma coisa.

# *Tecnicamente Não É Estatisticamente Significativo, Mas...*



Depois de definir seu modelo e testá-lo com seus dados, você deve sustentar suas conclusões, não importando o quanto acredita que elas estejam erradas. A Estatística deve emprestar objetividade a todos os processos.

Suponha que Bárbara, uma pesquisadora, acabou de coletar e analisar uma penca de dados, mas ainda não conseguiu encontrar nada. No entanto, no fundo de seu coração, ela sabe que sua teoria é verdadeira, mesmo que os dados não a confirmem. Sua teoria é a de que os cães têm PES (percepção extrassensorial) — ou seja, “sexto sentido”. Ela baseia sua teoria no fato de que seu cão parece saber quando ela vai sair da casa, quando ela vai levá-lo ao veterinário e quando um banho é iminente, pois ele fica triste e procura um canto para se esconder.

Bárbara testa sua teoria estudando dez cães. Ela coloca um pedaço de comida para cachorro em uma de duas tigelas e pede a cada cão que encontre o alimento empurrando uma tigela. (Assumindo que a tigela seja espessa o suficiente para evitar que os cães trapaceiem usando o faro para encontrar a comida.) Ela repete esse processo dez vezes com cada cão e registra o número de respostas corretas. Se os cães não tiverem PES, o esperado é que eles escolham a tigela correta em 50% das vezes, pois cada cão tem duas tigelas para escolher, e cada uma delas possui chances iguais de serem selecionadas.

Como era esperado, os cães do estudo de Bárbara acertaram 55% das vezes. Agora, esse percentual é tecnicamente superior ao valor esperado de 50%, mas não é o suficiente (especialmente com tão poucos cães e tão poucas tentativas) para justificar a significância estatística. Ou seja, Bárbara não tem provas suficientes para a teoria da PES. Mas, quando Bárbara apresenta seus resultados na conferência de que está participando, ela dá um toque em seus resultados, dizendo: “Os cães acertaram em 55% das vezes, valor superior a 50%, portanto, embora estes resultados não sejam *tecnicamente* suficientes para serem estatisticamente significativos, acredito que demonstram alguma evidência de que os cães têm PES”(fazendo com que todos os estatísticos presentes queiram gritar “NÃO!”).

Alguns pesquisadores utilizam esse tipo de conclusão o tempo todo — deslizando ao redor da estatística quando esta não segue seu caminho. Porém, este é um jogo muito perigoso, pois a próxima vez que alguém tentar repetir os resultados de Bárbara (e acredite, sempre há quem tente), essa pessoa vai descobrir o que você sabia desde o começo (talvez seja PES?): Quando Bárbara começa a pegar suas coisas para sair de casa, seu cão percebe que algo vai acontecer e se esconde. E é isso o que realmente acontece.

# Concluir que $x$ Causa $y$

Sabe qual é a palavra que deixa um estatístico nervoso? As duas primeiras parecem bastante inofensivas e  $x$  e  $y$  são apenas letras do alfabeto, portanto, só sobra a palavra *causa*. De todas as palavras usadas de forma vaga na Estatística, *causa* lidera a lista.

Veja aqui um exemplo do que quero dizer. Para seu relatório final de Estatística, você estuda os fatores que se relacionam com a nota dos alunos no exame final. Para isso, você coleta dados de 500 alunos de Estatística, perguntando a cada um deles uma variedade de questões, tais como: “Qual foi sua nota bimestral?”, “Quantas horas você dormiu na noite anterior à prova final?” e “Qual é sua nota média?”. Você, então, faz uma análise de regressão linear múltipla (usando as técnicas do Capítulo 5) e conclui que o tempo de estudo e a quantidade de sono na noite anterior ao teste são os fatores mais importantes para determinar a nota no exame. Você escreve todas suas análises em um papel e, no final, diz: “Estes resultados demonstram que mais tempo de estudo e uma boa noite de sono na noite anterior à prova são a causa de uma boa nota no exame final”.

Estava com você até dizer a palavra *causa*. Não se pode dizer que dormir mais tempo ou estudar mais tempo cause o aumento na nota da prova. Os dados coletados mostram que as pessoas que dormiram bem e estudaram muito conseguiram boas notas, e aqueles que não fizeram isso não tiveram boas notas. Mas esse resultado não significa que basta dormir e estudar mais para tudo ficar bem. Isso é o mesmo que dizer que uma vez que o aumento na altura se relaciona ao aumento do peso, você ficará mais alto se ganhar peso.

O problema é que você não fez com que a mesma pessoa mudasse seu tempo de sono e hábitos de estudo para ver o que aconteceria com seu desempenho no exame (usando dois exames diferentes com mesmo grau de dificuldade). Esse estudo requer um *experimento planejado*. Quando você realiza uma *pesquisa*, não tem como controlar os outros fatores relacionados, o que pode complicar um pouco as coisas, como a qualidade do estudo, a frequência nas aulas, as notas nos trabalhos de casa, e assim por diante.

A única maneira de controlar os outros fatores é fazendo um experimento casualizado (usando um grupo de tratamento, um grupo controle e os controles para outros fatores que poderiam influenciar o resultado). Afirmar causalidade sem realizar um experimento casualizado é um erro muito comum cometido por alguns pesquisadores na hora de tirar conclusões.



# *Supor que os Dados São Normais*

A palavra-chave aqui é *supor*. Para defini-la de forma simples, uma suposição é algo em que você acredita sem verificar. As suposições podem levar a análises erradas e a resultados incorretos — tudo sem que a pessoa suponha, ainda que conheça o perigo.

Muitas análises possuem certos requisitos. Por exemplo, os dados devem vir de uma distribuição normal (a clássica distribuição com forma de sino). Se alguém disser: “Supus que os dados eram normais”, isso quer dizer que ele apenas supôs que os dados vieram de uma distribuição normal. Mas será que basta presumir que uma distribuição é normal e seguir em frente, ou será que é preciso fazer algo mais? Tenho certeza de que você acertou — é preciso fazer algo mais.

Por exemplo, para realizar um teste-*t* para uma amostra (veja o Capítulo 3), os dados devem ser provenientes de uma distribuição normal a menos que o tamanho da amostra seja grande, caso em que você tem uma distribuição aproximadamente normal segundo o Teorema do Limite Central. (Você se lembra destas três palavras em Estatística I?) Aqui, não estamos fazendo uma suposição, mas examinando uma *condição* (algo que você verifica antes de prosseguir). Para isso, deverá traçar o gráfico dos dados, ver se eles atendem à condição e, caso atendam, devemos continuar. Caso os dados não atendam às condições, podemos usar os métodos não paramétricos (discutidos no Capítulo 16).

Quase todas as técnicas estatísticas para análise de dados têm algumas condições que devem ser atendidas antes de usá-las. Sempre descubra quais são essas condições e verifique se seus dados as satisfazem (e, se não, considere o uso de estatísticas não paramétricas, veja o Capítulo 16). Esteja ciente de que muitos livros didáticos de estatística utilizam a palavra *suposição* de forma incorreta, quando, na verdade, querem se referir à *condição*. É uma diferença sutil, mas muito importante.



# Relatar Apenas os Resultados “Importantes”



Como um analista de dados, você não deve apenas evitar a armadilha de relatar apenas os resultados significativos, emocionantes e importantes, mas também ser capaz de detectar quando alguém está fazendo isso. Alguns trituradores de números examinam todas as opções possíveis e examinam os dados de todas as formas possíveis antes de se decidir pela análise que vai levá-los ao resultado desejado.

Provavelmente, você já tenha identificado o problema dessa abordagem. Cada técnica carrega a probabilidade para o erro. Se você estiver fazendo um teste- $t$ , por exemplo, e o nível  $\alpha$  for de 0,05, ao longo do tempo, 5 de cada 100 testes- $t$  resultarão em um alarme falso (o que significa que você vai declarar um resultado estatisticamente significativo quando na verdade ele não o é) apenas por acaso. Assim, se um ávido pesquisador realizar 20 testes de hipóteses sobre o mesmo conjunto de dados, há a probabilidade de que, pelo menos um, em média, desses testes possa resultar em um alarme falso em virtude do acaso. À medida que este pesquisador realiza mais e mais testes, ele está injustamente aumentando suas chances de encontrar algo apenas por acaso e correndo o risco de chegar a uma conclusão errada durante esse processo.

No entanto, nem tudo é culpa do pesquisador. Geralmente, ele é pressionado por um sistema direcionado a resultados. É triste saber que os únicos resultados que viram notícia e aparecem em artigos de revistas científicas são os que apresentam um resultado estatisticamente significativo (quando a  $H_0$  é rejeitada). Talvez, tenha sido um erro os estatísticos terem adotado o termo *significância* para designar a rejeição da  $H_0$  — como se quisessem dizer que a rejeição de  $H_0$  é a única conclusão importante a que se pode chegar. E todas as vezes que a  $H_0$  não pode ser rejeitada? Por exemplo, quando os médicos não conseguem concluir que beber refrigerante *diet* engorda ou quando as pesquisas de satisfação não demonstram que as pessoas estão descontentes com o presidente? O público seria mais bem servido se pesquisadores e meios de comunicação fossem incentivados a comunicar os resultados estatisticamente insignificantes, mas ainda importantes, com os estatisticamente significativos.



A regra é a seguinte: a fim de descobrir se uma conclusão estatística é correta, não basta examinar apenas a análise que o pesquisador está mostrando. Você também deve conhecer as análises e os resultados *não* mostrados e questionar. Evite a pressa em rejeitar a  $H_0$ .

# Supor que uma Amostra Grande É Sempre Melhor

O ditado que diz quanto maior, melhor, serve para algumas coisas, mas nem sempre para os tamanhos amostrais. Por um lado, quanto maior a amostra, mais precisos são os resultados (se os dados não estiverem enviesados). Uma amostra grande também aumenta a capacidade de sua análise em detectar diferenças de um modelo ou negar uma afirmação sobre uma população (em outras palavras, em rejeitar a  $H_0$  quando devido). Essa capacidade de detectar diferenças verdadeiras a partir da  $H_0$  é chamada de *poder* de um teste (veja o Capítulo 3). No entanto, alguns pesquisadores podem (e frequentemente o fazem) expandir demais a ideia de poder. Eles aumentam o tamanho amostral até o ponto em que mesmo a diferença mais ínfima a partir da  $H_0$  faz com que eles gritem aos quatro ventos a rejeição da  $H_0$ .



Os tamanhos amostrais devem ser grandes o suficiente para fornecer precisão e repetibilidade dos resultados, mas há um problema, com o fato de ser grande demais, acredite ou não. Você pode sempre coletar tamanhos amostrais grandes o suficiente para rejeitar qualquer hipótese nula, mesmo quando o desvio real a partir dela seja embaraçosamente pequeno. Mas o que pode ser feito? Quando você ler ou ouvir que o resultado foi considerado estatisticamente significativo, pergunte qual foi a média da amostra (antes era colocada na fórmula  $t$ ) e avalie sua significância para você a partir de um ponto de vista prático. Cuidado com quem diz: “Estes resultados são estatisticamente significativos e o grande tamanho amostral (100.000) fornece evidências ainda mais fortes para isso.”

Suponha que uma pesquisa afirme que um típico cão doméstico assista a uma média de dez horas de TV por semana. Bob acredita que a verdadeira média é maior, com base no fato de que seu cão, Fido, assiste sozinho a, pelo menos, dez horas de programas de culinária toda semana. Bob, então, estabelece o seguinte teste de hipótese:  $H_0: \mu = 10$  versus  $H_a: \mu > 10$ . Ele coleta uma amostra aleatória de 100 cães e pede que seus donos registrem o quanto de TV os cães assistiram por semana. O resultado revela uma média amostral de 10,1 horas, e o desvio padrão da amostra é de 0,8 horas. Este resultado não é o que Bob esperava, pois 10,1 está muito próximo de 10. Ele calcula a estatística de teste usando a

fórmula  $t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$  e obtém um valor de  $t = \frac{(10,1 - 10,0)}{\frac{0,8}{\sqrt{100}}} = \frac{0,1}{0,08}$ , que equivale a 1,25 para  $t$ .

Como o teste é unicaudal à direita ( $>$  na  $H_a$ ), Bob pode rejeitar a  $H_0$  em  $\alpha$  se  $t$  for maior do que 1,645, e seu valor  $t$  de 1,25 estiver muito aquém desse valor. Observe que, como  $n = 100$  aqui, você encontra o valor de 1,645 na última linha da tabela de distribuição- $t$  (consulte a Tabela A-1 no apêndice). A linha está marcada com o símbolo do infinito para indicar uma amostra grande. Assim, Bob não pode rejeitar a  $H_0$ . Para colocar mais lenha na fogueira, o amigo de Bob, Joe, conduz o mesmo estudo e obtém a mesma média amostral e o mesmo desvio padrão de Bob, mas usa uma amostra aleatória de 500 cães,

em vez de 100. Consequentemente, o valor- $t$  de Joe é  $t = \frac{(10,1 - 10,0)}{\frac{0,8}{\sqrt{500}}} = \frac{0,1}{0,036}$ , que equivale a

2,78. Como 2,78 é maior do que 1,645, Joe consegue rejeitar a  $H_0$  (para desespero de Bob).



Por que o teste de Joe encontrou um resultado diferente do de Bob? A única diferença foi o tamanho da amostra. A amostra de Joe era maior e, quanto maior o tamanho amostral, menor o erro padrão (veja o Capítulo 3). Como o erro padrão fica no denominador da fórmula de  $t$ , quanto menor for seu valor, maior será o valor- $t$ . E, conseqüentemente, quanto maior for o valor- $t$ , mais fácil fica rejeitar a  $H_0$ . (Veja o Capítulo 3 para mais informações sobre precisão e margem de erro).

Agora, Joe poderia, tecnicamente, dar uma grande entrevista à imprensa ou escrever um artigo sobre seus resultados (sua mãe ficaria muito orgulhosa), mas você sabe a verdade. Você sabe que os resultados de Joe são tecnicamente *estatisticamente* significativos, mas não *praticamente significativos* — eles não significam nada para nenhuma pessoa ou cão. Afinal, quem se importa que ele foi capaz de mostrar evidências de que os cães assistem a um pouquinho mais do que dez horas de TV por semana versus exatamente dez horas por semana? Esta notícia não é exatamente um furo.



# *Não É Tecnicamente Aleatória, Mas...*

Quando você for coletar uma amostra para construir resultados estatísticos, a palavra de ordem é *aleatória*. Você deve fazer com que a amostra seja selecionada aleatoriamente da população. O problema é que as pessoas costumam coletar amostras do que acham ser mais aleatória, mais ou menos aleatória, ou aleatória o suficiente. O plano para a coleta de uma amostra é aleatório ou não.

Um dia, dei a cada um dos 50 alunos em minha classe um número de 1 a 50 e sorteei dois números aleatoriamente de um chapéu. Os dois alunos sorteados estavam sentados na primeira fila e, não só isso, eles estavam um ao lado do outro. Meus alunos imediatamente gritaram que eu estava roubando!

Após este resultado aparentemente estranho, aproveitei a oportunidade para falar com minha classe sobre as amostras verdadeiramente aleatórias. Uma *amostra aleatória* é escolhida de tal forma, que cada membro da população original tenha a mesma chance de ser selecionado. Às vezes, as pessoas que se sentam uma ao lado da outra podem ser escolhidas. Na verdade, se estes resultados aparentemente estranhos nunca acontecerem, deveríamos nos preocupar com o processo; em um processo verdadeiramente aleatório, podemos obter resultados que possam parecer estranhos, esquisitos ou mesmo fixos. Isso faz parte do jogo.

Em minhas experiências de consultoria, sempre pergunto como meus clientes escolhem ou pretendem escolher suas amostras. Eles sempre respondem que farão de tudo para que ela seja aleatória. Mas, quando lhes pergunto como vão fazer isso, por vezes, recebo respostas não satisfatórias. Por exemplo, uma pessoa precisava obter uma amostra aleatória de uma população de 500 galinhas caipiras em uma fazenda. Ela precisava de cinco galinhas e disse que as tinha selecionado aleatoriamente, escolhendo as cinco que primeiro se aproximaram dela. O problema é que os animais que se aproximam de você podem ser mais amigáveis, mais dóceis, mais velhos ou talvez mais mansos. Essas características não estão presentes em todas as galinhas da fazenda e, portanto, essa escolha não foi aleatória. Provavelmente, os resultados estarão enviesados neste caso.

Sempre pergunte ao pesquisador como ele ou ela selecionou a amostra e, quando você selecionar suas próprias amostras, permaneça fiel à definição de aleatório. Não use seu próprio julgamento para escolher uma amostra aleatória; use um computador para fazer isso por você!



# *Supor que 1.000 Respostas São 1.000 Respostas*

Um artigo de jornal sobre a última pesquisa diz que 50% dos entrevistados disseram blá-blá-blá. As letras miúdas dizem que os resultados são baseados em uma pesquisa com 1.000 adultos nos Estados Unidos. Mas, espere: 1.000 é o número real de pessoas selecionadas para a amostra ou é o número final de participantes? Talvez seja preciso dar uma segunda olhada, pois esses dois números quase nunca coincidem.

Por exemplo, Jenny quer saber a porcentagem de americanos que já sonegou impostos. Na aula de Estatística, ela descobriu que se conseguir uma amostra de 1.000 pessoas, a margem de erro para sua pesquisa é de apenas  $\pm 3\%$ , o que ela acredita ser excelente. Sendo assim, ela sai em busca de sua meta de 1.000 respostas à sua pesquisa. Ela sabe que atualmente é difícil conseguir que as pessoas respondam a uma pesquisa e, portanto, está preocupada em perder uma grande quantidade de sua amostra por causa disso. Então ela tem uma ideia: por que não distribuir mais questionários do que o necessário para que, assim, possa receber 1.000 questionários respondidos?

Jenny analisa os resultados de pesquisa em vários jornais, várias revistas e na Internet, e descobre que a taxa de resposta (a porcentagem de pessoas que realmente responderam à pesquisa) é, normalmente, cerca de 25%. (Em se tratando do mundo real, estou sendo generosa com esse número, acredite ou não. Mas pense nisso: quantas pesquisas você já jogou fora recentemente? Não se preocupe, também me sinto culpada por fazer isso). Jenny então faz os cálculos e descobre que se enviar 4.000 questionários e obtiver 25% deles de volta, ela terá os 1000 questionários de que precisa para fazer sua análise, responder a sua pergunta e conseguir aquela pequena margem de erro de  $\pm 3\%$ .

Jenny realiza sua pesquisa e, bem como o esperado, dos 4.000 questionários enviados, 1.000 voltaram. Ela, então, prossegue com sua análise e conclui que 400 dessas pessoas relataram ter sonegado seus impostos (40%). Depois, acrescenta a margem de erro e relata: “Com base nos dados de minha pesquisa, 40% dos norteamericanos sonegam seus impostos,  $\pm 3$  pontos percentuais.”

Agora, espere um pouco, Jenny. Ela só sabe o que essas 1.000 pessoas que responderam à pesquisa disseram e não tem ideia do que as outras 3.000 pessoas diriam. E aqui está o problema: o fato de alguém responder ou não a uma pesquisa está muitas vezes relacionado com o motivo pelo qual a pesquisa está sendo realizada. Não é algo aleatório. Os não respondentes (pessoas que não respondem a uma pesquisa) carregam um grande peso com relação ao que não estão tendo tempo para lhe contar.

Em nome da argumentação, vamos supor que 2.000 pessoas que inicialmente receberam a pesquisa se sentiram desconfortáveis com a pergunta, pois realmente sonegam seus impostos. E elas, simplesmente por não quererem que ninguém soubesse disso, jogaram a pesquisa no lixo. Suponha que as outras 1.000 pessoas não soneguem impostos e, portanto, não entenderam a pesquisa como algo importante e, por isso, não a responderam. Se esses dois cenários fossem verdadeiros, o resultado seria mais ou menos isto:

$$\text{Sonegadores} = 400 \text{ (respondentes)} + 2.000 \text{ (respondentes)} = 2400.$$

Tal resultado aumenta a porcentagem total de sonegadores para 2.400. dividido por 4.000 = 60%. Essa é uma diferença enorme!

Mas também é possível mudar totalmente o resultado se usássemos os outros 3.000 respondentes de outra forma. Poderíamos supor que nenhum deles sonega impostos, mas simplesmente não têm tempo para responder à pesquisa. Se soubesse dessa informação, você teria 600 (respondentes) + 3000 (respondentes) = 3.600 não sonegadores. De 4.000 entrevistados, esse número representaria 90% de não sonegadores e apenas 10% de sonegadores. É provável que a verdade esteja em algum lugar entre os dois exemplos dados, mas os não respondentes fazem com que fique muito difícil dizer onde exatamente.

E o pior é que a fórmula que Jenny usou para a margem de erro não sabe que a informação dada se baseia em dados enviesados e, assim, a margem de erro por ela relatada, de 3%, está errada. As fórmulas calculam os resultados independentemente de qualquer coisa, cabe a você se certificar de que esteja usando informações de qualidade.



Obter 1.000 resultados a partir do envio de 4.000 questionários não é tão bom quanto obter 1000 resultados de 1.000 questionários enviados (ou até mesmo 100 resultados de 100 enviados). Planeje sua pesquisa com base no acompanhamento que você pode fazer com as pessoas para conseguir concluir seu trabalho, e, se isso requer um tamanho amostral menor, que assim seja. Pelo menos, as chances para que os resultados estejam estatisticamente no alvo são melhores.

# *Naturalmente, os Resultados se Aplicam à População em Geral*

Tirar conclusões a respeito de uma população muito mais ampla do que a representada por sua amostra é um dos maiores erros em Estatística. Esse tipo de problema é chamado de *generalização* e ocorre mais frequentemente do que se pensa. As pessoas querem resultados instantâneos, pois não querem esperar por eles. Assim, pesquisas bem planejadas e os experimentos representam um contratempo para as pesquisas instantâneas da Internet e amostras convenientes.

Por exemplo, um pesquisador deseja saber como os canais de notícias da televisão a cabo têm influenciado a forma como os americanos recebem as notícias. Por acaso, ele também é professor de Estatística em uma grande instituição de pesquisa e tem 1.000 alunos em sua classe. Assim, ele decide que em vez de coletar uma amostra aleatória de americanos, o que seria muito mais difícil, demorado e caro, ele vai colocar uma pergunta em seu exame final para obter as respostas de seus alunos. Sua análise dos dados lhe mostra que apenas 5% de seus alunos leem o jornal e/ou assistem a noticiários na TV aberta, todos os demais assistem aos noticiários das TVs a cabo. Na classe desse professor, a proporção de alunos que exclusivamente assistem aos noticiários da TV a cabo em relação aos alunos que não assistem é de 20 para 1. O professor, então, relata isso e escreve um *press release* sobre o assunto. No dia seguinte, os canais de notícias a cabo se aproveitam dele e informam: “Os americanos preferem os canais de notícias a cabo aos jornais e canais de notícias da TV aberta em uma margem de 20 para 1!”.

Você consegue ver o que há de errado nesta situação? As conclusões do professor vão muito além de seu estudo, que está errado. Ele usou os alunos de Estatística para obter os dados que serviram como base para seu relatório e a manchete resultante. No entanto, o professor ainda relata que esses resultados são verdadeiros para todos os americanos. Acho que é seguro dizer que uma amostra de 1.000 estudantes universitários, todos tirados de uma turma de Estatística da mesma universidade, não representa os Estados Unidos.

Se a finalidade do professor é tirar conclusões sobre os Estados Unidos, ele tem que selecionar uma amostra aleatória de americanos para sua pesquisa. Se ele usar 1.000 estudantes de sua classe, suas conclusões só podem ser estendidas a sua classe e a mais ninguém.

Para evitar ou detectar a generalização, identifique a população sobre a qual se quer tirar conclusões e se certifique de que a amostra selecionada a represente de fato. Caso a amostra represente um grupo menor dentro dessa população, você também deve reduzir o alcance de suas conclusões.



Por vezes, parece mais fácil omitir algumas informações. Vejo isso com muita frequência quando leio artigos e reportagens baseadas em estatísticas. Mas esse erro não é culpa de uma única pessoa ou grupo. Os culpados podem incluir:

- ✓ **Os produtores:** Alguns pesquisadores podem omitir detalhes estatísticos em seus relatórios por uma variedade de razões, incluindo limitações de tempo e espaço. Afinal, não se pode escrever sobre todos os elementos do experimento do começo ao fim. No entanto, outros itens omitidos podem ser indicativos de um problema maior. Por exemplo, os relatórios costumam dizer muito pouco sobre como os dados foram coletados ou como a amostra foi selecionada. Ou podem discutir os resultados de uma pesquisa, mas não mostrarem as perguntas que foram feitas. Dez de 100 pessoas podem ter abandonado o experimento, mas os pesquisadores não lhe dizem por quê. É importante que você saiba se todos esses itens são importantes antes de decidir sobre a credibilidade ou não dos resultados mostrados por alguém.

Outra forma de omitir informações é através da remoção de dados que não se encaixam ao modelo pretendido (em outras palavras, “falsificação” dos dados). Suponha que um pesquisador registre a quantidade de tempo gasto na Internet e a relação com a idade. Ele ajusta uma bela reta a seus dados, indicando que as pessoas mais jovens navegam na Internet muito mais do que as pessoas mais velhas, e que o tempo de navegação diminui conforme aumenta a idade. Tudo certo, exceto por Claude, o outlier, que tem 80 anos e usa a Internet dia e noite, liderando sua própria sala de bingo, chat e mais. O que fazer com Claude? Se não fosse ele, essa relação teria um gráfico lindo. Que mal teria se o tirássemos? Afinal, é só uma pessoa, certo?

Nem pensar. Essa ideia tem tudo de errado. Remover os pontos indesejados de um conjunto de dados não é apenas errado, mas também muito arriscado. A remoção de uma observação de um conjunto de dados só é permitida quando se tem certeza de que a observação está mesmo errada. Por exemplo, alguém escreve uma pesquisa dizendo que passa 30 horas por dia na Internet ou que seu QI é 2.200.

- ✓ **Os comunicadores:** Ao relatar os resultados estatísticos, a mídia omite informações importantes o tempo todo, muitas vezes, devido a limitações de espaço e prazos apertados. No entanto, parte dela é resultado do ritmo acelerado da sociedade de hoje que se alimenta de frases de efeito. O melhor exemplo é resultado de pesquisas em que a margem de erro não é comunicada. Nesse caso, não se pode julgar a precisão dos resultados.
- ✓ **Os consumidores:** O público em geral também desempenha um papel na mentalidade de omissão. As pessoas ouvem uma notícia e, de imediato, acreditam que seja verdade, ignorando qualquer possibilidade de erro ou viés nos resultados. Por exemplo, você precisa decidir qual carro comprar e pede a opinião de seus



vizinhos e amigos em vez de examinar as pesquisas e as avaliações abrangentes e meticolosas resultantes. Em um momento ou outro, todos se esquecem de fazer as perguntas que deveriam, o que indiretamente alimenta todo o problema.

Na cadeia de informação estatística, os produtores (pesquisadores) precisam ser abrangentes e realistas quanto ao processo que conduziram e os resultados que obtiveram. Os comunicadores da informação (a mídia) necessitam avaliar criticamente a precisão das informações que estão recebendo e informá-las de forma justa. Os consumidores da informação estatística (nós) precisam parar de achar que todos os resultados são verdadeiros e buscar fontes confiáveis de estudos e análises estatísticas para tomar as decisões importantes em suas vidas.



No fim das contas, se um conjunto de dados parecer muito bom, provavelmente ele seja mesmo. Se o modelo se encaixar perfeitamente, desconfie. Se ele se encaixar bem, corra e não olhe para trás! Às vezes, o que é omitido fala muito mais do que o que é relatado.

## Capítulo 22

# Dez Formas de Chegar na Frente por Saber Estatística

---

### *Neste Capítulo*

- ▶ Sabendo o que procurar
  - ▶ Sendo cético e confiante
  - ▶ Juntando as peças do quebra-cabeça estatístico e verificando suas respostas
  - ▶ Sabendo a melhor forma de apresentar seus resultados
- 

**U**m dos meus objetivos pessoais no ensino de Estatística é ajudar as pessoas a serem capazes de dizer: “Espere um pouco!” e interromper uma análise errada ou um gráfico enganoso que cruzarem seus caminhos. Eu também quero ajudá-las a se tornarem os gurus da Estatística em seus locais de trabalho — aquelas pessoas que não têm medo de trabalhar com a Estatística e o fazem corretamente e com segurança (mas também sabem o momento de consultar um estatístico profissional). Este capítulo lhe apresenta dez maneiras de confiar em seus instintos estatísticos e aumentar seu valor profissional através de uma compreensão crítica da Estatística.

# *Faça as Perguntas Certas*

Todo estudo, experimento e toda pesquisa são feitos porque alguém tem uma pergunta e quer que ela seja respondida. Por exemplo: “Quanto tempo esta garantia deve durar?”, “Qual é a chance de que eu desenvolva complicações durante a cirurgia?”, “O que o povo americano acha sobre a proibição de fumar em público?” Só depois da clara definição de uma pergunta é que se pode começar a coleta adequada de dados.

Suponha que o dono de um restaurante me diga que quer realizar uma pesquisa para saber mais sobre a clientela do restaurante. Falamos sobre muitas variáveis, inclusive o número de pessoas na mesa, quantas vezes elas já estiveram lá, o tipo de comida pedido, a quantia paga, quanto tempo elas ficam, e assim por diante. Depois de coletar alguns dados e processar os resultados, de repente, ele tem um insight: o que ele realmente quer fazer é comparar a clientela do almoço à clientela do jantar. A clientela do jantar gasta mais? São mais velhos? Ficam mais tempo no restaurante? Mas, infelizmente, ele não pode responder a qualquer dessas perguntas, pois não mencionou na coleta de dados que queria saber se os clientes estavam lá para almoçar ou jantar.

O que aconteceu aqui é um erro muito comum. O dono do restaurante disse que “só queria estudar” sua clientela, mas nunca mencionou comparações, pois não havia pensado nisso antes. Se ele tivesse pensado nisso antes, teria percebido que a verdadeira questão era: “Como é a clientela do almoço em comparação com a clientela do jantar?” e, assim, a inclusão de uma questão sobre se os comensais estavam lá para almoçar ou jantar teria sido óbvia. Sempre faça as perguntas certas para obter as respostas de que precisa.



Testar as águas antes de mergulhar de cabeça em um estudo pode ser muito útil. Uma maneira de fazer isso é realizar o que os pesquisadores chamam de estudo piloto. Um *estudo piloto* é um pequeno estudo exploratório utilizado como teste para o estudo real. Por exemplo, você planeja uma pesquisa e a testa em um pequeno grupo para ver se os participantes encontram perguntas confusas, redundâncias, erros de ortografia, e assim por diante. Os estudos piloto são uma maneira rápida e barata de garantir que tudo saia bem quando o estudo real for colocado em prática.



# Seja Cético

Ser estatisticamente cético é uma coisa boa (dentro de um limite racional). Algumas pessoas desistiram da Estatística, pensando que as pessoas podem dizer o que quiserem se manipularem suficientemente os dados. Portanto, aqueles que têm um nível saudável de ceticismo podem chegar à frente no jogo.

Tabelas e gráficos coloridos podem chamar sua atenção, especialmente se tiverem legendas pequenas e organizadas, mas os longos e detalhados relatórios profissionais podem lhe mostrar informações além das que você deseja saber, todas dispostas em belas tabelas, página após página. O que é mais importante, no entanto, não é como a informação está disposta ou se ela soa profissional ou científica. O mais importante é o que aconteceu nos bastidores, estatisticamente falando, para produzir resultados corretos, justos e claros.



Muitas pessoas conhecem Estatística o suficiente para serem perigosas, e muitos resultados apresentados estão incorretos, por engano ou por planejamento (infelizmente). É melhor ser cético do que remediar!

Veja como fazer um bom uso de seu ceticismo:

- ✓ Obtenha uma cópia das perguntas do questionário usado na pesquisa. Se as perguntas forem enganosas, os resultados da pesquisa não são confiáveis.
- ✓ Saiba mais sobre o processo de coleta de dados. Quando a pesquisa foi realizada? Quem foi selecionado para participar? Como era a informação coletada? As pesquisas realizadas na Internet e as que se baseiam em ligações quase sempre são tendenciosas, e seus resultados devem ser jogados fora.
- ✓ Saiba mais sobre a taxa de resposta da pesquisa. Quantas pessoas foram inicialmente recrutadas? Quantas responderam? Se muitos foram recrutados, mas poucos responderam, é muito provável que os resultados sejam tendenciosos, pois os respondentes de uma pesquisa geralmente são pessoas que possuem opiniões mais fortes do que aqueles que não responderam.

# *Colete e Analise os Dados Corretamente*

Por um lado, é muito importante pensar de forma crítica e até mesmo ser cético a respeito dos resultados estatísticos com os quais você se depara em sua vida diária e no trabalho. Sempre questione antes de considerar os resultados confiáveis.

Por outro lado, é muito importante lembrar que os outros também avaliam criticamente seus resultados e, portanto, é preciso evitar o ceticismo que você vê os outros receberem. Para evitar potenciais críticas direcionadas a seus resultados, é preciso se certificar de ter feito tudo certo.

Agora que está lendo este livro, você vai ter muitas ferramentas para ajudá-lo a fazer a coleta e a análise dos dados de forma correta. Em cada capítulo, bato sempre na mesma tecla: usar a análise errada ou fazer análises demais não é bom. Para cada tipo de análise que apresento, também mostro como se certificar de que a análise em questão é a ideal para ser usada com os dados que você tem em mãos. Os Capítulos 1 e 2 servem como referência para as técnicas necessárias e lhe mostram onde encontrá-las no livro.

Noventa por cento do trabalho envolvido em uma análise estatística acontece antes mesmo dos dados serem inseridos no computador. Veja aqui uma lista básica do que é preciso verificar:

- ✓ Planeje sua pesquisa, experimento ou estudo para evitar o viés e garantir precisão.
- ✓ Certifique-se de realizar o estudo no momento certo e de selecionar uma amostra verdadeiramente aleatória de participantes.
- ✓ Acompanhe esses participantes para garantir que o resultado final tenha uma alta taxa de resposta.

Esta lista pode ser desafiante, mas, no final, você se sentirá seguro ao saber que seus resultados irão fazer frente às críticas, pois tudo foi feito da maneira correta.



# *Pedindo Ajuda*

Uma das coisas mais difíceis de ser entendida pelos não estatísticos é que eles não têm que fazer tudo sozinhos. Na verdade, não é o melhor caminho a percorrer, na maioria dos casos. As seis palavras mais importantes para qualquer não estatístico são “Saber quando consultar um estatístico profissional”. Saiba quando pedir ajuda. E o melhor momento para pedir ajuda é *antes* de coletar os dados.

Então, como saber seu limite e o momento que vai precisar que alguém lhe jogue uma tábua de salvação estatística? Aqui, vão alguns exemplos para ajudá-lo a ter uma ideia de quando pedir ajuda:

- ✓ Se seu chefe quer nada menos do que um relatório de 100 páginas sobre os resultados de marketing em sua mesa até segunda-feira e você ainda não coletou nenhum dado nº 1, PEÇA.
- ✓ Se estiver lendo *Cosmopolitan* na hora do almoço e pretende analisar como você e suas amigas se saíram no quiz “Quem é a rainha da fofoca em seu local de trabalho?”, NÃO PEÇA.
- ✓ Se a lista de perguntas EM sua pesquisa ficar maior do que você, PEÇA.
- ✓ Se quiser fazer um gráfico de barras de quantos dos seus amigos do Facebook são fãs da década de 1970, 80 ou 90, NÃO PEÇA.
- ✓ Se o diagrama de dispersão de seus dados se parecer com um teste de borrões de Rorschach, PEÇA (e rápido!).
- ✓ Se quiser saber as chances de que alguém que você não vê desde o Ensino Médio esteja no mesmo avião que você indo para a África, NÃO PEÇA.
- ✓ Se tiver que fazer um trabalho importante, o qual envolva Estatística, e não tiver certeza de como começar ou como analisar seus dados, uma vez que já os tenha em mãos, PEÇA. Quanto mais cedo você pedir ajuda, mais os profissionais poderão ajudá-lo!

# *Refazendo os Passos de Outras Pessoas*

Em algum ponto de sua vida profissional, você vai pegar um relatório, lê-lo e ficará com dúvidas. Então, vai encontrar os dados, e, depois de muito procurar, vai abrir uma planilha com linhas e mais linhas e colunas e mais colunas de números e caracteres. Estarrecido, você, então, se dá conta de que não tem ideia do que tudo aquilo significa. Depois de alguns minutos, vai dizer a si mesmo para não entrar em pânico e encontrar a pessoa que inseriu todos aqueles dados e descobrir o que está acontecendo afinal.

Mas aí vem a má notícia. Há alguns anos, um tal de Bob foi quem coletou os dados e os organizou nas planilhas, mas ele não trabalha mais na empresa. O que fazer agora? Mais do que nunca, você deve abandonar os dados e o relatório, começar tudo de novo do zero e perder tempo e dinheiro nesse processo.

Como este desastre poderia ter sido evitado? Todas as questões a seguir deveriam ter sido tratadas antes que Bob passasse seu relatório:

- ✓ O relatório deve incluir alguns parágrafos que descrevam como e quando os dados foram coletados, os nomes das variáveis no conjunto de dados, onde elas se localizam na planilha e como estão identificadas.
- ✓ O relatório deve incluir uma nota sobre qualquer falta de dados. A falta de dados às vezes é deixada em branco, mas também pode ser indicada por um sinal negativo (–) ou um ponto decimal. (O uso de zeros para indicar a falta de dados não é recomendado, pois pode se pensar que o valor real do dado é zero.)
- ✓ As linhas do conjunto de dados devem ser definidas. Por exemplo, cada linha representa uma pessoa? Elas têm números de identificação?



Infelizmente, muitas pessoas criam relatórios estatísticos e depois desaparecem sem deixar rastro, deixando para trás uma bagunça de dados que muitas vezes não tem conserto. É de bom senso que você tome algumas medidas para não deixar outras pessoas em apuros, como Bob fez. Sempre deixe uma pista para a pessoa que continuará de onde você parou. E, se estiver do outro lado, sempre peça explicações completas sobre o conjunto de dados antes de usá-lo.

# *Juntando as Peças*

Nunca caia de cabeça em uma análise esperando obter um único número como resposta e então sair fora. Estatística requer muito mais trabalho! Encare todos os problemas estatísticos como um quebra-cabeça cujas peças precisam ser juntadas antes que se possa ver tudo o que está realmente acontecendo.

Por exemplo, suponha que um vendedor de café queira prever a quantidade de café que deveria preparar para vender em um jogo de futebol em Buffalo, Nova York. Seu primeiro passo é pensar em quais variáveis podem estar relacionadas com as vendas de café. As variáveis podem ser o custo do café, a facilidade de carregá-lo, a localização do assento (quem vai querer andar um quilômetro para comprar café?) e a idade dos torcedores. O vendedor também suspeita que a temperatura no dia do jogo possa afetar as vendas de café, sendo que as baixas temperaturas se traduzem em mais vendas.

O vendedor, então, coleta dados sobre todas essas variáveis, explora as relações e descobre que as vendas de café e a temperatura se relacionam um pouco. Mas será que há algo mais além da temperatura?

Para descobrir, o vendedor compara as vendas de café em dois jogos com a mesma temperatura e percebe uma grande diferença. Analisando mais a fundo, ele percebe que um jogo caiu em um domingo e o outro, em uma segunda-feira. O comparecimento foi maior na segunda-feira, e a maioria eram adultos. Ao analisar os dados, o vendedor constatou que a temperatura realmente está relacionada com as vendas de café, mas o comparecimento do público, o dia da semana em que a partida acontece e a idade dos torcedores também estão. De posse desta informação, o vendedor foi capaz de prever as vendas de café com maior precisão, diminuindo, assim, as chances de que seu estoque de café acabe ou sobre demais. Esse exemplo demonstra que pode realmente valer a pena juntar as peças a fim de visualizar o todo.

# Verificando Suas Respostas

Depois que os dados forem analisados e você obtiver os resultados, é preciso dar mais um passo antes de correr eufórico para seu chefe, dizendo: “Veja isto!”. Você precisa ter certeza de que tem as respostas certas.



Quando digo respostas certas, não quero dizer que você precisa ter os resultados que seu chefe quer ouvir (apesar de que isso seria ótimo, é claro), mas, sim, que você precisa ter certeza de que sua análise de dados e cálculos estão corretos para que não fique sem chão quando as perguntas começarem a surgir. Siga estes passos básicos:

1. **Verifique se você digitou os dados corretamente e ceife os números que, obviamente, não fazem nenhum sentido (como alguém que diz ter 200 anos ou que vendeu 500 bilhões de lâmpadas em sua loja no ano passado).**

Os erros influenciam os dados e os resultados, então, acabe com eles antes que seja tarde demais.

2. **Certifique-se de que a soma dos números bate quando se espera que isso aconteça.**

Por exemplo, se você coletar dados sobre o número de funcionários em 100 empresas e não listar um número de grupos suficiente para abranger todos eles, estará em apuros! Também fique atento com dados de indivíduos que possam ter sido inseridos duas vezes. Esse erro aparece quando os dados são classificados por linhas.

3. **Se pretende tirar conclusões, certifique-se de estar usando os números certos para isso.**

Se quiser falar sobre como a criminalidade aumentou em seu bairro nos últimos cinco anos, mostrar um gráfico com o número de crimes não é a forma correta. O número de crimes pode aumentar simplesmente em virtude do aumento da população. Para chegar a conclusões estatísticas corretas sobre a criminalidade é necessário informar o índice de criminalidade, que é o número de crimes por pessoa (per capita), ou o número de crimes a cada 100.000 pessoas. Basta dividir o número de crimes pelo tamanho da população ou por 100.000, respectivamente. Esta abordagem não considera o tamanho da população.

# *Explicando a Saída*

Os computadores certamente desempenham um papel importante no processo de coleta e análise de dados. Existem diversos programas estatísticos, incluindo o MS Excel, o Minitab, o SAS, o SPSS, e vários outros. Cada tipo tem seu próprio estilo de imprimir os resultados. Entender, interpretar e explicar a saída desses softwares é uma arte e uma ciência que nem todo mundo conhece. Com seus conhecimentos em Estatística, no entanto, você pode ser uma dessas pessoas!

A saída do software é a forma bruta dos resultados de sínteses ou análises estatísticas. Podem ser gráficos, tabelas, diagramas de dispersão, resultados de análise de regressão, uma tabela de análise de variância ou um conjunto de estatísticas descritivas. Muitas vezes, a análise é nomeada pelo software; por exemplo, a ANOVA indica que uma análise da variância foi realizada (veja o Capítulo 9). Gráficos e tabelas, no entanto, exigem que o usuário informe ao software os nomes, títulos ou legendas (se houver) que devem ser incluídos, para que o público possa entender rapidamente o que é o quê.

Interpretar a saída desses softwares é peneirar o que pode parecer um monte de informações intimidantes. O truque é saber exatamente quais os resultados que você quer e onde o software os coloca na saída. Por exemplo, na saída de uma análise de regressão, você encontra a equação da reta de regressão na coluna COEF da saída (veja o Capítulo 4).

Na maioria das vezes, a saída traz informações não necessárias; às vezes, também traz informações que você não entende. Antes de pular tudo, consulte um estatístico para se certificar de que não está deixando de fazer algo importante, tal como examinar o coeficiente de correlação antes de fazer uma análise de regressão (veja o Capítulo 4).



Explicar o que você encontrou na saída do software também aumenta o conhecimento da pessoa a quem está se direcionando. Se você estiver escrevendo um relatório executivo, não precisa explicar passo a passo do que fez e por que, basta usar as partes da saída que dizem o essencial e explicar como isso afeta a empresa. Se estiver ajudando um colega a entender os resultados, forneça algumas informações de referência sobre as análises (você pode usar este livro). Por exemplo, pode discutir a função geral de um histograma antes de falar sobre os resultados de seu histograma.

Mas o mais importante é verificar se a análise está correta antes de explicá-la a outra pessoa. Às vezes, ao analisar os dados, é possível que você clique na variável errada ou selecione a coluna de dados errada, o que faz com que a análise saia totalmente errada.

# ***Fazendo Recomendações Convincentes***

À medida que se sobe na escala corporativa, menos tempo se tem para ler os relatórios e examinar cuidadosamente as estatísticas. A melhor análise de dados do mundo não significa nada se você não conseguir comunicar seus resultados a alguém que não tenha tempo ou interesse em conhecer os detalhes dela. Em um mundo orientado por dados, a Estatística pode desempenhar um papel importante no processo decisório. A habilidade de usar a Estatística para criar um argumento eficiente, fortalecer um exemplo ou fazer recomendações sólidas é fundamental.

Ponha-se na seguinte situação. Você fez o trabalho, coletou dados de marketing e vendas, fez as análises e processou os resultados. Baseado em seu estudo de posicionamento de produto para seu refrigerante, você determinou que a melhor estratégia de posicionamento para este produto nas prateleiras de supermercado é a de colocá-los no corredor do caixa ao nível dos olhos, para que as crianças possam vê-lo. (Afinal, você nunca vê cortadores de unha ou desinfetantes para as mãos nas prateleiras ao nível dos olhos das crianças no corredor do caixa, não é?) O fato é que seu chefe prefere colocar este produto no corredor de doces da loja. (É claro que ele não tem dados que sustentem essa decisão e está apenas se baseando em sua própria experiência de observação no corredor de doces.) Como convencê-lo a seguir sua recomendação?

Provavelmente, a pior coisa que você pode fazer é entrar em seu escritório com um relatório de cem páginas contando tudo nos mínimos detalhes. Cargas de informações complexas podem impressionar sua mãe, mas não vão impressionar seu chefe. Guarde o relatório caso ele o peça (ou caso você precise de aparador de porta). O que você precisa é de um relatório breve, sucinto e que vá direto ao ponto. Veja como fazer um:

- 1. Comece com uma descrição do problema.**

**“Queremos determinar qual local vai proporcionar uma maior venda do refrigerante.”**

- 2. Descreva sucintamente o processo de coleta de dados.**

**“Escolhemos aleatoriamente 50 lojas e colocamos o produto no corredor do caixa em 25 lojas e no corredor de doces nas outras 25. Também controlamos outros fatores como o número de produtos colocados.”**

- 3. Descrevas os dados coletados.**

**“Acompanhamos as vendas do produto durante um período de seis meses, calculando o total das vendas semanais de cada loja”. Neste ponto, mostre os gráficos e as tabelas para as vendas dos dois grupos ao longo do período estudado.**

- 4. Descreva resumidamente como você analisou os dados, mas gaste mais tempo com suas conclusões.**



Não mostre a saída do software — seu chefe não precisa ver isso. Você conhece a expressão “Nunca deixe vê-lo suar”? Isso é importante aqui. O que você quer dizer é: “Fizemos uma análise estatística comparando a venda média nesses locais e descobrimos que as vendas no corredor do caixa são significativamente maiores do que as vendas no corredor de doces”. Quantifique a diferença com porcentagens.

Continue com sua recomendação quanto à colocação do produto no corredor do caixa, ao nível dos olhos das crianças, certificando-se de responder à questão inicial, com a qual iniciou o primeiro passo. Então, a dica mais importante é deixar que seu chefe ache que a escolha pelo melhor posicionamento foi tudo ideia dele!

# ***Estabelecendo-se como o Cara da Estatística***

Em se tratando do trabalho, nada é mais valioso do que alguém que não tem medo de fazer a estatística. Cada escritório tem uma pessoa com coragem para calcular, a confiança para fazer os intervalos de confiança, a vontade de interpretar a saída e o bom senso para fazer um gráfico. Esta pessoa é, quase sempre, amiga de todos e a primeira pessoa a saber quando é preciso iniciar um novo trabalho.

Quais são as vantagens de ser o cara da Estatística? É a glória de saber que você está salvando o dia, fazendo um ponto para o time e se mantendo em pé diante do caos. Seus colegas vão dizer: “Te devo uma”, e você pode cobrá-los.

Mas, falando sério, o cara da Estatística tem mais segurança no emprego, pois seu chefe sabe que Estatística é essencial ao trabalho e que ter alguém que se arrisque quando necessário é inestimável.



A Estatística e as análises estatísticas podem intimidar, mas são fundamentais no trabalho. Para ocupar a maioria dos cargos de hoje em dia é preciso saber como selecionar amostras, escrever pesquisas, estabelecer o processo de coleta de dados e analisá-los.

# Capítulo 23

## Dez Empregos Legais que Usam Estatística

---

### *Neste Capítulo*

- ▶ Usando a estatística em uma ampla gama de empregos (o seu pode ser o próximo!) X Passando por pássaros, esportes e crimes
  - ▶ Levando a estatística para o mundo profissional da Medicina, do Direito e das Finanças
- 

**E**ste livro pretende servir de guia para aqueles que precisam saber Estatística em seu cotidiano (ou seja, todos nós), bem como aqueles que precisam dela no trabalho (a maioria de nós). Se parar para pensar, consigo encontrar alguma utilização da Estatística para quase todas as profissões (exceto, talvez, para um psicólogo).

Este capítulo apresenta dez carreiras que empregam a Estatística de alguma forma. Você pode se surpreender ao se dar conta da frequência com que a Estatística é usada em seu trabalho! Então, não queime este livro quando seu curso acabar; talvez, ele lhe seja útil na busca de emprego ou em seu trabalho. (Meu contador tem um exemplar deste livro na estante — o que isso quer dizer? Desde que ele não tenha uma cópia do *Matemática Básica Para Leigos*, acho que tudo bem.)

Um de meus objetivos pessoais, como professora de Estatística, é ajudar meus alunos a se destacarem no trabalho. Você sabe, quero ajudá-los a se tornarem aquela pessoa que entende de Estatística, que sabe o que está fazendo e quando fazer, pois sabe encontrar a estatística necessária de forma correta e com segurança. Com experiência e com a ajuda deste livro, você também poderá se tornar essa pessoa. Você vai se tornar um herói, e sua segurança no emprego será maior.

# *Pesquisadores de Opinião Pública*

Os pesquisadores de opinião pública coletam informações sobre pessoas de suas populações de interesse. Alguns dos grandes nomes na pesquisa de opinião pública profissional incluem a Organização Gallup, a Associated Press (AP), a Zogby International, a Harris Interactive e o Centro de Pesquisa Pew. Grandes agências de notícias tais como a NBC, CBS, CNN também realizam pesquisas, assim como muitas outras agências e organizações.

Os objetivos das pesquisas passam pela área médica, que tentam determinar o que está causando a obesidade, pelas pesquisas de opinião pública sobre política, que querem acompanhar o pulso diário da opinião pública *americana* e pelas pesquisas que fornecem feedback e ideias para as corporações.



O conhecimento em Estatística é considerado importantíssimo na indústria de pesquisa de opinião pública, pois os empregos nessa área podem incluir o planejamento das pesquisas, a seleção de amostras adequadas de participantes, a realização de uma pesquisa para coletar dados e o registro, a análise e a apresentação dos resultados.

Todas essas tarefas fazem parte da Estatística — a arte e a ciência de coletar e dar sentido a dados. Mas não sou só eu quem diz isso, veja abaixo um trecho de um anúncio de emprego para a Organização Gallup, que buscava um Analista de Pesquisa. Tenho que dizer que tudo o que pedem é ESTATÍSTICA!

Se você tem um forte histórico acadêmico em Ciências Sociais ou em Economia, está familiarizado com pesquisas quantitativas e categóricas e ferramentas estatísticas para Pesquisa de Mercado/Pesquisa de Opinião ou Consultoria, adora reunir dados de pesquisas e conceitos abstratos para descobrir algo significativo, enquanto continua a aprender, este é o lugar para gerenciar processos e projetos que resultem na conclusão perfeita garantida aos clientes.

E, aqui, temos algo que não se vê todos os dias. Encontrei um anúncio de emprego para analista de pesquisa de opinião com quase as mesmas exigências, mas para um ambiente de trabalho muito diferente. O emprego era para uma empresa que oferece segurança e inteligência para o governo dos Estados Unidos. Para este trabalho, é necessário ter uma habilitação de segurança federal. Você nunca sabe para onde o seu conhecimento em Estatística pode levá-lo!

Outros empregos relacionados à pesquisa de opinião pública que encontrei são especialistas em pesquisa quantitativa e analista de pesquisa de opinião pública.



Para saber mais sobre o que os pesquisadores de opinião pública fazem e como é o seu trabalho, visite o site [www.pollster.com](http://www.pollster.com) (conteúdo em inglês).

# Ornitólogo (*Observador de Pássaros*)

Todos nós, eventualmente, observamos pássaros. Admito ser uma observadora de pássaros semisséria, que sempre acompanha o Magee Marsh no Lago Erie em maio no Dia Internacional das Aves Migratórias. Mas você já pensou em ser pago para observar pássaros e outros animais selvagens? A conscientização cada vez maior sobre a preservação do meio ambiente foca a identificação, o estudo e a proteção de todos os tipos de vida selvagem.

A Ornitologia é a ciência que estuda os pássaros. Os ornitólogos estão sempre coletando dados, descobrindo e estudando as estatísticas sobre os pássaros — muitas vezes sobre um certo tipo de pássaro e seu comportamento. Alguns exemplos comuns de estatísticas sobre pássaros incluem:

- ✓ A contagem de pássaros (número de aves por unidade quadrada de espaço em um determinado dia)
- ✓ Localização de ninhos e mapas de território
- ✓ Número de ovos postos e eclodidos
- ✓ Preferências alimentares e técnicas de forrageamento
- ✓ Comportamentos filmados e quantificados



Você pode achar um site totalmente dedicado a empregos para ornitólogos e observadores da vida selvagem e verificar o uso de seus conhecimentos em Estatística. Veja aqui uma das ofertas de trabalho fornecida pela Sociedade de Ornitologia:

**TÉCNICOS EM PESQUISA SOBRE CORUJA-FLAMADA (2)** para Observatório de Pássaros de Idaho para estudo de corujas-flamadas e outros pássaros da floresta de Idaho (aprox. dois meses e meio). O trabalho consiste principalmente na realização de pesquisas padronizadas e na inserção de dados. Os candidatos devem ter as seguintes qualificações: 1) ter boa visão e audição, 2) proficiência nos procedimentos de pesquisas padronizadas, 3) capacidade para identificar as aves ocidentais através da visão e do som e 4) vontade de fazer o melhor. Os candidatos devem estar fisicamente aptos e não se intimidarem com as perspectivas de calor, umidade, insetos e lama. (Verdade!)

Depois que adquirir mais experiência e conhecimento, você poderá se tornar um biólogo pesquisador da vida selvagem no Departamento de Pesquisa Geológica Interior dos Estados Unidos. A descrição do emprego para este cargo colocado hoje, na verdade, exige 15 horas de créditos em Estatística, provando que o governo está por dentro dessa coisa de Estatística.

# *Comentarista ou Jornalista Esportivo*

Todo bom jornalista ou comentarista esportivo sabe que você não é nada se não tiver as estatísticas certas. Você faz seu dever de casa, estuda os campos de treinamento e se debruça sobre planilhas e dados históricos. Também lê os jornais, consulta os livros de recordes e assiste a filmes. A fonte de dados é inesgotável, mas seu público nunca está satisfeito.

Torcedores são viciados em estatísticas! (Sendo torcedora do Ohio State Buckeye, posso dizer que sou tão fanática quanto.) Coloco aqui apenas uma amostragem das estatísticas registradas e apresentadas pelo meu time de futebol americano:

- ✓ Pontos marcados
- ✓ Pontos contra
- ✓ Rushing yards
- ✓ Receiving yards
- ✓ Passing yards
- ✓ Intercepções
- ✓ Fumbles
- ✓ Punt and kick returns
- ✓ Número e distância de field goals (tentados e feitos)
- ✓ Kicker com a carreira mais longa
- ✓ Número de first downs
- ✓ Third down conversions
- ✓ Fourth down conversions
- ✓ Penalidades
- ✓ All-purpose yards
- ✓ Total offense
- ✓ Total defense
- ✓ Número de sacks
- ✓ Rushing defense
- ✓ Passing defense
- ✓ Turnover margins
- ✓ Eficiência de passe
- ✓ Pontuação do ataque

- ✓ Pontuação da defesa
- ✓ Pontuação dos times especiais
- ✓ Classificação dos técnicos
- ✓ Classificação do AP
- ✓ Classificação BCS (o que daria um outro livro!)
- ✓ Os 12 melhores + vencedor da temporada
- ✓ Registros dos técnicos
- ✓ As pontuações mais altas
- ✓ O público presente na partida
- ✓ Resistência da programação
- ✓ Ganho e perda de arranque
- ✓ Vencedores no cara e coroa

É óbvio que precisamos de um novo ditado: “Os que praticam esportes jogam, mas o que assistem calculam as estatísticas.”

# *Jornalista*

Os jornalistas de todos os tipos em uma ou outra situação têm que trabalhar com dados. E esse é um osso duro de roer, pois os dados que chegam a eles vêm de inúmeros canais e falam sobre uma infinidade de temas, mas os jornalistas precisam dar sentido a eles, escolher o que acreditam ser mais importante, peneirá-los, apresentar os resultados e ainda escrever uma história em torno deles, tudo dentro de um prazo muito apertado (por vezes, apenas algumas horas). Isso que é trabalho!

Como consumidora da mídia, vejo muito o uso de estatísticas claras, corretas e que chamam atenção para coisas interessantes e importantes. No entanto, também vejo muitas estatísticas sendo usadas de forma incorreta e enganosa pela mídia, e ainda me assusto.

Alguns dos problemas mais comuns incluem simples erros de matemática, como relatar porcentagens superiores a 100%, a suposição de relações de causa e efeito que não são comprovadas, o uso de gráficos enganosos e a omissão de informações (tais como o número de pessoas pesquisadas, a taxa de não respondentes e a margem de erro). Mas, para mim, o maior problema é ler uma manchete atraente, e até mesmo chocante, só para descobrir que ela não é corroborada pelas estatísticas apresentadas no artigo.

Ter concluído alguns cursos de Estatística o coloca à frente em qualquer entrevista de emprego para um cargo de jornalista. Os estatísticos de todo o mundo estão esperando que você chegue lá, montado em seu cavalo branco, e faça as coisas direito! (Não se esqueça de colocar este livro em seu alforje!).

Em reconhecimento da importância e em valorização da difícil tarefa que os jornalistas exercem ao usar e relatar as estatísticas, a Royal Statistical Society criou um prêmio para a Excelência em Estatística no Jornalismo. A seguir, veja a descrição do prêmio, e eu não poderia deixar de concordar com ele!

A Royal Statistical Society deseja incentivar a excelência no uso da estatística pelos jornalistas para questionar, analisar e investigar os problemas que afetam a sociedade em geral. A excelência jornalística em estatística vai ajudar no processo decisório de todos os setores considerados — por meio da comunicação acessível de informações complexas, destacando o acontecimento e a exposição de importantes informações omitidas.



# *Combatentes do Crime*

As estatísticas sobre a criminalidade ajudam os combatentes do crime, como os policiais, a determinar quais tipos de crimes ocorrem, onde ocorrem, com que frequência, contra quem e por quem são realizados. As estatísticas de criminalidade em toda a nação são compiladas e analisadas pelo Departamento de Justiça dos Estados Unidos. Pesquisas de Vitimização da Criminalidade Nacional também são realizadas para ajudar a entender as tendências dos crimes de vários tipos.

Os policiais registram todos os incidentes em que se envolvem, formando grandes bancos de dados que podem ser usados por oficiais da cidade e do Estado para determinar o número de policiais necessários e quais as áreas que devem receber mais atenção, além, também, de serem usados para as mudanças nas políticas e nos procedimentos de seus departamentos policiais. O Polícia Federal também pode usar esses enormes bancos de dados para rastrear criminosos, procurar por padrões nos tipos de crimes e ocorrências e acompanhar a tendência geral do número de crimes, bem como o tipo de crime que ocorre ao longo do tempo.

As pessoas que estiverem procurando uma casa ou uma escola nova podem consultar gratuitamente as informações disponíveis sobre as estatísticas de criminalidade. Já os políticos as usam para mostrar que a criminalidade está subindo ou caindo, que as verbas devem ou não serem gastas com mais policiais e o quão segura sua cidade ou Estado se tornou depois que eles assumiram o mandato.

Aqui vai uma visão geral de como o *site* do Departamento de Justiça dos EUA utiliza os dados e as estatísticas para ajudar a combater o crime:

Todos os Estados estabeleceram um registo criminal repositório que mantém os registros criminais e os dados de identificação e responde a inquéritos policiais e inquéritos para outros fins, como a verificação de antecedentes e de segurança nacional. Os registros criminais incluem os dados fornecidos por todos os componentes do sistema de justiça penal: aplicação da lei, Ministério Público, tribunais e corregedorias. Os registros desenvolvidos para fins estatísticos descrevem e classificam cada incidente criminal e incluem dados sobre as características do agressor, as relações entre o agressor e a vítima e o impacto do delito. Os dados estatísticos são extraídos dos registros operacionais de acordo com critérios uniformes de classificação e coleta. Os dados estatísticos detalhados permitem que as localidades identifiquem as áreas problemáticas e aloquem os recursos humanos e financeiros de forma eficiente e efetiva.

# *Profissional da Área Médica*

As pessoas que trabalham na área médica dependem das estatísticas para fazer pesquisas e encontrar novas curas, tratamentos, medicamentos e procedimentos para melhorar a saúde e o bem-estar das pessoas. Os pesquisadores médicos realizam ensaios clínicos para medir todos os efeitos colaterais possíveis de todos os medicamentos que passam pelo processo de aprovação. O tempo todo são realizados estudos comparativos para determinar quais fatores que influenciam o peso, a altura, o nível de inteligência e a capacidade de sobreviver a uma determinada doença. A Estatística é um meio para se ter mais certeza de que o que funciona para uma amostra de indivíduos também irá funcionar para a população à qual se destina.

Na área médica, o uso da Estatística começa logo que o nome de um paciente é chamado na sala de espera. Suponha que você seja um enfermeiro. A primeira coisa que pede ao paciente é “suba na balança, por favor”(cinco temidas palavras ouvidas em consultórios médicos ao redor do mundo). De lá, você verifica seus sinais vitais (também conhecidos como estatísticas vitais): temperatura, pressão arterial, frequência cardíaca e, por vezes, a taxa respiratória (a respiração). Você, então, registra esses números no computador e os compara ao que foi determinado como sendo o normal. O estabelecimento dos padrões de normalidade também envolve a estatística através da análise de dados históricos e da pesquisa médica.

Assim como outros tipos de estatísticas, a forma como as estatísticas vitais de uma pessoa são coletadas podem afetar significativamente os resultados. Por exemplo, existem diferentes tipos de instrumentos para aferir a pressão, e há uns melhores do que outros. E, além disso, todos sabemos que a balança no consultório do médico pesa sempre cinco quilos a mais!

# Executivo de Marketing

O marketing é fundamental para o sucesso de qualquer produto. É por isso que as empresas gastam milhões de reais em comerciais de 30 segundos durante o Campeonato Brasileiro. Pesquisar quem vai comprar seu produto, onde, quando e por quanto é um trabalho que emprega muito da Estatística.



Alguns dados são o que os estatísticos chamam de *quantitativos*, como pesquisas sobre clientes existentes, passados e potenciais, informações de vendas e tendências, informações econômicas e demográficas e dados sobre os concorrentes. Outros dados são chamados de *categóricos*, que incluem as entrevistas mais complexas (feitas uma a uma) e os grupos focais realizados com a finalidade de descobrir o que os consumidores pensam, como se sentem, quais são suas ideias e quais informações extras eles precisam ter sobre o produto. (Consulte o Capítulo 2 ou o seu livro de Estatística I para uma revisão sobre dados quantitativos e categóricos.)

Considere o exemplo da Mars, Incorporated, que fabrica os chocolates M&M's. Como o produto dessa empresa se tornou um ícone nacional para as crianças de todas as idades? O segredo só pode ser a capacidade inata da Mars de mudar com o tempo e continuar sabendo o que seus clientes querem.

A Estatística desempenha um papel enorme nesse sucesso através da coleta de dados sobre as vendas, mas, o mais importante, através da obtenção de feedback direto dos clientes por meio de entrevistas, grupos focais e pesquisas. (Nem consigo imaginar o esforço que é experimentar os M&M's e falar o que penso. "Oh espere, preciso de outra amostra antes que possa lhe dar uma boa resposta".) Ao analisar esses dados, a Mars consegue determinar alguns dos detalhes mais importantes e complicados responsáveis pelo sucesso e pela longevidade de qualquer empresa.

Por exemplo, em 1995, a Mars conduziu uma pesquisa nacional pedindo que os clientes escolhessem a mais nova cor dos M&M's. Foi quando o azul entrou em cena. Em 2002, a pesquisa se tornou global e o roxo se tornou a mais nova aquisição da paleta dos M&M's. A Mars utiliza a Estatística para encontrar maneiras de ser inovadora na fabricação de novas cores, novos sabores e estilos e, até mesmo, permitindo a personalização dos M&M's, mas ainda mantém a clássica essência do produto, a responsável pelo começo de tudo.

# Advogado

Com certeza, você já ouviu a frase “acima de qualquer suspeita”. É o código que os jurados usam para decidir se o réu é culpado ou inocente. O domínio da Estatística desempenha um papel importante para determinar se as leis estão sendo cumpridas ou se um réu é culpado ou inocente e se as leis precisam ser criadas ou alteradas. A informação estatística é uma prova muito forte.

As ações judiciais são muitas vezes resolvidas com base em dados estatísticos coletados em diversas situações semelhantes ao longo dos anos. A Estatística também permite que os legisladores decomponham as informações a fim de propor novas leis. Por exemplo, o uso da Estatística para demonstrar que as duas primeiras horas são as mais importantes para se encontrar uma criança desaparecida levou à transmissão dos Amber Alerts<sup>1</sup> pela TV e pelo rádio e à sua publicação em rodovias quando crianças desaparecem.

Promotores e advogados de defesa frequentemente usam a probabilidade e a estatística também em seus casos. Eles também devem tornar tais estatísticas compreensíveis ao júri. (Talvez, a distribuição de cópias deste livro deva ser uma exigência para o júri selecionado!) Muitas vezes, estatísticos são solicitados pelas equipes de advogados para ajudá-los a coletar informações, decifrar resultados e utilizar os dados em situação de júri popular.

Os advogados podem usar a correlação para demonstrar que certas variáveis possuem uma relação linear, como a distância de derrapagem e a quantidade de um determinado tipo de concreto no pavimento, ou a força da viga de uma ponte relacionada ao peso colocado sobre ela.

A estatística também pode ajudar a testar os argumentos. Por exemplo, suponha que a empresa de entregas A declara que seus pacotes são entregues, em média, duas horas mais rápido do que a empresa B. Se uma amostra aleatória de pacotes demorar mais do que dois dias para chegar e a diferença for grande o suficiente para gerar fortes evidências contra a afirmação da empresa A, ela poderia ser processada por propaganda enganosa. É claro que qualquer decisão baseada em estatísticas pode estar errada simplesmente em virtude do acaso. Neste caso, se a amostra aleatória de pacotes apenas levou mais tempo do que o normal e, portanto, não representa o tempo médio de entrega, a empresa de entregas A pode apelar, dizendo que foi injustamente acusada por propaganda enganosa. É como andar na corda bamba, mas os estatísticos se esforçam ao máximo para definir os procedimentos que ajudarão a verdade a vir à tona.

# *Corretor de Ações*

Os apostadores profissionais são pessoas que ganham a vida jogando. Eles conhecem os caminhos, andam por todos os lugares e são bons no que fazem: jogar os jogos que acreditam poderem ganhar por causa de sua habilidade e sorte. Existem diferentes estilos de jogadores profissionais. Um tipo prefere os jogos de cassino, como poker e blackjack.

Outro tipo de jogador é o corretor de ações que usa terno e gravata, tem almoços de negócios e está sempre mudando suas decisões ao longo do dia, dependendo do estado atual do sistema. Esses profissionais fazem previsões e decidem sobre a compra e venda de ações minuto a minuto.

A Estatística desempenha um papel fundamental para que o corretor de ações tome uma decisão bem-sucedida. Para ficar em vantagem, os corretores usam sofisticados procedimento de coleta de dados e softwares de análise, modelos financeiros e números de todas as partes do mundo, mas têm que saber analisar essas informações, interpretá-las e agir rapidamente.

No entanto, é preciso tomar cuidado com os corretores que ignoram as estatísticas reais ou inventam suas próprias estatísticas para dar a seus clientes uma visão não real do que está acontecendo com o dinheiro deles. Já houve casos em que os corretores roubaram milhões e até bilhões de dólares de seus clientes bem embaixo de seus narizes. Embora tais situações não sejam comuns, geram um enorme impacto sobre a confiança dos investidores e, em última instância, até mesmo sobre a Economia. A coleta e utilização de dados verificáveis, além da análise de dados através de técnicas legítimas, deveriam ser o objetivo de todo bom corretor de ações (e também o de todos que usam estatísticas no trabalho).

---

<sup>1</sup> N.E.: O programa Amber Alerts é uma parceria voluntária entre órgãos de combate ao crime e várias empresas americanas para divulgar casos de desaparecimento de crianças

# Apêndice

## Tabelas de Referência

***E***ste apêndice inclui as tabelas usadas por cinco importantes distribuições na Estatística II: a distribuição- $t$ , a distribuição binomial, a distribuição do Qui-quadrado, a distribuição para a estatística de teste da soma de postos e a distribuição- $F$ .

# Tabela-*t*

A Tabela A-1 mostra as probabilidades unicaudais à direita da distribuição-*t* (consulte o Capítulo 3). Para usar a Tabela A 1, você precisará de quatro informações retiradas do problema com o qual estiver lidando:

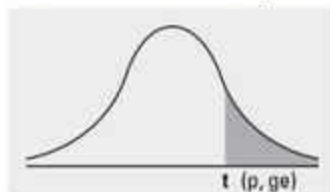
- ✓ O tamanho amostral,  $n$ .
- ✓ A média de  $x$ , denotada por  $\mu$ .
- ✓ O desvio padrão dos dados,  $s$ .
- ✓ O valor de  $x$  para o qual deseja a probabilidade da cauda direita.

Depois de obter essas informações, transforme o valor de  $x$  em uma estatística-*t* (ou valor  $t$ ), subtraindo o valor de  $x$  da média e dividindo o resultado pelo desvio padrão (veja o

Capítulo 3) através da fórmula 
$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Em seguida, encontre este valor de  $t$  na Tabela A-1, procurando a linha correspondente aos graus de liberdade para a estatística-*t* ( $n - 1$ ). Examine a linha toda até encontrar dois valores entre os quais se encontra sua estatística-*t*. Depois, vá ao topo das colunas e encontre as probabilidades. A probabilidade de que  $t$  esteja além do valor de  $x$  (probabilidade da cauda direita) estará entre estas duas probabilidades. Observe que a última coluna da tabela-*t* mostra  $gl = \infty$ , o que representa os valores da distribuição-*Z*, pois, para os tamanhos amostrais grandes,  $t$  e  $Z$  são próximas.

Área direita da distribuição-t (p, gl)



gl/p	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0005
1	0,324920	1,000000	3,077684	6,313752	12,70620	31,82052	63,65674	636,6192
2	0,288675	0,816497	1,885618	2,919986	4,30265	6,96456	9,92484	31,5991
3	0,276671	0,764892	1,637744	2,353363	3,18245	4,54070	5,84091	12,9240
4	0,270722	0,740697	1,533206	2,131847	2,77645	3,74695	4,60409	8,6103
5	0,267181	0,726687	1,475884	2,015048	2,57058	3,36493	4,03214	6,8688
6	0,264835	0,717558	1,439756	1,943180	2,44691	3,14267	3,70743	5,9588
7	0,263167	0,711142	1,414924	1,894579	2,36462	2,99795	3,49948	5,4079
8	0,261921	0,706387	1,396815	1,859548	2,30600	2,89646	3,35539	5,0413
9	0,260955	0,702722	1,383029	1,833113	2,26216	2,82144	3,24984	4,7809
10	0,260185	0,699812	1,372184	1,812461	2,22814	2,76377	3,16927	4,5869
11	0,259556	0,697445	1,363430	1,795885	2,20099	2,71808	3,10581	4,4370
12	0,259033	0,695483	1,356217	1,782288	2,17881	2,68100	3,05454	4,3178
13	0,258591	0,693829	1,350171	1,770933	2,16037	2,65031	3,01228	4,2208
14	0,258213	0,692417	1,345030	1,761310	2,14479	2,62449	2,97684	4,1405
15	0,257885	0,691197	1,340606	1,753050	2,13145	2,60248	2,94671	4,0728
16	0,257599	0,690132	1,336757	1,745884	2,11991	2,58349	2,92078	4,0150
17	0,257347	0,689195	1,333379	1,739607	2,10982	2,56693	2,89823	3,9651
18	0,257123	0,688364	1,330391	1,734064	2,10092	2,55238	2,87844	3,9216
19	0,256923	0,687621	1,327728	1,729133	2,09302	2,53948	2,86093	3,8834
20	0,256743	0,686954	1,325341	1,724718	2,08596	2,52798	2,84534	3,8495
21	0,256580	0,686352	1,323188	1,720743	2,07961	2,51765	2,83136	3,8193
22	0,256432	0,685805	1,321237	1,717144	2,07387	2,50832	2,81876	3,7921
23	0,256297	0,685306	1,319460	1,713872	2,06866	2,49987	2,80734	3,7676
24	0,256173	0,684850	1,317836	1,710882	2,06390	2,49216	2,79694	3,7454
25	0,256060	0,684430	1,316345	1,708141	2,05954	2,48511	2,78744	3,7251
26	0,255955	0,684043	1,314972	1,705618	2,05553	2,47863	2,77871	3,7066
27	0,255858	0,683685	1,313703	1,703288	2,05183	2,47266	2,77068	3,6896
28	0,255768	0,683353	1,312527	1,701131	2,04841	2,46714	2,76326	3,6739
29	0,255684	0,683044	1,311434	1,699127	2,04523	2,46202	2,75639	3,6594
30	0,255605	0,682756	1,310415	1,697261	2,04227	2,45726	2,75000	3,6460
∞	0,253347	0,674490	1,281552	1,644854	1,95996	2,32635	2,57583	3,2905



# ***Tabela Binomial***

A Tabela A-2 mostra as probabilidades para a distribuição binomial (consulte o Capítulo 17). Para usar a Tabela A-2, você precisará de três informações retiradas do problema com o qual estiver lidando:

- ✓ O tamanho amostral,  $n$ .
- ✓ A probabilidade de ocorrência,  $p$ .
- ✓ O valor de  $x$  para o qual deseja a probabilidade cumulativa.

Encontre a parte da Tabela A-2 dedicada a seu  $n$  e procure o  $x$  na linha e o  $p$  na coluna. Cruze a linha com a coluna e descubra a probabilidade para  $x$ . Para obter a probabilidade de ser estritamente menor que, maior que, maior ou igual a, ou estar entre dois valores de  $x$ , some os valores apropriados da Tabela A-2, usando o procedimento descrito no Capítulo 16.

## Tabela A-2

## A Tabela Binomial

Os números na tabela representam as probabilidades para os valores de  $x$  de 0 a  $n$ .

Probabilidades binomiais:

$$\binom{n}{x} p^x (1-p)^{n-x}$$

		$p$										
$n$	$x$	0,1	0,2	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,8	0,9
1	0	0,900	0,800	0,750	0,700	0,600	0,500	0,400	0,300	0,250	0,200	0,100
	1	0,100	0,200	0,250	0,300	0,400	0,500	0,600	0,700	0,750	0,800	0,900
2	0	0,810	0,640	0,563	0,490	0,360	0,250	0,160	0,090	0,063	0,040	0,010
	1	0,180	0,320	0,375	0,420	0,480	0,500	0,480	0,420	0,375	0,320	0,180
	2	0,010	0,040	0,063	0,090	0,160	0,250	0,360	0,490	0,563	0,640	0,810
3	0	0,729	0,512	0,422	0,343	0,216	0,125	0,064	0,027	0,016	0,008	0,001
	1	0,243	0,384	0,422	0,441	0,432	0,375	0,288	0,189	0,141	0,096	0,027
	2	0,027	0,096	0,141	0,189	0,288	0,375	0,432	0,441	0,422	0,384	0,243
	3	0,001	0,008	0,016	0,027	0,064	0,125	0,216	0,343	0,422	0,512	0,729
4	0	0,656	0,410	0,316	0,240	0,130	0,063	0,026	0,008	0,004	0,002	0,000
	1	0,292	0,410	0,422	0,412	0,346	0,250	0,154	0,076	0,047	0,026	0,004
	2	0,049	0,154	0,211	0,265	0,346	0,375	0,346	0,265	0,211	0,154	0,049
	3	0,004	0,026	0,047	0,076	0,154	0,250	0,346	0,412	0,422	0,410	0,292
	4	0,000	0,002	0,004	0,008	0,026	0,063	0,130	0,240	0,316	0,410	0,656
5	0	0,590	0,328	0,237	0,168	0,078	0,031	0,010	0,002	0,001	0,000	0,000
	1	0,328	0,410	0,396	0,360	0,259	0,156	0,077	0,028	0,015	0,006	0,000
	2	0,073	0,205	0,264	0,309	0,346	0,312	0,230	0,132	0,088	0,051	0,008
	3	0,008	0,051	0,088	0,132	0,230	0,312	0,346	0,309	0,264	0,205	0,073
	4	0,000	0,006	0,015	0,028	0,077	0,156	0,259	0,360	0,396	0,410	0,328
	5	0,000	0,000	0,001	0,002	0,010	0,031	0,078	0,168	0,237	0,328	0,590
6	0	0,531	0,262	0,178	0,118	0,047	0,016	0,004	0,001	0,000	0,000	0,000
	1	0,354	0,393	0,356	0,303	0,187	0,094	0,037	0,010	0,004	0,002	0,000
	2	0,098	0,246	0,297	0,324	0,311	0,234	0,138	0,060	0,033	0,015	0,001
	3	0,015	0,082	0,132	0,185	0,276	0,313	0,276	0,185	0,132	0,082	0,015
	4	0,001	0,015	0,033	0,060	0,138	0,234	0,311	0,324	0,297	0,246	0,098
	5	0,000	0,002	0,004	0,010	0,037	0,094	0,187	0,303	0,356	0,393	0,354
	6	0,000	0,000	0,000	0,001	0,004	0,016	0,047	0,118	0,178	0,262	0,531
7	0	0,478	0,210	0,133	0,082	0,028	0,008	0,002	0,000	0,000	0,000	0,000
	1	0,372	0,367	0,311	0,247	0,131	0,055	0,017	0,004	0,001	0,000	0,000
	2	0,124	0,275	0,311	0,318	0,261	0,164	0,077	0,025	0,012	0,004	0,000
	3	0,023	0,115	0,173	0,227	0,290	0,273	0,194	0,097	0,058	0,029	0,003
	4	0,003	0,029	0,058	0,097	0,194	0,273	0,290	0,227	0,173	0,115	0,023
	5	0,000	0,004	0,012	0,025	0,077	0,164	0,261	0,318	0,311	0,275	0,124
	6	0,000	0,000	0,001	0,004	0,017	0,055	0,131	0,247	0,311	0,367	0,372
	7	0,000	0,000	0,000	0,000	0,002	0,008	0,028	0,082	0,133	0,210	0,478

(continua)

# Tabela A-2 (continuação)

Probabilidades binomiais:		$p$										
$\binom{n}{x} p^x (1-p)^{n-x}$	$n \quad x$	0,1	0,2	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,8	0,9
	8 0	0,430	0,168	0,100	0,058	0,017	0,004	0,001	0,000	0,000	0,000	0,000
	1	0,383	0,336	0,267	0,198	0,090	0,031	0,008	0,001	0,000	0,000	0,000
	2	0,149	0,294	0,311	0,296	0,209	0,109	0,041	0,010	0,004	0,001	0,000
	3	0,033	0,147	0,208	0,254	0,279	0,219	0,124	0,047	0,023	0,009	0,000
	4	0,005	0,046	0,087	0,136	0,232	0,273	0,232	0,136	0,087	0,046	0,005
	5	0,000	0,009	0,023	0,047	0,124	0,219	0,279	0,254	0,208	0,147	0,033
	6	0,000	0,001	0,004	0,010	0,041	0,109	0,209	0,296	0,311	0,294	0,149
	7	0,000	0,000	0,000	0,001	0,008	0,031	0,090	0,198	0,267	0,336	0,383
	8 0	0,000	0,000	0,000	0,000	0,001	0,004	0,017	0,058	0,100	0,168	0,430
	9 0	0,387	0,134	0,075	0,040	0,010	0,002	0,000	0,000	0,000	0,000	0,000
	1	0,387	0,302	0,225	0,156	0,060	0,018	0,004	0,000	0,000	0,000	0,000
	2	0,172	0,302	0,300	0,267	0,161	0,070	0,021	0,004	0,001	0,000	0,000
	3	0,045	0,176	0,234	0,267	0,251	0,164	0,074	0,021	0,009	0,003	0,000
	4	0,007	0,066	0,117	0,172	0,251	0,246	0,167	0,074	0,039	0,017	0,001
	5	0,001	0,017	0,039	0,074	0,167	0,246	0,251	0,172	0,117	0,066	0,007
	6	0,000	0,003	0,009	0,021	0,074	0,164	0,251	0,267	0,234	0,176	0,045
	7 0	0,000	0,000	0,001	0,004	0,021	0,070	0,161	0,267	0,300	0,302	0,172
	8	0,000	0,000	0,000	0,000	0,004	0,018	0,060	0,156	0,225	0,302	0,387
	9	0,000	0,000	0,000	0,000	0,000	0,002	0,010	0,040	0,075	0,134	0,387
	10 0	0,349	0,107	0,056	0,028	0,006	0,001	0,000	0,000	0,000	0,000	0,000
	1	0,387	0,268	0,188	0,121	0,040	0,010	0,002	0,000	0,000	0,000	0,000
	2	0,194	0,302	0,282	0,233	0,121	0,044	0,011	0,001	0,000	0,000	0,000
	3	0,057	0,201	0,250	0,267	0,215	0,117	0,042	0,009	0,003	0,001	0,000
	4	0,011	0,088	0,146	0,200	0,251	0,205	0,111	0,037	0,016	0,006	0,000
	5	0,001	0,026	0,058	0,103	0,201	0,246	0,201	0,103	0,058	0,026	0,001
	6	0,000	0,006	0,016	0,037	0,111	0,205	0,251	0,200	0,146	0,088	0,011
	7	0,000	0,001	0,003	0,009	0,042	0,117	0,215	0,267	0,250	0,201	0,057
	8	0,000	0,000	0,000	0,001	0,011	0,044	0,121	0,233	0,282	0,302	0,194
	9	0,000	0,000	0,000	0,000	0,002	0,010	0,040	0,121	0,188	0,268	0,387
	10	0,000	0,000	0,000	0,000	0,000	0,001	0,006	0,028	0,056	0,107	0,349
	11 0	0,314	0,086	0,042	0,020	0,004	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,384	0,236	0,155	0,093	0,027	0,005	0,001	0,000	0,000	0,000	0,000
	2	0,213	0,295	0,258	0,200	0,089	0,027	0,005	0,001	0,000	0,000	0,000
	3	0,071	0,221	0,258	0,257	0,177	0,081	0,023	0,004	0,001	0,000	0,000
	4	0,016	0,111	0,172	0,220	0,236	0,161	0,070	0,017	0,006	0,002	0,000
	5	0,002	0,039	0,080	0,132	0,221	0,226	0,147	0,057	0,027	0,010	0,000
	6	0,000	0,010	0,027	0,057	0,147	0,226	0,221	0,132	0,080	0,039	0,002
	7	0,000	0,002	0,006	0,017	0,070	0,161	0,236	0,220	0,172	0,111	0,016
	8	0,000	0,000	0,001	0,004	0,023	0,081	0,177	0,257	0,258	0,221	0,071
	9	0,000	0,000	0,000	0,001	0,005	0,027	0,089	0,200	0,258	0,295	0,213
	10	0,000	0,000	0,000	0,000	0,001	0,005	0,027	0,093	0,155	0,236	0,384
	11	0,000	0,000	0,000	0,000	0,000	0,000	0,004	0,020	0,042	0,086	0,314

(continua)



# Tabela A-2 (continuação)

Probabilidades binomiais:			p										
n	x		0,1	0,2	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,8	0,9
$\binom{n}{x} p^x(1-p)^{n-x}$													
12	0		0,282	0,069	0,032	0,014	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	1		0,377	0,206	0,127	0,071	0,017	0,003	0,000	0,000	0,000	0,000	0,000
	2		0,230	0,283	0,232	0,168	0,064	0,016	0,002	0,000	0,000	0,000	0,000
	3		0,085	0,236	0,258	0,240	0,142	0,054	0,012	0,001	0,000	0,000	0,000
	4		0,021	0,133	0,194	0,231	0,213	0,121	0,042	0,008	0,002	0,001	0,000
	5		0,004	0,053	0,103	0,158	0,227	0,193	0,101	0,029	0,011	0,003	0,000
	6		0,000	0,016	0,040	0,079	0,177	0,226	0,177	0,079	0,040	0,016	0,000
	7		0,000	0,003	0,011	0,029	0,101	0,193	0,227	0,158	0,103	0,053	0,004
	8		0,000	0,001	0,002	0,008	0,042	0,121	0,213	0,231	0,194	0,133	0,021
	9		0,000	0,000	0,000	0,001	0,012	0,054	0,142	0,240	0,258	0,236	0,085
	10		0,000	0,000	0,000	0,000	0,002	0,016	0,064	0,168	0,232	0,283	0,230
	11		0,000	0,000	0,000	0,000	0,000	0,003	0,017	0,071	0,127	0,206	0,377
12		0,000	0,000	0,000	0,000	0,000	0,000	0,002	0,014	0,032	0,069	0,282	
13	0		0,254	0,055	0,024	0,010	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1		0,367	0,179	0,103	0,054	0,011	0,002	0,000	0,000	0,000	0,000	0,000
	2		0,245	0,268	0,206	0,139	0,045	0,010	0,001	0,000	0,000	0,000	0,000
	3		0,100	0,246	0,252	0,218	0,111	0,035	0,006	0,001	0,000	0,000	0,000
	4		0,028	0,154	0,210	0,234	0,184	0,087	0,024	0,003	0,001	0,000	0,000
	5		0,006	0,069	0,126	0,180	0,221	0,157	0,066	0,014	0,005	0,001	0,000
	6		0,001	0,023	0,056	0,103	0,197	0,209	0,131	0,044	0,019	0,006	0,000
	7		0,000	0,006	0,019	0,044	0,131	0,209	0,197	0,103	0,056	0,023	0,001
	8		0,000	0,001	0,005	0,014	0,066	0,157	0,221	0,180	0,126	0,069	0,006
	9		0,000	0,000	0,001	0,003	0,024	0,087	0,184	0,234	0,210	0,154	0,028
	10		0,000	0,000	0,000	0,001	0,006	0,035	0,111	0,218	0,252	0,246	0,100
	11		0,000	0,000	0,000	0,000	0,001	0,010	0,045	0,139	0,206	0,268	0,245
	12		0,000	0,000	0,000	0,000	0,000	0,002	0,011	0,054	0,103	0,179	0,367
	13		0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,010	0,024	0,055	0,254
14	0		0,229	0,044	0,018	0,007	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1		0,356	0,154	0,083	0,041	0,007	0,001	0,000	0,000	0,000	0,000	0,000
	2		0,257	0,250	0,180	0,113	0,032	0,006	0,001	0,000	0,000	0,000	0,000
	3		0,114	0,250	0,240	0,194	0,085	0,022	0,003	0,000	0,000	0,000	0,000
	4		0,035	0,172	0,220	0,229	0,155	0,061	0,014	0,001	0,000	0,000	0,000
	5		0,008	0,086	0,147	0,196	0,207	0,122	0,041	0,007	0,002	0,000	0,000
	6		0,001	0,032	0,073	0,126	0,207	0,183	0,092	0,023	0,008	0,002	0,000
	7		0,000	0,009	0,028	0,062	0,157	0,209	0,157	0,062	0,028	0,009	0,000
	8		0,000	0,002	0,008	0,023	0,092	0,183	0,207	0,126	0,073	0,032	0,001
	9		0,000	0,000	0,002	0,007	0,041	0,122	0,207	0,196	0,147	0,086	0,008
	10		0,000	0,000	0,000	0,001	0,014	0,061	0,155	0,229	0,220	0,172	0,035
	11		0,000	0,000	0,000	0,000	0,003	0,022	0,085	0,194	0,240	0,250	0,114
	12		0,000	0,000	0,000	0,000	0,001	0,006	0,032	0,113	0,180	0,250	0,257
	13		0,000	0,000	0,000	0,000	0,000	0,001	0,007	0,041	0,083	0,154	0,356
	14		0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,007	0,018	0,044	0,229

(continua)

# Tabela A-2 (continuação)

Probabilidades binomiais:			$p$										
$n$	$x$		$0,1$	$0,2$	$0,25$	$0,3$	$0,4$	$0,5$	$0,6$	$0,7$	$0,75$	$0,8$	$0,9$
$\binom{n}{x} p^x(1-p)^{n-x}$													
15	0		0,206	0,035	0,013	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1		0,343	0,132	0,067	0,031	0,005	0,000	0,000	0,000	0,000	0,000	0,000
	2		0,267	0,231	0,156	0,092	0,022	0,003	0,000	0,000	0,000	0,000	0,000
	3		0,129	0,250	0,225	0,170	0,063	0,014	0,002	0,000	0,000	0,000	0,000
	4		0,043	0,188	0,225	0,219	0,127	0,042	0,007	0,001	0,000	0,000	0,000
	5		0,010	0,103	0,165	0,206	0,186	0,092	0,024	0,003	0,001	0,000	0,000
	6		0,002	0,043	0,092	0,147	0,207	0,153	0,061	0,012	0,003	0,001	0,000
	7		0,000	0,014	0,039	0,081	0,177	0,196	0,118	0,035	0,013	0,003	0,000
	8		0,000	0,003	0,013	0,035	0,118	0,196	0,177	0,081	0,039	0,014	0,000
	9		0,000	0,001	0,003	0,012	0,061	0,153	0,207	0,147	0,092	0,043	0,002
	10		0,000	0,000	0,001	0,003	0,024	0,092	0,186	0,206	0,165	0,103	0,010
	11		0,000	0,000	0,000	0,001	0,007	0,042	0,127	0,219	0,225	0,188	0,043
	12		0,000	0,000	0,000	0,000	0,002	0,014	0,063	0,170	0,225	0,250	0,129
	13		0,000	0,000	0,000	0,000	0,000	0,003	0,022	0,092	0,156	0,231	0,267
	14		0,000	0,000	0,000	0,000	0,000	0,000	0,005	0,031	0,067	0,132	0,343
15		0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,005	0,013	0,035	0,206	
20	0		0,122	0,012	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1		0,270	0,058	0,021	0,007	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2		0,285	0,137	0,067	0,028	0,003	0,000	0,000	0,000	0,000	0,000	0,000
	3		0,190	0,205	0,134	0,072	0,012	0,001	0,000	0,000	0,000	0,000	0,000
	4		0,090	0,218	0,190	0,130	0,035	0,005	0,000	0,000	0,000	0,000	0,000
	5		0,032	0,175	0,202	0,179	0,075	0,015	0,001	0,000	0,000	0,000	0,000
	6		0,009	0,109	0,169	0,192	0,124	0,037	0,005	0,000	0,000	0,000	0,000
	7		0,002	0,055	0,112	0,164	0,166	0,074	0,015	0,001	0,000	0,000	0,000
	8		0,000	0,022	0,061	0,114	0,180	0,120	0,035	0,004	0,001	0,000	0,000
	9		0,000	0,007	0,027	0,065	0,160	0,160	0,071	0,012	0,003	0,000	0,000
	10		0,000	0,002	0,010	0,031	0,117	0,176	0,117	0,031	0,010	0,002	0,000
	11		0,000	0,000	0,003	0,012	0,071	0,160	0,160	0,065	0,027	0,007	0,007
	12		0,000	0,000	0,001	0,004	0,035	0,120	0,180	0,114	0,061	0,022	0,000
	13		0,000	0,000	0,000	0,001	0,015	0,074	0,166	0,164	0,112	0,055	0,002
	14		0,000	0,000	0,000	0,000	0,005	0,037	0,124	0,192	0,169	0,109	0,009
	15		0,000	0,000	0,000	0,000	0,001	0,015	0,075	0,179	0,202	0,175	0,032
	16		0,000	0,000	0,000	0,000	0,000	0,005	0,035	0,130	0,190	0,218	0,090
	17		0,000	0,000	0,000	0,000	0,000	0,001	0,012	0,072	0,134	0,205	0,190
	18		0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,028	0,067	0,137	0,285
	19		0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,007	0,021	0,058	0,270
	20		0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,003	0,012	0,122

(continua)



# *Tabela do Qui-quadrado*

A Tabela A-3 mostra as probabilidades da cauda direita da distribuição do Qui-quadrado (você pode usar o Capítulo 14 como referência para o teste do Qui-quadrado). Para usar a Tabela A-3, você precisará de três informações retiradas do problema com o qual estiver lidando:

- ✓ O tamanho amostral,  $n$ .
- ✓ O valor do Qui-quadrado para o qual deseja a probabilidade da cauda direita.
- ✓ Se estiver trabalhando com uma tabela de dupla entrada, precisará de  $l$  = número de linhas e  $c$  = número de colunas. Se você estiver trabalhando com o teste de qualidade de ajuste, precisará de  $k - 1$ , onde  $k$  é o número de categorias.

Os graus de liberdade para a estatística de teste do Qui-quadrado é  $(l - 1) * (c - 1)$ , caso esteja testando uma associação entre duas variáveis, onde  $l$  e  $c$  são o número de linhas e colunas em uma tabela de dupla entrada, respectivamente. Ou os graus de liberdade são calculados por  $k - 1$  em um teste de qualidade de ajuste, onde  $k$  é o número de categorias, veja o Capítulo 15.

Acompanhe a linha para os graus de liberdade até encontrar o valor mais próximo da estatística de teste do Qui-quadrado. Quando o encontrar, consulte a parte superior da coluna. Esse valor será a área à direita (além) da estatística Qui-quadrado em questão.

**Tabela A-3****A Tabela do Qui-quadrado**

Os números na tabela representam os valores do Qui-quadrado cuja área à direita equivale a  $p$ .

$gl/p$	0,10	0,05	0,025	0,01	0,005
1	2,71	3,84	5,02	6,64	7,88
2	4,61	5,99	7,38	9,21	10,60
3	6,25	7,82	9,35	11,35	12,84
4	7,78	9,49	11,14	13,28	14,86
5	9,24	11,07	12,83	15,09	16,75
6	10,65	12,59	14,45	16,81	18,55
7	12,02	14,07	16,01	18,48	20,28
8	13,36	15,51	17,54	20,09	21,96
9	14,68	16,92	19,02	21,67	23,59
10	15,99	18,31	20,48	23,21	25,19
11	17,28	19,68	21,92	24,73	26,76
12	18,55	21,03	23,34	26,22	28,30
13	19,81	22,36	24,74	27,69	29,819
14	21,06	23,69	26,12	29,14	31,32
15	22,31	25,00	27,49	30,58	32,80
16	23,54	26,30	28,85	32,00	34,27
17	24,77	27,59	30,19	33,41	35,72
18	25,99	28,87	31,53	34,81	37,16
19	27,20	30,14	32,85	36,19	38,58
20	28,41	31,41	34,17	37,57	40,00
21	29,62	32,67	35,48	38,93	41,40
22	30,81	33,92	36,78	40,29	42,80
23	32,01	35,17	38,08	41,64	44,18
24	33,20	36,42	39,36	42,98	45,56
25	34,38	37,65	40,65	44,31	46,93
26	35,56	38,89	41,92	45,64	48,29
27	36,74	40,11	43,20	46,96	49,65
28	37,92	41,34	44,46	48,28	50,99
29	39,09	42,56	45,72	49,59	52,34
30	40,26	43,77	46,98	50,89	53,67
40	51,81	55,76	59,34	63,69	66,77
50	63,17	67,51	71,42	76,15	79,49





# Tabela da Soma de Postos

As Tabelas A-4 (a) e A-4 (b) mostram os valores críticos para o teste da soma de postos para  $\alpha = 0,05$  e  $\alpha = 0,10$ , respectivamente; consulte o Capítulo 18 para obter mais informações sobre este teste. Para usar a Tabela A-4, você precisará de duas informações retiradas do problema com o qual estiver lidando:

- ✓ A estatística da soma de postos,  $T$ .
- ✓ Os tamanhos amostrais das duas amostras,  $n_1$  e  $n_2$ .

Para encontrar o valor crítico para sua estatística da soma de postos utilizando a Tabela A-4, vá à coluna que representa  $n_1$  e a linha que representa  $n_2$ . Cruze a linha e a coluna para encontrar os valores críticos inferiores e superiores (indicados por  $VCI$  e  $VCS$ ) para o teste da soma de postos.

Tabela A-4(a)			Tabela da Soma de Postos ( $\alpha = 0,05$ )															
$n_1 \backslash n_2$			3		4		5		6		7		8		9		10	
	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33		
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44		
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56		
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70		
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83		
8	8	28	14	38	21	49	29	61	39	73	49	87	53	93	54	98		
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114		
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131		

Tabela A-4(b)

Tabela da Soma de Postos ( $\alpha = 0,10$ )

<div><div><div><div><math>n_1</math></div><div><math>n_2</math></div></div></div><div></div></div>	3		4		5		6		7		8		9		10	
	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS	VCI	VCS
3	6	15	7	17	7	20	8	22	9	24	9	27	10	29	11	31
4	7	17	12	24	13	27	14	30	15	33	16	36	17	39	18	42
5	7	20	13	37	19	36	20	40	22	43	24	46	25	50	26	54
6	8	22	14	30	20	40	28	50	30	54	32	58	33	63	35	67
7	9	24	15	33	22	43	30	54	39	66	41	71	43	76	46	80
8	9	27	16	36	24	46	32	58	41	71	52	84	54	90	57	95
9	10	29	17	39	25	50	33	63	43	76	54	90	66	105	69	111
10	11	31	18	42	26	54	35	67	46	80	57	95	69	111	83	127

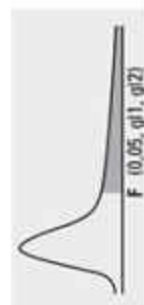
# Tabela- $F$

A Tabela A-5 mostra os valores críticos para a distribuição- $F$ , onde  $\alpha$  é igual a 0,05. (Os *valores críticos* são os valores que representam a fronteira entre rejeitar ou não rejeitar a  $H_0$ ; consulte o Capítulo 9.) Para usar a Tabela A-5, você precisará de três informações retiradas do problema com o qual está lidando:

- ✓ O tamanho amostral,  $n$ .
- ✓ O número de populações (ou tratamentos que estão sendo comparados),  $k$ .
- ✓ O valor de  $F$  para o qual deseja a probabilidade cumulativa.

Para encontrar o valor crítico para a estatística- $F$  utilizando a Tabela A-5, vá para a coluna que representa os graus de liberdade de que você precisa ( $k - 1$  e  $n - k$ ). Cruze a coluna de graus de liberdade ( $k - 1$ ) com a linha de graus de liberdade ( $n - k$ ) e encontre o valor crítico na distribuição- $F$ .

Tabela A-5

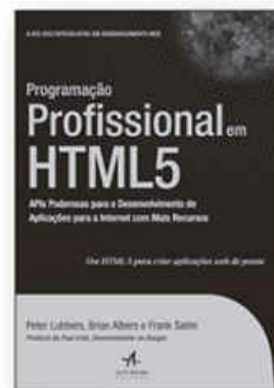
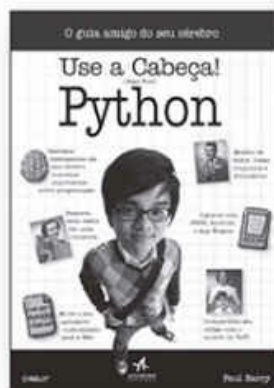
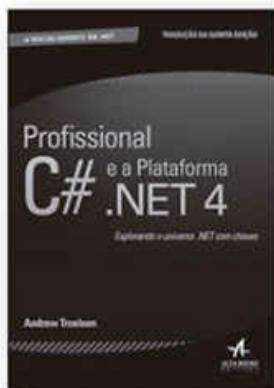
A Tabela-F ( $\alpha = 0,05$ )

q2/q1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161,4476	199,5000	215,7073	224,5832	230,1619	233,9860	236,7684	238,8827	240,5433	241,8817	243,9060	245,9499	248,0131	249,0518	250,0951	251,1432	252,1957	253,252
2	18,5128	19,0000	19,1543	19,2468	19,2964	19,3295	19,3532	19,3710	19,3848	19,3959	19,4125	19,4291	19,4458	19,4541	19,4624	19,4707	19,4791	19,487
3	10,1280	9,5521	9,2766	9,1172	9,0135	8,9406	8,867	8,8452	8,8123	8,7855	8,7446	8,7029	8,6602	8,6385	8,6166	8,5944	8,5720	8,549
4	7,7096	6,9443	6,5914	6,3882	6,2561	6,1631	6,0942	6,0410	5,9988	5,9644	5,9117	5,8578	5,8025	5,7744	5,7459	5,7170	5,6877	5,658
5	6,6079	5,7861	5,4095	5,1922	5,0503	4,9503	4,8759	4,8183	4,7725	4,7351	4,6777	4,6188	4,5581	4,5272	4,4957	4,4638	4,4314	4,398
6	5,9874	5,1433	4,7571	4,5337	4,3874	4,2839	4,2067	4,1468	4,0980	4,0600	3,9999	3,9381	3,8742	3,8415	3,8082	3,7743	3,7398	3,704
7	5,5914	4,7374	4,3468	4,1203	3,9715	3,8660	3,7870	3,7257	3,6767	3,6365	3,5747	3,5107	3,4445	3,4105	3,3758	3,3404	3,3043	3,267
8	5,3177	4,4590	4,0662	3,8379	3,6875	3,5806	3,5005	3,4381	3,3881	3,3472	3,2839	3,2184	3,1503	3,1152	3,0794	3,0428	3,0053	2,966
9	5,1174	4,2565	3,8625	3,6331	3,4817	3,3738	3,2927	3,2296	3,1789	3,1373	3,0729	3,0061	2,9365	2,9005	2,8637	2,8259	2,7872	2,747
10	4,9646	4,1028	3,7083	3,4780	3,3258	3,2172	3,1355	3,0717	3,0204	2,9782	2,9130	2,8450	2,7740	2,7372	2,6996	2,6609	2,6211	2,580
11	4,8443	3,9823	3,5874	3,3567	3,2039	3,0946	3,0123	2,9480	2,8962	2,8536	2,7876	2,7186	2,6464	2,6090	2,5705	2,5309	2,4901	2,448
12	4,7472	3,8853	3,4903	3,2592	3,1059	2,9961	2,9134	2,8486	2,7964	2,7534	2,6866	2,6169	2,5436	2,5055	2,4663	2,4259	2,3842	2,341
13	4,6672	3,8056	3,4105	3,1791	3,0254	2,9153	2,8321	2,7669	2,7144	2,6710	2,6037	2,5331	2,4589	2,4202	2,3803	2,3392	2,2966	2,252
14	4,6001	3,7389	3,3439	3,1122	2,9582	2,8477	2,7642	2,6987	2,6458	2,6022	2,5342	2,4630	2,3879	2,3487	2,3082	2,2664	2,2229	2,177
15	4,5431	3,6823	3,2874	3,0556	2,9013	2,7905	2,7066	2,6408	2,5876	2,5437	2,4753	2,4034	2,3275	2,2878	2,2468	2,2043	2,1601	2,114
16	4,4940	3,6337	3,2389	3,0069	2,8524	2,7413	2,6572	2,5911	2,5377	2,4935	2,4247	2,3522	2,2756	2,2354	2,1938	2,1507	2,1058	2,058
17	4,4513	3,5915	3,1968	2,9647	2,8100	2,6987	2,6143	2,5480	2,4943	2,4499	2,3807	2,3077	2,2304	2,1898	2,1477	2,1040	2,0584	2,010
18	4,4139	3,5546	3,1599	2,9277	2,7729	2,6613	2,5767	2,5102	2,4563	2,4117	2,3421	2,2686	2,1906	2,1497	2,1071	2,0629	2,0166	1,968
19	4,3807	3,5219	3,1274	2,8951	2,7401	2,6283	2,5435	2,4768	2,4227	2,3779	2,3080	2,2341	2,1555	2,1141	2,0712	2,0264	1,9795	1,930
20	4,3512	3,4928	3,0984	2,8661	2,7109	2,5990	2,5140	2,4471	2,3928	2,3479	2,2776	2,2033	2,1242	2,0825	2,0391	1,9938	1,9464	1,896
21	4,3248	3,4668	3,0725	2,8401	2,6848	2,5727	2,4876	2,4205	2,3660	2,3210	2,2504	2,1757	2,0960	2,0540	2,0102	1,9645	1,9165	1,865
22	4,3009	3,4434	3,0491	2,8167	2,6613	2,5491	2,4638	2,3965	2,3419	2,2967	2,2258	2,1508	2,0707	2,0283	1,9842	1,9380	1,8894	1,838
23	4,2793	3,4221	3,0280	2,7955	2,6400	2,5277	2,4422	2,3748	2,3201	2,2747	2,2036	2,1282	2,0476	2,0050	1,9605	1,9139	1,8648	1,812
24	4,2597	3,4028	3,0088	2,7763	2,6207	2,5082	2,4226	2,3551	2,3002	2,2547	2,1834	2,1077	2,0267	1,9838	1,9390	1,8920	1,8424	1,789
25	4,2417	3,3852	2,9912	2,7587	2,6030	2,4904	2,4047	2,3371	2,2821	2,2365	2,1649	2,0889	2,0075	1,9643	1,9192	1,8718	1,8217	1,768
26	4,2252	3,3690	2,9752	2,7426	2,5868	2,4741	2,3883	2,3205	2,2655	2,2197	2,1479	2,0716	1,9898	1,9464	1,9010	1,8533	1,8027	1,748
27	4,2100	3,3541	2,9604	2,7278	2,5719	2,4591	2,3732	2,3053	2,2501	2,2043	2,1323	2,0558	1,9736	1,9299	1,8842	1,8361	1,7851	1,730
28	4,1960	3,3404	2,9467	2,7141	2,5581	2,4453	2,3593	2,2913	2,2360	2,1900	2,1179	2,0411	1,9586	1,9147	1,8687	1,8203	1,7689	1,713
29	4,1830	3,3277	2,9340	2,7014	2,5454	2,4324	2,3463	2,2783	2,2229	2,1768	2,1045	2,0275	1,9446	1,9005	1,8543	1,8055	1,7537	1,698
30	4,1709	3,3158	2,9223	2,6896	2,5336	2,4205	2,3343	2,2662	2,2107	2,1646	2,0921	2,0148	1,9317	1,8874	1,8409	1,7918	1,7396	1,683
40	4,0847	3,2317	2,8387	2,6060	2,4495	2,3359	2,2490	2,1802	2,1240	2,0772	2,0035	1,9245	1,8389	1,7929	1,7444	1,6928	1,6373	1,576
60	4,0012	3,1504	2,7581	2,5252	2,3683	2,2541	2,1665	2,0970	2,0401	1,9926	1,9174	1,8364	1,7480	1,7001	1,6491	1,5943	1,5343	1,467
120	3,9201	3,0718	2,6802	2,4472	2,2899	2,1750	2,0868	2,0164	1,9588	1,9105	1,8337	1,7505	1,6587	1,6084	1,5543	1,4952	1,4290	1,351





# Conheça alguns de nossos outros livros sobre informática\_



Todas as imagens são meramente ilustrativas



ALTA BOOKS  
EDITORA

- Idiomas
- Culinária
- Informática
- Negócios
- Guias de Viagem
- Interesse Geral



Visite também nosso site para conhecer  
lançamentos e futuras publicações!

[www.altabooks.com.br](http://www.altabooks.com.br)



/alta\_books



/altabooks



## Seja autor da Alta Books

Todo o custo de produção fica por conta da editora e você ainda recebe direitos autorais pela venda no período de contrato.\*

Envie a sua proposta para [autoria@altabooks.com.br](mailto:autoria@altabooks.com.br) ou encaminhe o seu texto\*\* para:  
Rua Viúva Cláudio 291 - CEP: 20970-031 Rio de Janeiro

\*Caso o projeto seja aprovado pelo Conselho Editorial.

\*\*Qualquer material encaminhado à editora não será devolvido.





## O jeito mais fácil e divertido de melhorar suas habilidades em Estatística

Precisa expandir seus conhecimentos em Estatística e avançar para a Estatística II? Este amigável manual fornece o que você precisa para entender o que é regressão múltipla, análise de variância (ANOVA), testes do Qui-quadrado, procedimentos não paramétricos e outros tópicos relevantes. *Estatística II Para Leigos* também oferece várias dicas para a hora da prova, além de aplicações reais que vão fazer com que você tire de letra a análise de dados, tanto na sala de aula quanto no trabalho.

- **Começando com o básico** — revisão dos principais pontos de Estatística I, expandindo de regressão linear simples a intervalos de confiança e teste de hipótese
- **Começando a fazer previsões** — compreensão dos conceitos de regressão múltipla, não linear e logística; verificação das condições e interpretação dos resultados
- **Analisando a variância com ANOVA** — decomposição da tabela ANOVA, ANOVA com um e com dois fatores, teste-F e comparações múltiplas
- **Relacionando os testes do Qui-quadrado** — análise das tabelas de dupla entrada e teste dos dados categóricos para independência e qualidade de ajuste
- **Prosseguindo com os procedimentos não paramétricos** — técnicas usadas quando não se pode assumir que os dados possuem uma distribuição normal

**Deborah Rumsey** é especialista em ensino de Estatística e membro auxiliar do Departamento de Estatística da Ohio State University. Também é membro da Associação Americana de Estatística e recebeu o prêmio Presidential Teaching da Kansas State University. A Dra. Rumsey publicou inúmeros artigos e ministrou muitas palestras sobre o ensino da Estatística.



### Abra este livro e descubra:

- Métodos atualizados de análise de dados
- Explicações completas sobre os conceitos de Estatística II
- Instruções passo a passo concisas e objetivas
- Dissecção das ajudas do computador
- Muitas dicas, estratégias e alertas
- Os dez erros mais comuns nas conclusões estatísticas
- Aplicações da Estatística no dia a dia
- Tabelas para completar os cálculos usados no livro

Acesse o site

[www.paraleigos.com.br](http://www.paraleigos.com.br)  
e conheça outros títulos!

FOR  
DUMMIES



ALTA BOOKS  
EDITORA  
[www.altabooks.com.br](http://www.altabooks.com.br)

f /paraleigos

t /para\_leigos

ISBN 978-85-7608-636-9



9 788576 086369 >





# Índice

Estatística II Para Leigos	3
Rosto	9
Creditos	10
Dedicatória	13
Sumário Resumido	16
Sumário	18
Introdução	27
Sobre Este Livro	28
Convenções Usadas Neste Livro	30
Só de Passagem	31
Penso que	32
Como Este Livro Está Organizado	33
Parte I: Encarando os Fundamentos da Análise de Dados e da Construção de Modelos	33
Parte II: Usando Diferentes Tipos de Regressão para Fazer Previsões	33
Parte III: Analisando a Variância com ANOVA	33
Parte IV: Construindo Fortes Ligações com os Testes Qui-quadrado	33
Parte V: Estatística Não Paramétrica: Rebeldes sem Distribuição	34
Parte VI: A Parte dos Dez	34
Ícones Usados Neste Livro	35
De Lá para Cá, Daqui para Lá	36
Parte I: Encarando os Fundamentos da Análise de Dados e da Construção de Modelos	37
Capítulo 1: Além das Operações Numéricas: A Arte e a Ciência da Análise de Dados	40
Análise de Dados: Olhe Antes de Mastigar	41
Nada(nem mesmo uma reta) dura para sempre	42
Bisbilhotar os dados não é coisa que se faça!	42
Proibido pescar(dados)	43
Veja o Quadro como um Todo: Um Panorama sobre Estatística II	45
Parâmetro da população	45
Estatística amostral	45
Intervalo de confiança	46
Teste de hipótese	46
Análise de variância (ANOVA)	47
Comparações múltiplas	47

Efeitos de interação	48
Correlação	48
Regressão linear	49
Testes Qui-quadrados	50
Estatística não paramétrica	51
Capítulo 2: Encontre a Análise Certa para o Problema	52
Variáveis Categóricas versus Variáveis Quantitativas	53
Estatísticas para Variáveis Categóricas	55
Estimando uma proporção	55
Comparando proporções	55
Procurando relações entre variáveis categóricas	56
Construindo modelos para fazer previsões	57
Estatísticas para Variáveis Quantitativas	59
Fazendo estimativas	59
Fazendo comparações	59
Explorando relações	60
Prevendo y através de x	62
Evitando o Viés	63
Medindo a Precisão Através da Margem de Erro	65
Conhecendo Seus Limites	67
Capítulo 3: Revendo Intervalos de Confiança e Testes de Hipótese	68
Estimando Parâmetros Usando os Intervalos de Confiança	69
Entendendo o básico: A forma geral de um intervalo de confiança	69
Encontrando o intervalo de confiança para uma média populacional	70
O que altera a margem de erro?	71
Interpretando um intervalo de confiança	73
O que É que os Testes de Hipótese Têm?	75
O que $H_0$ e $H_a$ realmente representam?	75
Reunindo evidências em uma estatística de teste	75
Determinando a força da evidência através do valor-p	76
Alarmes falsos e oportunidades perdidas: Erros Tipo I e Tipo II	77
O poder de um teste de hipótese	78
Parte II: Usando Diferentes Tipos de Regressão para Fazer Previsões	83
Capítulo 4: Em Linha com a Regressão Linear Simples	86
Investigando Relações com Diagramas de Dispersão e Correlações	87
Usando diagramas de dispersão para investigar relações	88
Comparando informações através do coeficiente de correlação	89
Construindo um Modelo de Regressão Linear Simples	91

Encontrando a reta certa para modelar seus dados	91
O intercepto y da reta de regressão	92
O coeficiente angular da reta de regressão	93
Estimando pontos através da regressão linear	93
Sem Deixar Nenhuma Conclusão para Trás: Testes e Intervalos de Confiança para a Regressão	95
Analisando o coeficiente angular	95
Inspecionando o intercepto y	98
Construindo intervalos de confiança para a resposta média	99
Prevendo o futuro com os intervalos de previsão	100
Checando a Adequação do Modelo (dos Dados, Não das Roupas!)	103
Definindo as condições	103
Encontrando e investigando os resíduos	104
Usando $r^2$ para medir o ajuste do modelo	108
Analisando outliers	109
Conhecendo as Limitações de Sua Análise de Regressão	111
Evitando cair no modo causa e efeito	111
Extrapolação: N-A-O-Til, NUNCA!	112
Às vezes é preciso ter mais do que uma variável	112
Capítulo 5: Regressão Múltipla com Duas Variáveis X	114
Conhecendo o Modelo de Regressão Múltipla	115
Descobrimos os usos da regressão múltipla	115
A fórmula geral do modelo de regressão múltipla	115
Seguindo os passos rumo a uma análise	116
Observando x's e y's	117
Coletando Dados	118
Identificando Possíveis Relações	120
Construindo diagramas de dispersão	120
Correlações: Examinando os vínculos	121
Checando a Multicolinearidade	124
Encontrando o Modelo sob Medida para Duas Variáveis X	126
Obtendo os coeficientes de regressão múltipla	126
Interpretando os coeficientes	127
Testando os coeficientes	128
Prevendo y Através das Variáveis x	131
Verificando o Ajuste do Modelo de Regressão Múltipla	133
Observando as condições	133
Traçando um plano para checar as condições	133

Verificando as três condições	135
-------------------------------	-----

Capítulo 6: Como Vou Sentir Sua Falta se Você Não Sair? Escolha do Modelo de Regressão	138
--	-----

Dando o Pontapé Inicial na Estimativa para a Distância de um Punt	139
---	-----

Fazendo o brainstorm das variáveis e coletando os dados	139
---	-----

Examinando diagramas de dispersão e correlações	141
---	-----

Igual a Comprar Sapatos: O Modelo É Lindo, Mas Serve?	145
---	-----

Avaliando o ajuste do modelo de regressão múltipla	145
--	-----

Processo de seleção de modelo	146
-------------------------------	-----

Capítulo 7: Subindo na Curva de Aprendizagem com a Regressão Não Linear	151
---	-----

Antecipando a Regressão Não Linear	152
------------------------------------	-----

Começando com Diagramas de Dispersão	154
--------------------------------------	-----

Nas Curvas da Estrada com os Polinômios	156
---	-----

Relembrando o que é um polinômio	156
----------------------------------	-----

Em busca do melhor modelo polinomial	158
--------------------------------------	-----

Usando um polinômio de segundo grau para passar na prova	159
--	-----

Avaliando o ajuste de um modelo polinomial	162
--	-----

Fazendo previsões	165
-------------------	-----

Subiu? Desceu? Então É Exponencial!	167
-------------------------------------	-----

Recordando os modelos exponenciais	167
------------------------------------	-----

Em busca do melhor modelo exponencial	168
---------------------------------------	-----

Espalhando segredos de forma exponencial	170
--	-----

Capítulo 8: Sim, Não, Talvez: Fazendo Previsões Usando a Regressão Logística	173
--	-----

Entendendo o Modelo de Regressão Logística	174
--	-----

Qual é a diferença entre a regressão logística e as outras regressões?	174
--	-----

Utilizando uma curva em S para estimar as probabilidades	175
--	-----

Interpretando os coeficientes do modelo de regressão logística	176
--	-----

O modelo de regressão linear em ação	176
--------------------------------------	-----

Fazendo uma Análise de Regressão Logística	178
--	-----

Fazendo a análise no Minitab	178
------------------------------	-----

Encontrando os coeficientes e construindo o modelo	179
--	-----

Estimando p	181
-------------	-----

Verificando o ajuste do modelo	181
--------------------------------	-----

Ajustando o modelo	182
--------------------	-----

Parte III: Analisando a Variância com ANOVA	185
---	-----

Capítulo 9: Precisando Testar Várias Médias? Venha para a ANOVA!	188
--	-----

Comparando Duas Médias com um Teste-t	189
---------------------------------------	-----

Avaliando Mais Médias com ANOVA	191
---------------------------------	-----

Cuspe de sementes: Uma situação perfeita para a ANOVA	191
Seguindo os passos da ANOVA	192
Verificando as Condições	193
Verificando a independência	193
Procurando o que é normal	193
Notando a dispersão	194
Estabelecendo as Hipóteses	198
Realizando o Teste-F	199
ANOVA no Minitab	199
Desmembrando a variância em somas de quadrados	200
Localizando as médias das somas de quadrados	201
Chegando à estatística-F	202
Tirando conclusões a partir da ANOVA	203
O que fazer agora?	205
Verificando o Ajuste do Modelo ANOVA	206
Capítulo 10: Organizando as Médias Através das Comparações Múltiplas	208
Acompanhando a ANOVA	209
Comparando o uso de minutos no celular: Um exemplo	209
Preparando o terreno para os procedimentos de comparação múltipla	211
Identificando as Médias Diferentes com Fisher e Tukey	213
Pescando diferenças com o LSD de Fisher	213
Usando o novo e aperfeiçoado LSD de Fisher	214
O teste de Tukey	216
Examinando a Saída para Determinar a Análise	218
Tantos Outros Procedimentos, Tão Pouco Tempo!	219
Cortando a conversa fiada com o ajuste de Bonferroni	219
Comparando combinações usando o método de Scheffe	220
O teste de Dunnett	220
O teste de Student Newman-Keuls	221
O teste de Duncan	221
Ficando não paramétrico com o teste de Kruskal-Wallis	222
Capítulo 11: Percorrendo os Caminhos da ANOVA com Dois Fatores	224
Configurando o Modelo ANOVA com Dois Fatores	225
Determinando os tratamentos	225
Em busca das somas de quadrados	225
Entendendo os Efeitos da Interação	228
Mas, afinal, o que é interação?	228
Interagindo com os gráficos de interação	229

Testando os Termos na ANOVA com Dois Fatores	232
Executando uma Tabela ANOVA	233
Interpretando os resultados: Números e gráficos	233
O Branco Fica Mais Branco na Água Quente? Mais um Caso para a ANOVA com Dois Fatores	236

## Capítulo 12: Regressão e ANOVA: Uma Relação Inesperada! 240

Vendo a Regressão Através dos Olhos da Variação	241
Localizando a variabilidade e encontrando uma “x-plicação”	241
Chegando aos resultados com a regressão	242
Avaliando o ajuste do modelo de regressão	243
Regressão e ANOVA: O Encontro dos Modelos	245
Comparando as somas de quadrados	245
Dividindo os graus de liberdade	247
Levando a regressão até a tabela ANOVA	248
Relacionando as estatísticas F e t: a última fronteira	249

## Parte IV: Construindo Fortes Ligações com os Testes Qui-quadrado 251

### Capítulo 13: Fazendo Associações com Tabelas de Dupla Entrada 254

Decompondo uma Tabela de Dupla Entrada	256
Organizando dados em uma tabela de dupla entrada	256
Preenchendo as células	256
Totais marginais	257
Desmembrando as Probabilidades	259
Probabilidades marginais	259
Probabilidades conjuntas	260
Probabilidades condicionais	261
Tentando Ser Independente	267
Verificando a independência entre duas categorias	267
Verificando a independência entre duas variáveis	268
Desmistificando o Paradoxo de Simpson	270
Experimentando o Paradoxo de Simpson	270
Descobrimo o porquê do Paradoxo de Simpson	272
De olho no Paradoxo de Simpson	274

### Capítulo 14: Independente o Suficiente para o Teste do Qui-quadrado 275

O Teste do Qui-quadrado para a Independência	276
Coletando e organizando os dados	277
Determinando as hipóteses	278
Calculando as frequências esperadas	279
Verificando as condições para o teste	280

Calculando a estatística Qui-quadrado	280
Encontrando seus resultados na tabela do Qui-quadrado	283
Tirando conclusões	286
Colocando o Qui-quadrado à prova	288
Comparando Dois Testes para Comparar Duas Proporções	290
Refamiliarizando-se com o teste-Z para duas proporções populacionais	290
Igualando os testes do Qui-quadrado e testes-Z para uma tabela dois por dois	291
Capítulo 15: Usando os Testes do Qui-quadrado para Qualidade de Ajuste (dos Dados, e Não de Seu Jeans)	295
Encontrando a Estatística de Qualidade de Ajuste	296
O observado versus o esperado	296
Calculando a estatística de qualidade de ajuste	298
Interpretando a Estatística da Qualidade de Ajuste Através do Qui-quadrado	301
Verificando as condições antes de começar	302
Os passos para o teste Qui-quadrado de qualidade de ajuste	302
<b>Parte V: Estatística Não Paramétrica: Rebeldes sem Distribuição</b>	<b>306</b>
Capítulo 16: Ficando Não Paramétrico	309
Em Favor da Estatística Não Paramétrica	310
Não precisa se preocupar se as condições não forem atendidas	310
Uma chance para a mediana mostrar seu potencial	311
Então, qual é a pegadinha?	313
Dominando o Básico das Estatísticas Não Paramétricas	314
Sinal	314
Postos	316
Postos com sinais	317
Soma de postos	318
Capítulo 17: Todos os Sinais Apontam para o Teste dos Sinais e o Teste de Postos Sinalizados	320
Interpretando os Sinais: O Teste dos Sinais	321
Testando a mediana	322
Estimando a mediana	325
Testando os pares combinados	327
Um Passo Adiante com o Teste de Postos Sinalizados	330
Uma limitação do teste dos sinais	330
Seguindo os passos para realizar um teste de postos sinalizados	330
Emagrecendo com os postos sinalizados	331
Capítulo 18: Subindo de Posto com o Teste das Somas dos Postos	335
Realizando o Teste da Soma dos Postos	336



Verificando as condições	336
Seguindo os passos para a realização de um teste	336
Aumentando o tamanho da amostra	338
Realizando um Teste da Soma dos Postos: Qual Corretor de Imóveis Vende Casas Mais Rápido?	340
Verificando as condições para este teste	340
Testando a hipótese	341
Capítulo 19: Faça o Kruskal-Wallis e Ordene as Somas com Wilcoxon	346
Fazendo o Teste de Kruskal-Wallis para Comparar Mais de Duas Populações	347
Verificando as condições	348
Estabelecendo o teste	350
Realizando o teste passo a passo	350
Localizando as Diferenças: O Teste da Soma dos Postos de Wilcoxon	354
Comparações pareadas	354
Realizando testes de comparação para ver quem é diferente	355
Examinando as medianas para ver como elas se diferem	356
Capítulo 20: Apontando Correlações com o Posto de Spearman	358
Pearson e Suas Preciosas Condições	359
Correlação de Posto de Spearman	361
Calculando a correlação de posto de Spearman	361
Spearman em ação: Relacionando aptidão ao desempenho	362
Parte VI: A Parte dos Dez	366
Capítulo 21: Os Dez Erros Mais Comuns nas Conclusões Estatísticas	369
Dizer o que as Estatísticas Provam	370
Tecnicamente Não É Estatisticamente Significativo, Mas	371
Concluir que x Causa y	372
Supor que os Dados São Normais	373
Relatar Apenas os Resultados “Importantes”	374
Supor que uma Amostra Grande É Sempre Melhor	375
Não É Tecnicamente Aleatória, Mas	377
Supor que 1.000 Respostas São 1.000 Respostas	378
Naturalmente, os Resultados se Aplicam à População em Geral	380
Omitir	381
Capítulo 22: Dez Formas de Chegar na Frente por Saber Estatística	383
Faça as Perguntas Certas	384
Seja Cético	385
Colete e Analise os Dados Corretamente	386
Pedindo Ajuda	387

Refazendo os Passos de Outras Pessoas	388
Juntando as Peças	389
Verificando Suas Respostas	390
Explicando a Saída	391
Fazendo Recomendações Convincentes	392
Estabelecendo-se como o Cara da Estatística	394

Capítulo 23: Dez Empregos Legais que Usam Estatística	395
Pesquisadores de Opinião Pública	396
Ornitólogo (Observador de Pássaros)	397
Comentarista ou Jornalista Esportivo	398
Jornalista	400
Combatentes do Crime	401
Profissional da Área Médica	402
Executivo de Marketing	403
Advogado	404
Corretor de Ações	405

Apêndice: Tabelas de Referência	406
Tabela-t	407
Tabela Binomial	409
Tabela do Qui-quadrado	414
Tabela da Soma de Postos	417
Tabela-F	419