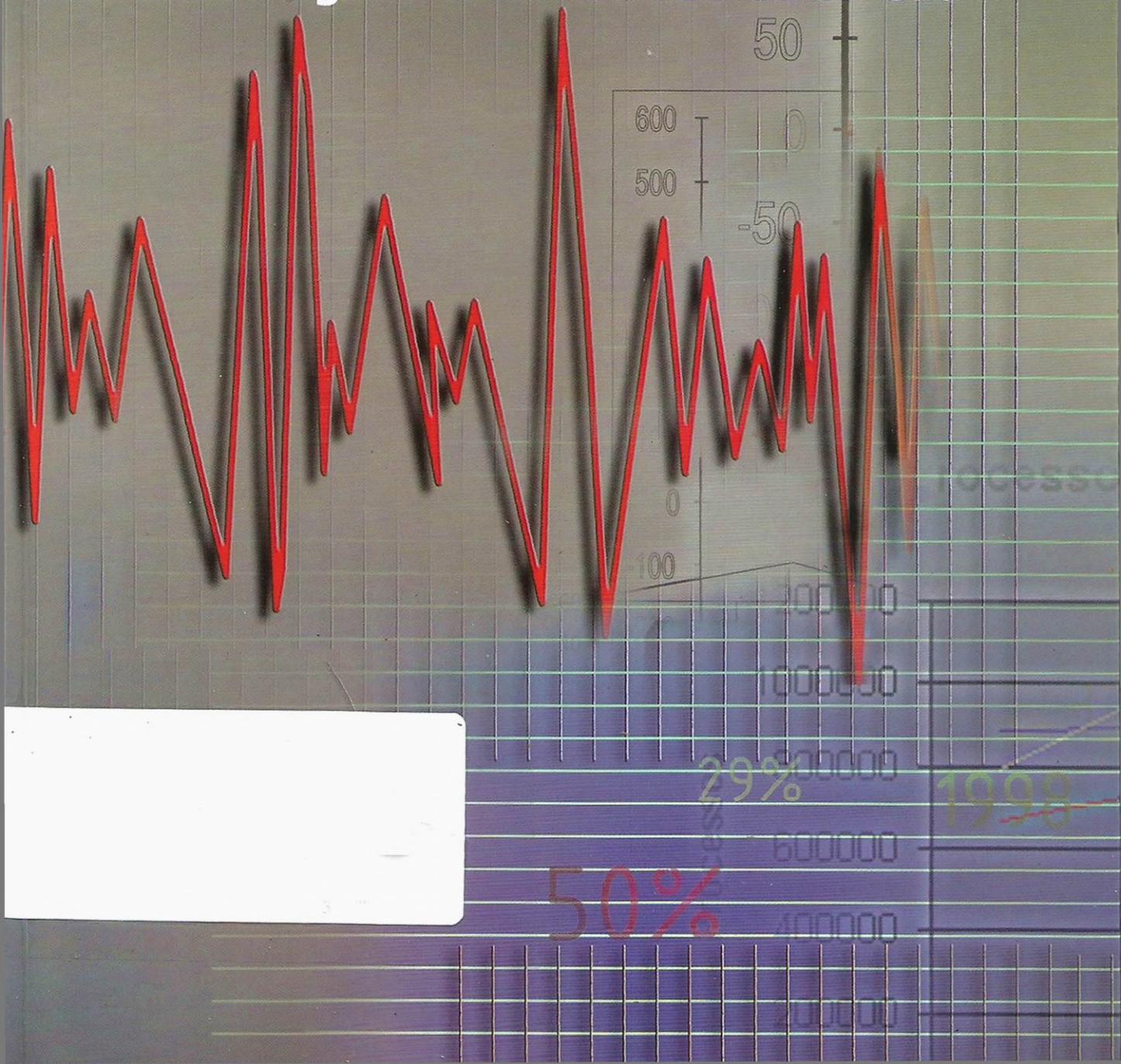


THOMSON

principios de

Bio estatística

Marcello Pagano & Kimberlee Gauvreau



Sumário

Prefácio	XIII
1 Introdução	1
1.1 Resumo do Texto.....	2
1.2 Exercícios de Revisão	5
Bibliografia.....	5
2 Apresentação de Dados	6
2.1 Tipos de Dados Numéricos	6
2.1.1 Dados Nominais.....	6
2.1.2 Dados Ordinais	8
2.1.3 Dados Substituídos por Postos.....	8
2.1.4 Dados Discretos	9
2.1.5 Dados Contínuos.....	10
2.2 Tabelas.....	10
2.2.1 Distribuições de Freqüências	10
2.2.2 Freqüência Relativa	12
2.3 Gráficos	14
2.3.1 Gráficos de Barras	14
2.3.2 Histogramas	15
2.3.3 Polígonos de Freqüência.....	16
2.3.4 Gráficos de Dispersão Unidimensionais.....	19
2.3.5 Box Plots.....	19
2.3.6 Gráficos de Dispersão Bidimensionais	20
2.3.7 Gráficos de Linha:.....	21
2.4 Aplicações Adicionais.....	22
2.5 Exercícios de Revisão	28
Bibliografia.....	33
3 Medidas-Resumo Numéricas	35
3.1 Medidas de Tendência Central	35
3.1.1 Média	35
3.1.2 Mediana	37
3.1.3 Moda	38

3.2 Medidas de Dispersão	39
3.2.1 Amplitude	39
3.2.2 Intervalo Interquartil	40
3.2.3 Variância e Desvio-Padrão	42
3.2.4 Coeficiente de Variação	44
3.3 Dados Agrupados	44
3.3.1 Média de Dados Agrupados	45
3.3.2 Variância de Dados Agrupados	47
3.4 Desigualdade de Chebychev	47
3.5 Aplicações Adicionais	49
3.6 Exercícios de Revisão	54
Bibliografia.....	59
4 Taxas e Padronização	60
4.1 Taxas	60
4.2 Padronização de Taxas	64
4.2.1 Método Direto de Padronização	66
4.2.2 Método Indireto de Padronização	67
4.2.3 Uso de Taxas Padronizadas.....	68
4.3 Aplicações Adicionais.....	77
4.3.1 Método Direto de Padronização	78
4.3.2 Método Indireto de Padronização	80
4.4 Exercícios de Revisão	81
Bibliografia.....	87
5 Tábuas de Vida.....	88
5.1 Cálculo da Tábua de Vida	88
5.1.1 Coluna 1	88
5.1.2 Coluna 2	90
5.1.3 Colunas 3 e 4	91
5.1.4 Coluna 5	93
5.1.5 Coluna 6	93
5.1.6 Coluna 7	93
5.2 Aplicações da Tábua de Vida	94
5.3 Anos Potenciais de Vida Perdidos	96
5.4 Aplicações Adicionais.....	100
5.5 Exercícios de Revisão	104
Bibliografia.....	112
6 Probabilidade	113
6.1 Operações sobre Eventos e Probabilidade	113
6.2 Probabilidade Condicional	117
6.3 Teorema de Bayes	118
6.4 Testes de Diagnósticos	123
6.4.1 Sensibilidade e Especificidade.....	123
6.4.2 Aplicações do Teorema de Bayes	124

6.4.3 Curvas ROC	127
6.4.4 Cálculos de Prevalência	129
6.5 O Risco Relativo e a Razão de Chances	131
6.6 Aplicações Adicionais	136
6.7 Exercícios de Revisão	141
Bibliografia.....	145
7 Distribuições Teóricas de Probabilidade.....	147
7.1 Distribuições de Probabilidade	147
7.2 A Distribuição Binomial	149
7.3 A Distribuição de Poisson	155
7.4 A Distribuição Normal	159
7.5 Aplicações Adicionais	167
7.6 Exercícios de Revisão	172
Bibliografia.....	175
8 Distribuição Amostral da Média	177
8.1 Distribuições Amostrais	177
8.2 O Teorema Central do Limite	178
8.3 Aplicações do Teorema Central do Limite.....	179
8.4 Aplicações Adicionais	184
8.5 Exercícios de Revisão	190
Bibliografia.....	192
9 Intervalos de Confiança	193
9.1 Intervalos de Confiança Bilaterais	193
9.2 Intervalos de Confiança Unilaterais	198
9.3 Distribuição t de Student	199
9.4 Aplicações Adicionais	202
9.5 Exercícios de Revisão	205
Bibliografia.....	207
10 Testes de Hipóteses	209
10.1 Conceitos Gerais	209
10.2 Testes de Hipóteses Bilaterais.....	211
10.3 Testes de Hipóteses Unilaterais	214
10.4 Tipos de Erro	215
10.5 Poder	218
10.6 Estimação do Tamanho da Amostra.....	221
10.7 Aplicações Adicionais	223
10.8 Exercícios de Revisão	228
Bibliografia.....	230

11 Comparação de Duas Médias	232
11.1 Amostras Pareadas	233
11.2 Amostras Independentes	237
11.2.1 Variâncias Iguais	238
11.2.2 Variâncias Desiguais	242
11.3 Aplicações Adicionais	244
11.4 Exercícios de Revisão	248
Bibliografia	252
12 Análise de Variância	254
12.1 Análise de Variância com um Fator	254
12.1.1 O Problema	254
12.1.2 Fontes de Variação	257
12.2 Procedimentos de Comparações Múltiplas	260
12.3 Aplicações Adicionais	262
12.4 Exercícios de Revisão	266
Bibliografia	268
13 Métodos Não-paramétricos	269
13.1 O Teste do Sinal	269
13.2 O Teste de Postos Sinalizados de Wilcoxon	271
13.3 O Teste da Soma de Postos de Wilcoxon	274
13.4 Vantagens e Desvantagens dos Métodos Não-paramétricos	277
13.5 Aplicações Adicionais	278
13.6 Exercícios de Revisão	282
Bibliografia	286
14 Inferência sobre Proporções	287
14.1 Aproximação Normal para a Distribuição Binomial	287
14.2 Distribuição Amostral de uma Proporção	289
14.3 Intervalos de Confiança	290
14.4 Testes de Hipóteses	292
14.5 Estimação do Tamanho da Amostra	293
14.6 Comparação de Duas Proporções	294
14.7 Aplicações Adicionais	297
14.8 Exercícios de Revisão	299
Bibliografia	302
15 Tabelas de Contingência	304
15.1 O Teste Qui-Quadrado	304
15.1.1 Tabelas 2×2	304
15.1.2 Tabelas $r \times c$	309
15.2 Teste de McNemar	310
15.3 A Razão de Chances	312
15.4 Falácia de Berkson	317
15.5 Aplicações Adicionais	319
15.6 Exercícios de Revisão	324
Bibliografia	330

16 Tabelas de Contingência 2 X 2 Múltiplas.....	332
16.1 Paradoxo de Simpson.....	332
16.2 O Método de Mantel-Haenszel	333
16.2.1 Teste de Homogeneidade	335
16.2.2 Razão de Chances Resumo.....	338
16.2.3 Teste de Associação	341
16.3 Aplicações Adicionais.....	343
16.4 Exercícios de Revisão.....	348
Bibliografia.....	351
17 Correlação	352
17.1 O Gráfico de Dispersão Bidimensional	352
17.2 Coeficiente de Correlação de Pearson	354
17.3 Coeficiente de Correlação de Postos de Spearman.....	357
17.4 Aplicações Adicionais.....	360
17.5 Exercícios de Revisão.....	364
Bibliografia.....	366
18 Regressão Linear Simples	367
18.1 Conceitos da Regressão	367
18.2 O Modelo	371
18.2.1 A Linha de Regressão da População	371
18.2.2 O Método dos Mínimos Quadrados.....	373
18.2.3 Inferência para os Coeficientes da Regressão.....	376
18.2.4 Inferência para Valores Previstos.....	379
18.3 Avaliação do Modelo	381
18.3.1 O Coeficiente de Determinação	381
18.3.2 Gráficos de Resíduos	382
18.3.3 Transformações.....	384
18.4 Aplicações Adicionais.....	386
18.5 Exercícios de Revisão.....	391
Bibliografia.....	395
19 Regressão Múltipla	396
19.1 O Modelo	396
19.1.1 A Equação da Regressão de Mínimos Quadrados	397
19.1.2 Inferência para os Coeficientes da Regressão.....	398
19.1.3 Avaliação do Modelo	400
19.1.4 Variáveis Indicadoras	401
19.1.5 Termos de Interação	403
19.2 Seleção do Modelo.....	404
19.3 Aplicações Adicionais.....	406
19.4 Exercícios de Revisão	410
Bibliografia.....	414

20 Regressão Logística	415
20.1 O Modelo	415
20.1.1 A Função Logística	416
20.1.2 A Equação Ajustada	418
20.2 Regressão Logística Múltipla	419
20.3 Variáveis Indicadoras	422
20.4 Aplicações Adicionais.....	424
20.5 Exercícios de Revisão	427
Bibliografia.....	430
21 Análise de Sobrevida	431
21.1 O Método da Tábua de Vida	432
21.2 O Método do Produto-Limite	437
21.3 O Teste Log-Rank	440
21.4 Aplicações Adicionais.....	444
21.5 Exercícios de Revisão	451
Bibliografia.....	453
22 Teoria da Amostragem	454
22.1 Esquemas de Amostragem	454
22.1.1 Amostragem Aleatória Simples	455
22.1.2 Amostragem Sistemática.....	455
22.1.3 Amostragem Estratificada.....	456
22.1.4 Amostragem por Conglomerados	457
22.1.5 Amostragem Não-Probabilística	457
22.2 Fontes de Tendência.....	457
22.3 Aplicações Adicionais.....	459
22.4 Exercícios de Revisão	463
Bibliografia.....	464
Apêndice A <i>Tabelas</i>	465
Apêndice B	491
Índice	503

2

Apresentação de Dados

Todo estudo ou experimento produz um conjunto de dados cujo tamanho pode variar desde poucas medidas até muitos milhares de observações. Um conjunto completo de dados, no entanto, não proverá, necessariamente, um investigador com informações que possam ser facilmente interpretadas. Por exemplo: a Tabela 2.1 relaciona por linha os primeiros 2.560 casos de síndrome de imunodeficiência adquirida (Aids) registrados nos Centros de Controle e Prevenção de Doenças [1]. Cada indivíduo foi classificado como portador ou paciente de sarcoma de Kaposi, designado por 1, ou como não sofrendo da doença, representado por 0 (o sarcoma de Kaposi é um tumor que afeta a pele, as membranas mucosas e os nós linfáticos). Embora a Tabela 2.1 exiba o conjunto inteiro de resultados, é extremamente difícil caracterizar-se os dados. Não podemos sequer identificar as proporções relativas de 0s e 1s. Entre os dados brutos e os resultados reportados do estudo, existe manipulação inteligente e imaginativa dos números, realizada por meio de métodos de estatísticas descritivas.

Estatísticas descritivas são um meio de se organizar e resumir as observações. Elas nos provêm com um resumo das características gerais de um conjunto de dados e podem assumir várias formas, entre as quais as tabelas, os gráficos e as medidas-resumo numéricas. Neste capítulo, discutiremos os vários métodos de se exibir um conjunto de dados. No entanto, antes que decidamos qual técnica é a mais apropriada em determinada situação, precisamos primeiramente determinar que tipo de dados temos.

2.1 Tipos de Dados Numéricos

2.1.1 Dados Nominais

No estudo da bioestatística, encontramos muitos tipos diferentes de dados numéricos que têm variados graus de estrutura na relação entre os valores possíveis. Um dos mais simples são os dados nominais, nos quais os valores são classificados em categorias ou classes não-ordenadas. Tal como na Tabela 2.1, os números são usados freqüentemente para representar as categorias. Em certo estudo, por exemplo, os homens podem ser assinalados com o valor 1 e as mulheres com o valor 0.

TABELA 2.1

Resultados indicando se um indivíduo teve o sarcoma de Kaposi para os primeiros 2.560 pacientes de Aids, registrados pelos Centros de Controle e Prevenção de Doenças em Atlanta, Geórgia.

00000000	00010100	00000010	00001000	00000001	00000000	10000000	00000000
00101000	00000000	00000000	00011000	00100001	01001100	00000000	00000010
00000001	00000000	00000010	01100000	00000000	00000100	00000000	00000000
00100010	00100000	00000101	00000000	00000000	00000001	00001001	00000000
00000000	00010000	00010000	00010000	00000000	00000000	00000000	00000000
00000000	00000000	00000000	00001000	00000000	00010000	10000000	00000000
00100000	00000000	00001000	00000010	00000000	00000100	00000000	00010000
00000000	00000000	00000000	00001000	00000000	00000100	00000000	01000000
00010000	00000000	00010000	01000000	00000000	00000000	00000101	00100000
00000000	00000000	00000100	00000000	01000100	00000000	00000001	10100000
00000100	00000000	00010000	00000000	00001000	00000000	00000010	00100000
00000000	00000000	00000000	10001000	00001000	00000000	01000000	00000000
00000000	00001100	00000000	00000000	10000011	00000001	11000000	00001000
00000000	00000000	00000000	00000000	01000000	00000001	00010001	00000000
10000000	00000000	01000000	00000000	00000000	01010100	00000000	00010100
00000000	00000000	00000000	00001010	00000101	00000000	00000000	00010000
00000000	00000000	00000000	00000001	00000100	00000000	00000000	00001000
11000000	00000100	00000000	00000000	00000000	00000000	00000000	00001000
11000000	00010010	00000000	00001000	00000000	00111000	00000001	01001100
00000000	01100000	00100010	10000000	00000000	00000010	00000001	00000000
01000010	01000100	00000000	00010000	00000000	01000000	00000001	00000000
01000000	00000001	00000000	10000000	01000000	00000000	00000000	00000100
00000000	00000000	01000010	00000000	00000000	00000000	00000000	00000000
00000000	00000000	00000000	00001010	00001001	10000000	00000000	00000010
00000000	01000000	00000000	00001000	00000000	01000000	00000000	00000000
00001000	01000010	01001111	00100000	00000000	00100000	00000000	10000001
00000001	00000000	01000000	00000000	00000000	00000000	00000000	01000000
00000000	00000000	00100000	01000000	00100000	00000000	00000000	00000011
01000000	000000100	10000001	00000001	00001000	00000100	00001000	00001000
00100000	00000000	00000000	00000000	00000000	01000001	00010011	00000000
00000000	10000000	10000000	00000000	00000000	00001000	01000000	00000000
00001000	00000000	01000010	00011000	00000001	00001001	00000000	00000001
01000010	01001000	01000000	00000010	00000000	10000000	00000100	00000000
00000010	00000000	00000000	00000010	00000000	00100100	00000000	10110100
00001100	00000100	000001010	00000000	00000000	00000000	00000000	00000000
00000010	00000000	00000000	00000000	00100000	10100000	00001000	00000000
01000000	00000000	00000000	00100000	00000000	01000001	00010010	00010001
00000000	00100000	00110000	00000000	00010000	00000000	00000100	00000000
00010100	00000000	000001001	00000001	00000000	00000000	00000000	00000000
00000010	000000100	01010100	10000001	00001000	00000000	00010010	00010000

Embora os atributos estejam rotulados com números em vez de palavras, tanto sua ordem como as magnitudes não são importantes. Poderíamos facilmente deixar 1 representando as mulheres e 0 designando os homens. Os números são usados principalmente com o fim de conveniência; os valores numéricos permitem-nos usar computadores para realizar análises complexas dos dados.

Os dados nominais que assumem um entre dois valores distintos — tal como macho ou fêmea — são chamados *dicotômicos* ou *binários*, dependendo de qual raiz — grega ou latina — para dois seja preferida. No entanto, nem todos os dados nominais precisam ser di-

cotônicos. Freqüentemente, existem três ou mais categorias possíveis, nas quais as observações podem ser classificadas. Por exemplo: as pessoas podem estar agrupadas de acordo com seu tipo sanguíneo, tal que 1 represente o tipo O, 2 é o tipo A, 3 é o tipo B e 4 é o tipo AB. Novamente a seqüência desses valores não é importante. Os números simplesmente servem como rótulos para os diferentes tipos de sangue, tal como o fazem as letras. Precisamos ter isso em mente quando realizarmos as operações aritméticas sobre os dados. Para uma determinada população, um tipo médio de sangue de 1,8 é sem sentido. No entanto, uma operação aritmética que pode ser interpretada é a proporção de indivíduos classificados em cada grupo. Uma análise dos dados na Tabela 2.1 mostra que 9,6% dos pacientes de Aids sofrem de sarcoma de Kaposi e 90,4% não.

2.1.2 Dados Ordinais

Quando a ordem entre as categorias se torna importante, as observações são referenciadas como *dados ordenados*. Por exemplo: as lesões podem ser classificadas de acordo com seu nível de severidade, de modo que 1 representa uma lesão fatal, 2 é severa, 3 é moderada e 4 é pequena. Aqui existe uma ordem natural entre os agrupamentos: um número menor representa uma lesão mais séria. No entanto, ainda não estamos preocupados com a magnitude desses números. Poderíamos ter deixado que 4 representasse uma lesão fatal e 1 uma lesão pequena. Além disso, a diferença entre uma lesão fatal e uma severa não é necessariamente a mesma que entre uma lesão moderada e uma pequena, ainda que ambos os pares de resultados estejam distanciados de uma unidade. Como resultado, muitas operações aritméticas ainda não fazem sentido quando aplicadas a dados ordinais.

A Tabela 2.2 fornece um segundo exemplo de dados ordinais: a escala exibida é usada por oncologistas para classificar o status de desempenho de pacientes registrados em ensaios clínicos [2]. Um *ensaio clínico* é o estudo experimental que envolve indivíduos humanos. Seu objetivo usual é facilitar a comparação de tratamentos alternativos para alguma doença, tal como o câncer. Os indivíduos são aleatoriamente alocados nos diferentes grupos de tratamento e acompanhados até um específico ponto final.

TABELA 2.2

Classificação do status de desempenho de pacientes do Eastern Cooperative Oncology Group.

Status	Definição
0	Paciente totalmente ativo, capaz de ter todo desempenho pré-doença sem restrição.
1	Paciente restrito em atividade fisicamente enérgica, exceto ambulatorial, e capaz de realizar trabalho de natureza leve ou sedentária.
2	Paciente ambulatorial e capaz de todo autocuidado, mas incapaz de realizar qualquer atividade de trabalho; até 50% ou mais das horas acordado.
3	Paciente capaz de somente autocuidado limitado; confinado na cama ou cadeira; mais de 50% das horas acordado.
4	Paciente completamente incapaz, inclusive de qualquer autocuidado; totalmente confinado em cama ou cadeira.

2.1.3 Dados Substituídos por Postos

Em algumas situações, temos um grupo de observações que primeiramente são arranjadas a partir da mais alta para a mais baixa, de acordo com sua magnitude; então, lhes são atribuídos números que correspondem a cada posição da observação na seqüência. Esses tipos de dados são conhecidos como *postos*. Considere, como um exemplo, todas as causas possíveis

de morte nos Estados Unidos. Poderíamos fazer uma lista de todas elas, junto com o número de vidas perdidas, em 1992. Se as causas forem ordenadas a partir da que resultou em maior número de mortes até a que causou o menor e lhes atribuímos números inteiros consecutivos, diz-se que os dados foram substituídos por postos. A Tabela 2.3 lista as dez principais causas de morte nos Estados Unidos em 1992 [3]. Note que as doenças cerebrovasculares estariam ordenadas em terceiro, quer causem 480.000 quer 98.000 mortes. Ao lhes atribuirmos os postos, desprezamos as magnitudes das observações e consideramos somente suas posições relativas. Mesmo com essa imprecisão, é espantoso o volume de informações que os postos contêm. De fato, algumas vezes é melhor trabalharmos com postos do que com os dados originais. Essa questão será explorada posteriormente no Capítulo 13.

TABELA 2.3

As dez causas principais de morte nos Estados Unidos, 1992.

Ordem	Causa da Morte	Total de Mortes
1	Doenças do coração	717.706
2	Neoplasmas malignos	520.578
3	Doenças cerebrovasculares	143.769
4	Doenças pulmonares obstrutivas crônicas	91.938
5	Acidentes e efeitos adversos	86.777
6	Pneumonia e gripe	75.719
7	<i>Diabetes mellitus</i>	50.067
8	Infecção por vírus de imunodeficiência humana	33.566
9	Suicídio	30.484
10	Homicídio e intervenção legal	25.488

2.1.4 Dados Discretos

Para dados discretos, tanto a ordenação como a magnitude são importantes. Nesse caso, os números representam quantidades mensuráveis reais em vez de meros rótulos, e os dados discretos estão restritos a ter somente valores específicos — freqüentemente inteiros ou contagens — que diferem por quantidades fixadas; nenhum valor intermediário é possível. Exemplos de dados discretos incluem o número de acidentes com veículos motorizados em Massachusetts em um mês específico, o número de vezes que uma mulher deu à luz, o número de novos casos de tuberculose registrado nos Estados Unidos durante um período de um ano e o número de camas disponíveis em um hospital particular.

Observe que para os dados discretos existe uma ordem natural entre os valores possíveis. Se estamos interessados no número de vezes que uma mulher deu à luz, por exemplo, um número maior indica que uma mulher teve mais filhos. Além disso, a diferença entre um ou dois nascimentos é a mesma do que entre quatro e cinco nascimentos. Finalmente, o número de nascimentos está restrito a inteiros não-negativos; uma mulher não pode dar à luz 3.4 vezes. Por ser significativo medir a distância entre os possíveis valores de dados para as observações discretas, as regras aritméticas podem ser aplicadas. No entanto, o resultado de uma operação aritmética realizada sobre dois valores de variáveis discretas não é necessariamente discreto. Suponha, por exemplo, que uma mulher tenha dado à luz três vezes, enquanto outra somente duas vezes. O número médio de nascimentos para essas duas mulheres é 2.5, o que não é ele próprio um inteiro.

2.1.5 Dados Contínuos

Dados que representam quantidades mensuráveis, mas que não estão restritos a assumir certos valores especificados (tais como inteiros), são conhecidos como *dados contínuos*. Nesse caso, a diferença entre quaisquer dois valores de dados possíveis pode ser arbitrariamente pequena. Exemplos de dados contínuos incluem o tempo, o nível sérico de colesterol de um paciente, a concentração de um poluente e a temperatura. Em todos eles, os valores fracionais são possíveis. Desde que seja possível medir-se a distância entre duas observações de uma maneira significativa, as operações aritméticas podem ser aplicadas. O único fator que limita uma observação contínua é o grau de precisão com o qual pode ser medida; consequentemente, vemos com freqüência o tempo ser arredondado para o mais próximo segundo e o peso para a mais próxima libra ou grama. Quanto mais precisos forem os instrumentos de medida, maior a quantidade de detalhes que pode ser obtida nos dados registrados.

Às vezes podemos querer um menor grau de detalhe do que o proporcionado pelos dados contínuos; por isso, ocasionalmente transformamos as observações contínuas em discretas, ordinais ou mesmo dicotônicas. Em um estudo dos efeitos do fumo materno nos recém-nascidos, por exemplo, poderíamos primeiro registrar peso ao nascer de um grande número de bebês e então categorizar os bebês em três grupos: aqueles que pesam menos do que 1.500 gramas, aqueles que pesam entre 1.500 e 2.500 gramas e aqueles que pesam mais do que 2.500 gramas. Embora tenhamos a medida real do peso ao nascer, não estamos preocupados se um determinado bebê pesa 1.560 gramas ou 1.580 gramas; estamos interessados somente no número de bebês que fica dentro de cada categoria. A partir da experiência prévia, não podemos esperar diferenças substanciais entre os bebês dentro dos grupos de pesos ao nascer muito baixos, pesos ao nascer baixos e pesos ao nascer normais. Além disso, os dados ordinais são freqüentemente mais fáceis de se manusear do que os contínuos e assim simplificam a análise. No entanto, há uma consequente perda de detalhes na informação sobre os bebês. Geralmente, o grau de precisão exigido em um determinado conjunto de dados depende das questões que estão sendo estudadas.

A Seção 2.1 descreveu uma graduação dos dados numéricos desde os nominais até os contínuos. Conforme prosseguimos, a natureza da relação entre os possíveis valores de dados tornou-se crescentemente complexa. Entre os vários tipos de dados é preciso fazer distinções, pois são usadas técnicas diferentes para analisá-los. Como mencionado anteriormente, não faz sentido falar de um tipo de sangue médio de 1,8; no entanto, faz sentido nos referirmos a uma temperatura média de 24,55 °C.

2.2 Tabelas

Agora que somos capazes de diferenciar os vários tipos de dados, precisamos aprender como identificar as técnicas estatísticas mais apropriadas para descrever cada tipo. Embora um certo volume de informação seja perdido quando os dados são resumidos, um grande volume pode também ser ganho. Uma *tabela* talvez seja o meio mais simples de se resumir um conjunto de observações e pode ser usada para todos os tipos de dados numéricos.

2.2.1 Distribuições de Freqüências

Uma tabela comumente usada para avaliar dados é chamada de *distribuição de freqüências*, que consiste de um conjunto de classes ou de categorias junto com contagens numéricas que correspondam a cada conjunto para dados nominais e ordinais. Como uma ilustração deste formato, a Tabela 2.4 exibe os números de indivíduos (contagens numéricas) que sofriam e não sofriam de

sarcoma de Kaposi (classes ou categorias) para os primeiros 2.560 casos de Aids registrados nos Centros de Controle de Doenças. Um exemplo mais complexo é dado na Tabela 2.5, que especifica o número de cigarros fumados por adulto nos Estados Unidos em vários anos [4].

Para exibir os dados discretos ou contínuos na forma de uma distribuição de freqüências, precisamos dividir o intervalo de valores das observações em uma série de intervalos não-sobrepostos distintos. Se houver muitos intervalos, o resumo não constituirá grande melhoria com relação aos dados brutos. Se houver muito poucos, um grande volume de informação se perderá. Embora não seja necessário, os intervalos são freqüentemente construídos de modo que todos tenham larguras iguais, o que facilita as comparações entre as classes. Uma vez que o limite superior e o inferior tenham sido selecionados, o número de observações cujos valores estejam dentro de cada par de limites é contado e os resultados são arranjados na forma de tabela. Como parte do National Health Examination Survey, por exemplo, os níveis séricos de colesterol de 1.067 homens de 25–34 anos foram registrados para o mais próximo miligrama por 100 mililitros [5]. As observações foram subdivididas então em intervalos de larguras iguais; as freqüências que correspondem a cada intervalo são apresentadas na Tabela 2.6.

TABELA 2.4

Casos de Sarcoma de Kaposi para os primeiros 2.560 pacientes de Aids registrados nos Centros de Controle de Doenças em Atlanta, Geórgia.

Sarcoma de Kaposi	Número de Indivíduos
Sim	246
Não	2.314

TABELA 2.5

Consumo de cigarros por pessoa na idade de 18 anos ou mais velha, Estados Unidos, 1900–1990.

Ano	Número de Cigarros
1900	54
1910	151
1920	665
1930	1.485
1940	1.976
1950	3.522
1960	4.171
1970	3.985
1980	3.851
1990	2.828

TABELA 2.6

Freqüências absolutas dos níveis séricos de colesterol para 1.067 homens dos Estados Unidos, com idades entre 25 e 34 anos, 1976–1980.

Nível de Colesterol (mg/100 ml)	Número de Homens
80–119	13
120–159	150
160–199	442
200–239	299
240–279	115
280–319	34
320–359	9
360–399	5
Total	1.067

A Tabela 2.6 nos dá um quadro global de como os dados se parecem; mostra como os valores do nível sérico de colesterol estão distribuídos pelos intervalos. Note que as observações variam de 80 até 399 mg/100 ml, com relativamente poucas medidas nas extremidades do intervalo e uma grande proporção dos valores situados entre 120 e 279 mg/100ml. O intervalo 160–199 mg/100ml contém o maior número de observações. A Tabela 2.6 nos dá um entendimento muito melhor dos dados se comparada à lista de 1.067 leituras de níveis de colesterol. Embora tenhamos perdido alguma informação — dada a tabela, não podemos recriar os valores brutos dos dados — extraímos também informações importantes que nos auxiliam a entender a distribuição de níveis séricos de colesterol para esse grupo de homens.

O fato de ganhar um tipo de informação enquanto outra se perde permanece verdadeira, mesmo para os dados dicotômicos simples das Tabelas 2.1 e 2.4. Poderíamos achar que não perdíamos qualquer informação ao se resumir esses dados e contar os números de 0s e de 1s, mas realmente perdemos. Por exemplo: se há algum tipo de tendência nas observações no decorrer do tempo — talvez a proporção de pacientes com Aids portadores de sarcoma de Kaposi esteja aumentando ou diminuindo conforme a epidemia amadureça — essa informação é perdida no resumo.

Tabelas são mais informativas quando não se tornam excessivamente complexas. Como uma regra geral, as tabelas e as colunas nelas contidas devem ser sempre claramente rotuladas. Se unidades de medida estiverem envolvidas, tal como mg/100ml para os níveis séricos de colesterol na Tabela 2.6, devem ser especificadas.

2.2.2 Freqüência Relativa

Algumas vezes é útil conhecer a proporção dos valores situados em um determinado intervalo de uma distribuição de freqüências em vez do número absoluto. A *freqüência relativa* para um intervalo é a proporção do número total de observações que nele aparece. Ela é calculada ao dividir-se o número de valores dentro do intervalo pelo número total de valores na tabela. A proporção pode ser deixada como está ou ser multiplicada por 100% para se obter a porcentagem de valores no intervalo. Na Tabela 2.6, por exemplo, a freqüência relativa na classe 80–119 ml/100 ml é $(13/1067) \times 100\% = 1,2\%$; analogamente, a freqüência relativa na classe 120–159 mg/100 ml é $(150/1067) \times 100\% = 14,1\%$. As freqüências relativas para todos os intervalos em uma tabela somam 100%.

Freqüências relativas são úteis para se comparar conjuntos de dados que contenham números desiguais de observações. A Tabela 2.7 exibe as freqüências absolutas e relativas das leituras de níveis séricos de colesterol para os 1.067 homens de 25–34 anos descritas na Tabela 2.6, assim como para um grupo de 1.227 homens de 55–64 anos. Por haver mais homens no grupo de mais idade, é inapropriado comparar as colunas de freqüências absolutas para os dois conjuntos de homens. No entanto, comparar as freqüências relativas é significativo. Podemos ver que, no geral, os homens mais velhos têm maiores níveis séricos de colesterol do que os mais novos; os homens mais novos têm maior proporção de observações em cada um dos intervalos abaixo de 200 mg/100 ml, enquanto os homens mais velhos têm uma maior proporção em cada uma das classes acima desse valor.

A *freqüência relativa acumulada* para um intervalo é a porcentagem do número total de observações que tem um valor menor ou igual ao limite superior do intervalo. A freqüência relativa acumulada é calculada pela soma das freqüências relativas para o intervalo especificado e todas as outras anteriores. Assim, para o grupo de idade de 25–34 anos da Tabela 2.7, a freqüência relativa acumulada do segundo intervalo é $1,2 + 14,1 = 15,3\%$; analogamente, a freqüência relativa acumulada do terceiro intervalo é $1,2 + 14,1 + 41,4 = 56,7\%$. Tal como as freqüências relativas, as freqüências relativas acumuladas são úteis para comparar conjuntos de dados que contenham números desiguais de observações. A Tabela 2.8 lista as freqüências relativas acumuladas para os níveis séricos de colesterol dos dois grupos de homens da Tabela 2.7.

TABELA 2.7

Freqüências absolutas e relativas dos níveis séricos de colesterol para 2.294 homens dos Estados Unidos, 1976–1980.

Nível de Colesterol (mg/100 ml)	Idades 25-34		Idades 55-64	
	Número de Homens	Freqüência Relativa (%)	Número de Homens	Freqüência Relativa (%)
80–119	13	1,2	5	0,4
120–159	150	14,1	48	3,9
160–199	442	41,4	265	21,6
200–239	299	28,0	458	37,3
240–279	115	10,8	281	22,9
280–319	34	3,2	128	10,4
320–359	9	0,8	35	2,9
360–399	5	0,5	7	0,6
Total	1.067	100,0	1.227	100,0

TABELA 2.8

Freqüências relativas e freqüências relativas acumuladas de níveis séricos de colesterol para 2.294 homens dos Estados Unidos, 1976–1980.

Nível de Colesterol (mg/100 ml)	Idades 25-34		Idades 55-64	
	Número de Homens	Freqüência Relativa (%)	Número de Homens	Freqüência Relativa (%)
	Acumulada		Acumulada	
80–119	1,2	1,2	0,4	0,4
120–159	14,1	15,3	3,9	4,3
160–199	41,4	56,7	21,6	25,9
200–239	28,0	84,7	37,3	63,2
240–279	10,8	95,5	22,9	86,1
280–319	3,2	98,7	10,4	96,5
320–359	0,8	99,5	2,9	99,4
360–399	0,5	100,0	0,6	100,0

De acordo com a Tabela 2.7, os homens mais velhos tendem a ter níveis séricos de colesterol mais altos do que os mais jovens. Este é o tipo de generalização que ouvimos muito freqüentemente. Por exemplo, pode-se também dizer que os homens são mais magros do que as mulheres ou que as mulheres vivem mais do que os homens. A generalização com relação ao nível sérico de colesterol não significa que cada homem entre 55 a 64 anos tenha um nível de colesterol mais alto do que cada homem entre 25 a 34 anos, nem significa que o nível sérico de colesterol de cada homem aumente com a idade. O que a declaração implica é que para um determinado nível de colesterol, a proporção de homens mais jovens com leitura menor ou igual a esse valor é maior do que a proporção de homens mais velhos com uma leitura menor ou igual a esse valor. Esse padrão é mais óbvio na Tabela 2.8 do que na Tabela 2.7. Por exemplo, 56,7% dos homens de 25–34 anos têm um nível sérico de colesterol menor ou igual a 199 mg/100 ml, enquanto somente 25,9% dos homens de 55–64 anos estão nessa categoria. Por essas proporções relativas para os dois grupos seguirem tal tendência em cada intervalo na tabela, as duas distribuições são denominadas *estocasticamente ordenadas*. Para qualquer

nível especificado, uma maior proporção de homens mais velhos tem leituras de nível sérico de colesterol acima deste valor do que os mais jovens; em consequência, a distribuição de níveis para os homens mais velhos é estocasticamente maior do que para os mais jovens. Essa definição fará mais sentido quando estudarmos as variáveis aleatórias e as distribuições de probabilidade no Capítulo 7. Lá, as implicações dessa ordenação se tornarão mais aparentes.

2.3 Gráficos

Um segundo modo para resumir e exibir os dados é pelo uso de gráficos ou representações pictográficas dos dados numéricos. Os gráficos devem ser concebidos de modo a transmitirem os padrões gerais de um conjunto de observações em uma simples visualização. Embora sejam mais fáceis para se ler do que as tabelas, os gráficos freqüentemente fornecem menor grau de detalhe. Entretanto, a perda de detalhes pode ser acompanhada por um ganho no entendimento dos dados. Os gráficos mais informativos são relativamente simples e auto-explicativos. Tal como as tabelas, devem ser claramente rotulados e as unidades de medida devem ser indicadas.

2.3.1 Gráficos de Barras

Os gráficos de barras são um tipo popular de gráfico usados para exibir uma distribuição de freqüências para os dados nominais e ordinais. Em um gráfico de barras, as várias categorias nas quais as observações são classificadas estão apresentadas ao longo de um eixo horizontal. Uma barra vertical é desenhada por cima de cada categoria de tal modo que a altura da barra represente a freqüência ou a freqüência relativa de observações dentro daquela classe. As barras devem ser de igual largura e separadas uma da outra de modo a não implicar continuidade. Como exemplo, temos a Figura 2.1, um gráfico de barras que exibe os dados relativos ao consumo de cigarros nos Estados Unidos apresentados na Tabela 2.5. Note-se que quando é representada na forma de um gráfico, a tendência ao consumo de cigarros no decorrer dos anos é ainda mais aparente do que o que é na tabela.

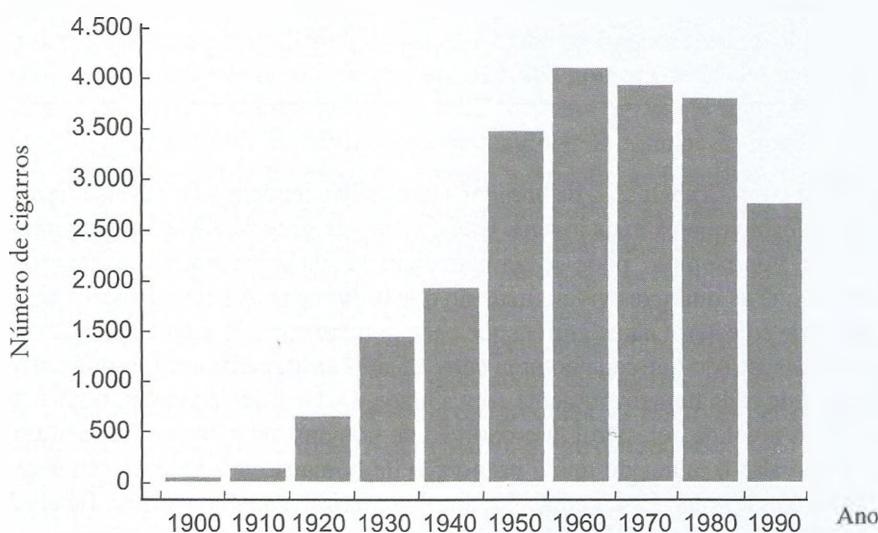


FIGURA 2.1

Gráfico de barras: consumo de cigarros por pessoa na idade de 18 anos ou mais velha, Estados Unidos, 1900–1990.

2.3.2 Histogramas

Talvez o tipo de gráfico mais comumente utilizado seja o *histograma*. Enquanto um gráficos de barras configura a representação pictográfica de uma distribuição de freqüências tanto para os dados nominais como ordinais, um histograma mostra uma distribuição de freqüências para os dados discretos ou contínuos. O eixo horizontal exibe os limites verdadeiros dos vários intervalos, que são os pontos que os separam dos outros intervalos em ambos os lados. Por exemplo, a fronteira entre as duas primeiras classes de nível sérico de colesterol da Tabela 2.6 é 119,5 mg/100 ml; ele é o limite superior verdadeiro do intervalo 80–119 e o limite inferior verdadeiro de 120–159. O eixo vertical de um histograma mostra a freqüência ou a freqüência relativa das observações dentro de cada intervalo.

A primeira etapa na construção de um histograma é traçar as escalas dos eixos. A escala vertical deve começar do zero; se isso não é feito, as comparações visuais entre os intervalos podem ficar distorcidas. Uma vez que os eixos tenham sido desenhados, uma barra vertical centrada no ponto médio é colocada sobre cada intervalo. A altura da barra demarca a freqüência associada com o intervalo. Como exemplo, a Figura 2.2 exibe um histograma construído a partir dos dados dos níveis séricos de colesterol na Tabela 2.6.

Na realidade, a freqüência associada a cada intervalo em um histograma é representada não pela altura da barra acima dela, mas pela área da barra. Assim, na Figura 2.2, 1,2% da área total corresponde às 13 observações que existem entre 79,5 e 119,5 mg/100 ml e 14,1% da área corresponde às 150 observações entre 119,5 e 159,5 mg/100 ml. A área do histograma inteiro soma 100% ou 1. Note-se que a proporção da área total que corresponde a um intervalo é igual à freqüência relativa daquele intervalo. Como resultado, um histograma que exibe freqüências relativas — tal como a Figura 2.3 — terá a mesma forma de um histograma com freqüências absolutas. Porque é a área de cada barra que representa a proporção relativa de observações em um intervalo, é preciso tomar cuidado quando se constrói um histograma com larguras de intervalos diferentes; a altura precisa variar junto com a largura, de modo que a área de cada uma das barras permaneça em proporção apropriada.

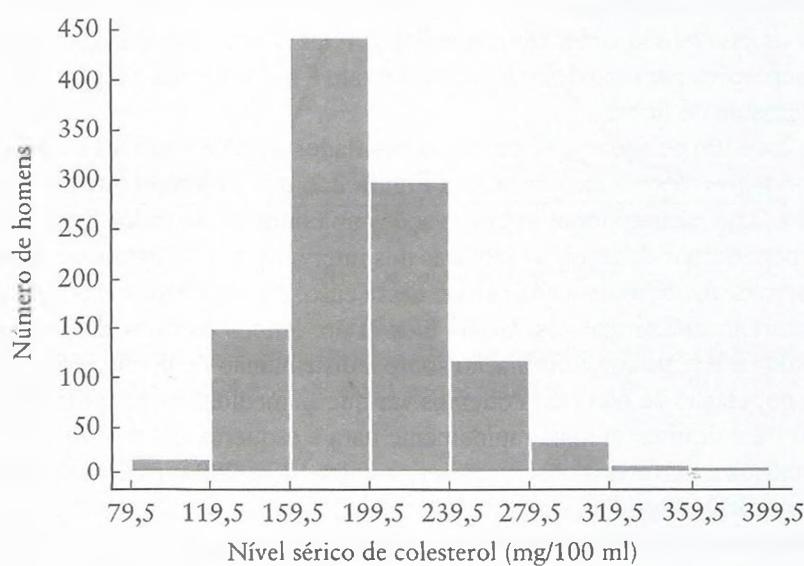
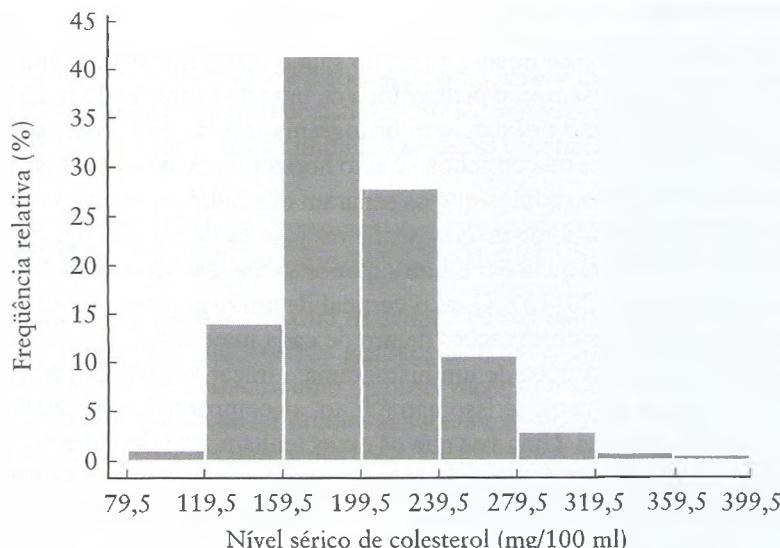


FIGURA 2.2

Histograma: freqüências absolutas de níveis séricos de colesterol para 1.067 homens dos Estados Unidos, com idade entre 25 e 34 anos, 1976–1980.

**FIGURA 2.3**

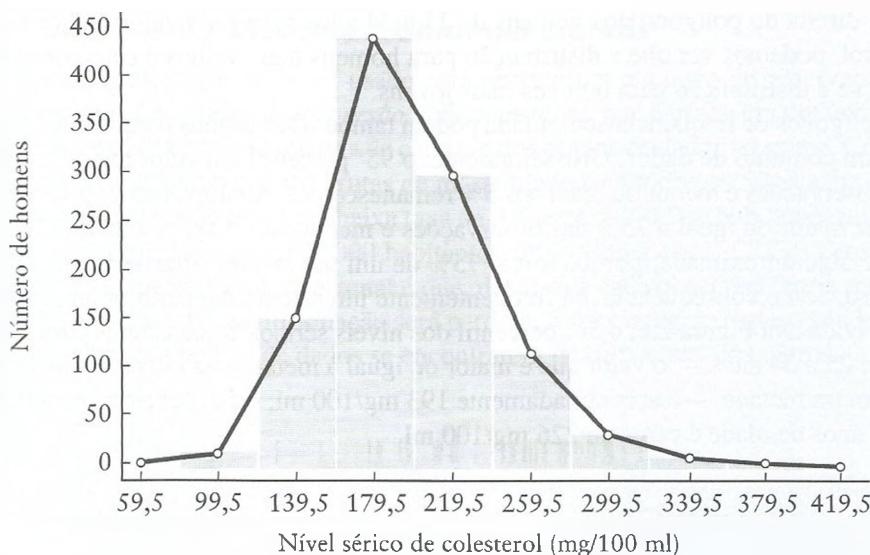
Histograma: freqüências relativas de níveis séricos de colesterol para 1.067 homens dos Estados Unidos, com idade entre 25 e 34 anos, 1976–1980.

2.3.3 Polígonos de Freqüência

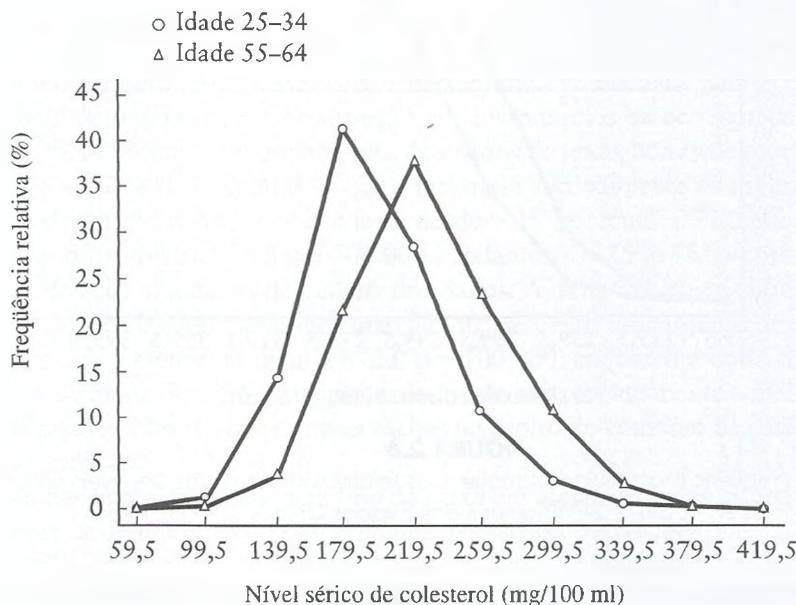
O *polígono de freqüência*, outro gráfico comumente utilizado, é similar ao histograma em muitos aspectos. Um polígono de freqüência usa os mesmos dois eixos que um histograma, e é construído ao se colocar um ponto no centro de cada um dos intervalos de forma tal que a altura do ponto seja igual à freqüência ou freqüência relativa associada com o intervalo. Pontos também são colocados no eixo horizontal nos pontos médios dos intervalos imediatamente precedentes e imediatamente seguintes aos intervalos que contêm as observações. Então, os pontos são conectados por linhas retas. Como em um histograma, a freqüência de observações para um determinado intervalo é representada pela área dentro dele e abaixo do segmento de linha.

A Figura 2.4 é um polígono de freqüência dos dados de níveis séricos de colesterol na Tabela 2.6. Compare-o com o histograma na Figura 2.2, que está reproduzido muito levemente no fundo. Se o número total de observações no conjunto de dados fosse aumentado regularmente, poderíamos diminuir as larguras dos intervalos no histograma e ainda ter um adequado número de medidas em cada classe; nesse caso, o histograma e o polígono de freqüência se tornariam indistinguíveis. Como eles estão, ambos os tipos de gráficos transmitem essencialmente a mesma informação sobre a distribuição de níveis séricos de colesterol para essa população de homens. Podemos ver que as medidas estão centradas ao redor de 180 mg/100 ml e diminuem mais rapidamente para a esquerda desse valor do que o fazem para a direita. A maioria das observações fica entre 120 e 280 mg/100 ml e todas estão entre 80 e 400 mg/100 ml.

Por poderem ser facilmente superpostos, os polígonos de freqüência são superiores aos histogramas para se comparar dois ou mais conjuntos de dados. A Figura 2.5 exibe os polígonos de freqüência dos dados de níveis séricos de colesterol apresentados na Tabela 2.7. Como os homens mais velhos tendem a ter níveis séricos de colesterol mais altos, seu polígono fica à direita do polígono dos homens mais jovens.

**FIGURA 2.4**

Polígono de freqüência: freqüências absolutas de níveis séricos de colesterol para 1.067 homens dos Estados Unidos, com idade entre 25 e 34 anos, 1976–1980.

**FIGURA 2.5**

Polígono de Freqüência: freqüências relativas de níveis séricos de colesterol para 2.294 homens dos Estados Unidos, 1976–1980.

Embora seu eixo horizontal seja o mesmo de um polígono de freqüência padrão, o eixo vertical de um *polígono de freqüência acumulada* exibe freqüências relativas acumuladas. Um ponto é colocado no limite superior verdadeiro de cada intervalo; a altura do ponto representa a freqüência relativa acumulada associada ao intervalo. Os pontos são então conectados por linhas retas. Como os polígonos de freqüência, os polígonos de freqüência acumulada podem ser usados para comparar conjuntos de dados, conforme é ilustrado na Figura 2.6. Notando-se que o polígono de freqüência acumulada dos homens de 55 a 64 anos se

encontra à direita do polígono dos homens de 25 a 34 anos para cada valor de nível sérico de colesterol, podemos ver que a distribuição para homens mais velhos é estocasticamente maior do que a distribuição para homens mais jovens.

Os polígonos de freqüência acumulada podem também ser usados para se obter os *percentis* de um conjunto de dados. Grosseiramente, o 95º percentil é o valor maior ou igual a 95% das observações e menor ou igual aos 5% remanescentes. Analogamente, o 75º percentil é o valor maior ou igual a 75% das observações e menor ou igual aos outros 25%. Essa definição é algo aproximada, porque tomar 75% de um inteiro tipicamente não resulta em outro inteiro; como consequência, há freqüentemente um arredondamento ou uma interpolação envolvida. Na Figura 2.6, o 50º percentil dos níveis séricos de colesterol para o grupo de idade de 25 a 34 anos — o valor que é maior ou igual à metade das observações e menor ou igual à outra metade — é aproximadamente 193 mg/100 ml; o 50º percentil para o grupo de 55 a 64 anos de idade é cerca de 226 mg/100 ml.

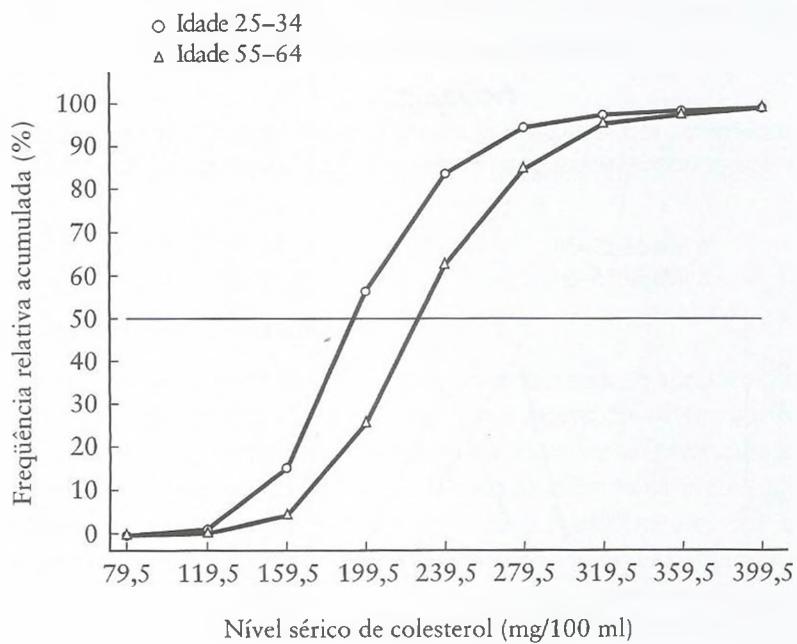


FIGURA 2.6

Polígono de freqüência acumulada: freqüências relativas acumuladas de níveis séricos de colesterol para 2.294 homens dos Estados Unidos, 1976–1980.

Os percentis são úteis para descrever a forma de uma distribuição. Por exemplo, se o 40º percentil e o 60º percentil de um conjunto de dados se encontram em distâncias iguais do ponto médio, e isso também é verdadeiro para o 30º percentil e o 70º percentil, para o 20º percentil e o 80º percentil e para todos os outros percentis que somam 100, os dados são *simétricos*, isto é, a distribuição de valores tem a mesma forma de cada lado do 50º percentil. Alternativamente, se há diversas observações afastadas de apenas um lado do ponto médio, diz-se que os dados são *assimétricos*. Se essas observações são menores do que os valores restantes, os dados são assimétricos à esquerda ; se eles são maiores do que as outras medidas, os dados são assimétricos à direita. As várias formas que uma distribuição de dados pode assumir serão discutidas posteriormente no Capítulo 3.

2.3.4 Gráficos de Dispersão Unidimensionais

Outro tipo de gráfico que pode ser usado para resumir um conjunto de observações discreteas ou contínuas é o *gráfico de dispersão unidimensional*, que é usado em um único eixo horizontal para exibir a posição relativa de cada um dos pontos de dados no grupo. Como exemplo, a Figura 2.7 mostra as taxas brutas de mortalidade para todos os 50 estados e o distrito de Colúmbia em 1992, desde uma baixa taxa de 319,8 para 100.000 habitantes no Alasca até uma alta taxa de 1.214,9 por 100.000 habitantes em Washington, D.C. [3]. Uma vantagem do gráfico de dispersão unidimensional é que, desde que cada observação seja representada individualmente, nenhuma informação será perdida: a desvantagem é que a sua leitura pode ser difícil se muitos pontos de dados se encontrarem próximos uns dos outros.

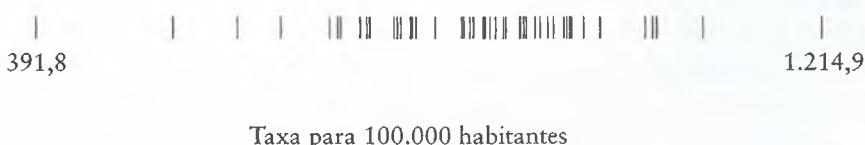


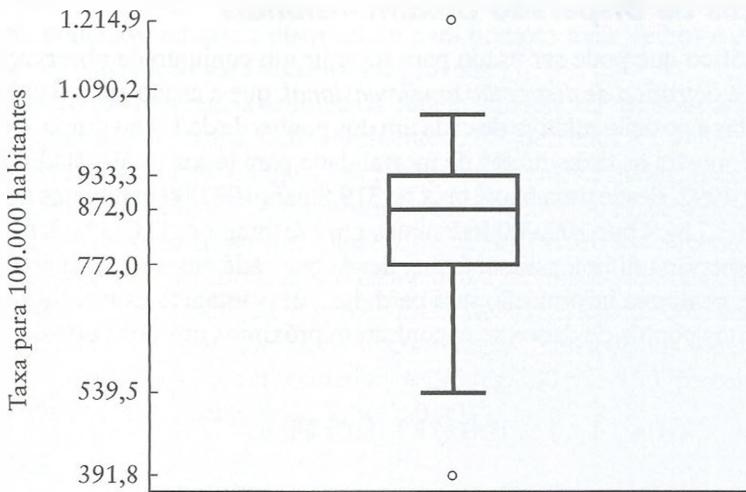
FIGURA 2.7

Gráfico de dispersão unidimensional: taxas brutas de mortalidade para os Estados Unidos, 1992.

2.3.5 Box Plots

Box plots são similares aos gráficos de dispersão unidimensionais, pois exigem um único eixo; em vez de se plotar cada observação, no entanto, eles exibem somente um resumo dos dados [6]. A Figura 2.8 é um box plot dos dados de taxas brutas de mortalidade exibidos na Figura 2.7. A caixa central — que é mostrada verticalmente na Figura 2.8, mas que também pode ser horizontal — estende-se desde o 25º percentil, 772,0 por 100.000 habitantes até o 75º percentil, 933,3 por 100.000 habitantes. Os 25º e 75º percentis de um conjunto de dados são chamados de *quartis* dos dados. A linha que corre entre os quartis em 872,0 mortes por 100.000 habitantes marca o 50º percentil do conjunto de dados; metade das observações é menor ou igual a 872,0 por 100.000, enquanto a outra metade é maior ou igual a esse valor. Se o 50º percentil encontra-se aproximadamente a meio caminho entre os dois quartis, implica que as observações no centro do conjunto de dados são grosseiramente simétricas.

As linhas que se projetam para fora da caixa em ambos os lados estendem-se para valores adjacentes do gráfico. Os *valores adjacentes* são as observações mais extremas no conjunto de dados que não estão a mais de 1,5 vez a altura da caixa além dos quartis. Na Figura 2.8, 1,5 vez a altura da caixa é $1,5 \times (933,3 - 772,0) = 242,0$ por 100.000 da população. Conseqüentemente, os valores adjacentes são as observações menores e maiores no conjunto de dados que não são mais extremos que $772,0 - 242,0 = 530,0$ e $933,3 + 242,0 = 1.175,3$ por 100.000 habitantes respectivamente ou 539,5 por 100.000 e 1.090,2 por 100.000 habitantes. Nos conjuntos razoavelmente simétricos, os valores adjacentes devem conter aproximadamente 99% das medidas. Todos os pontos fora desse intervalo são representados por círculos; essas observações são consideradas fora do padrão ou pontos dos dados que são atípicos dos valores restantes.

**FIGURA 2.8**

Box plot: taxas brutas de mortalidade para os Estados Unidos, 1992.

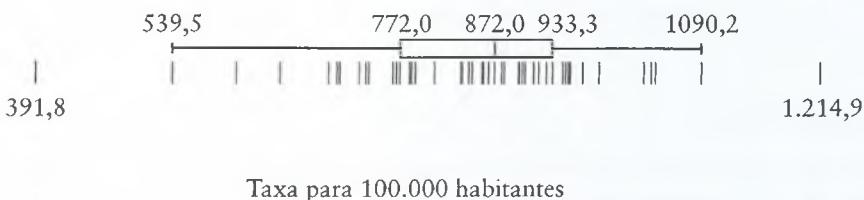
**FIGURA 2.9**

Gráfico de dispersão unidimensional e box plot:
taxas brutas de mortalidade para os Estados Unidos, 1992.

Deve-se notar que a explanação anterior é meramente um modo de definir um box plot: outras definições existem e exibem variados graus de complexidade [7]. Além disso, embora um box plot transmita uma clara quantidade de informação sobre a distribuição de um conjunto de números, um volume de informação ainda maior pode ser exibido ao se combinar o gráfico de dispersão unidimensional e o box plot, como na Figura 2.9.

2.3.6 Gráficos de Dispersão Bidimensionais

Diferentemente dos outros gráficos que discutimos até agora, um gráfico de dispersão bidimensional é usado para mostrar a relação entre duas medidas contínuas distintas. Cada um dos pontos no gráfico representa um par de valores; a escala para uma quantidade está marcada no eixo horizontal, ou eixo x, e a escala da outra no eixo vertical, ou eixo y. Por exemplo, a Figura 2.10 plota duas medidas simples da função do pulmão — capacidade vital forçada (em inglês, FVC — *forced vital capacity*) e o volume expiratório forçado em um segundo (FEV₁ — *forced expiratory volume in one second*, em inglês) — para 19 indivíduos asmáticos que participaram de um estudo que investigou os efeitos físicos do dióxido sulfúrico [8]. A capacidade vital forçada é o volume de ar que pode ser expelido dos pulmões em seis segundos e o volume expiratório forçado em um segundo é o que pode ser expelido depois de um segundo de esforço constante. Note-se que o indivíduo representado pelo ponto mais afastado para a esquerda tem uma medida de FEV₁ de 2,0 litros e uma medida de FVC de 2,8 litros. (Somente 18 pontos estão marcados no gráfico em vez de 19).

porque dois indivíduos tiveram valores idênticos de FVC e FEV₁; como consequência, um ponto encontra-se diretamente em cima de outro.) Como se poderia esperar, o gráfico indica que há uma forte relação entre essas duas quantidades; a FVC aumenta em magnitude quando a FEV₁ cresce.

2.3.7 Gráficos de Linha

O gráfico de linha é similar ao gráfico de dispersão bidimensional, pois pode ser usado para ilustrar a relação entre quantidades contínuas. Uma vez mais, cada ponto no gráfico representa um par de valores. Nesse caso, no entanto, cada valor no eixo x tem uma única medida correspondente no eixo y. Pontos adjacentes estão conectados por linhas retas. Mais comumente, a escala ao longo do eixo horizontal representa o tempo. Dessa forma, somos capazes de traçar a mudança cronológica na quantidade no eixo vertical em um período de tempo especificado. Como exemplo, a Figura 2.11 exibe a tendência nas taxas registradas de malária, inclusive as mudanças oriundas de fontes identificáveis, que ocorreram nos Estados Unidos.

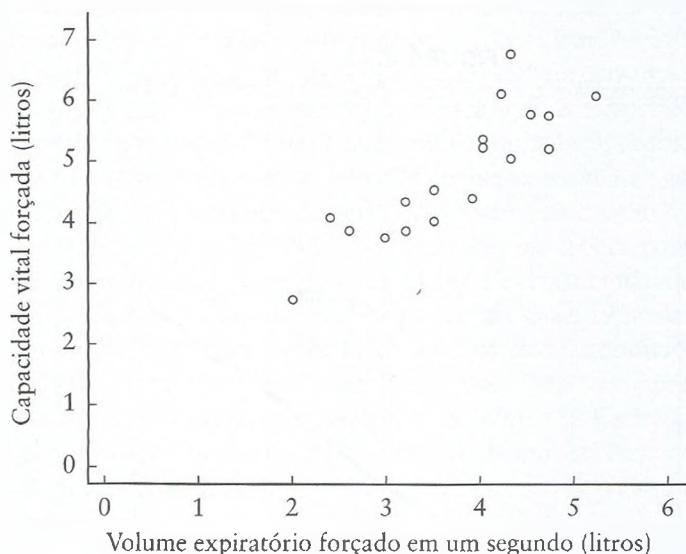


FIGURA 2.10

Gráfico de dispersão bidimensional e plotagem de caixa: capacidade vital forçada versus volume expiratório forçado em um segundo para 19 indivíduos asmáticos.

Unidos entre 1940 e 1989 [9]. Observe a escala logarítmica no eixo vertical; esta escala nos permite mostrar um grande intervalo de observações, embora mostre ainda a variação entre os valores menores.

Para comparar dois grupos ou mais em relação a uma determinada quantidade, é possível plotar mais do que uma medida ao longo do eixo y. Suponha que estamos interessados nos crescentes custos de cuidados com a saúde. Para investigar esse problema, poderíamos querer comparar as variações no custo ocorridas em dois sistemas de cuidados com a saúde diferentes em anos recentes. A Figura 2.12 mostra a tendência nos gastos de cuidados com a saúde tanto nos Estados Unidos como no Canadá, entre 1970 e 1989 [10].

Nesta seção, não tentamos examinar todos os possíveis tipos de gráficos. Em vez disso, incluímos somente uma seleção dos mais comuns. Deve-se notar que existem muitas outras representações imaginativas [11]. Como regra geral, no entanto, não se deve colocar muita informação dentro de um simples gráfico. Com freqüência, uma ilustração relativamente simples é a mais efetiva.

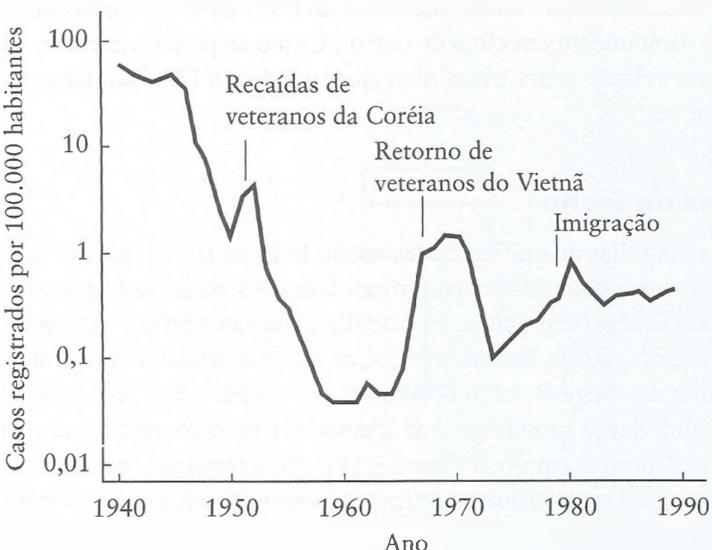
**FIGURA 2.11**

Gráfico de linha: taxas registradas de malária por ano, Estados Unidos, 1940–1989.

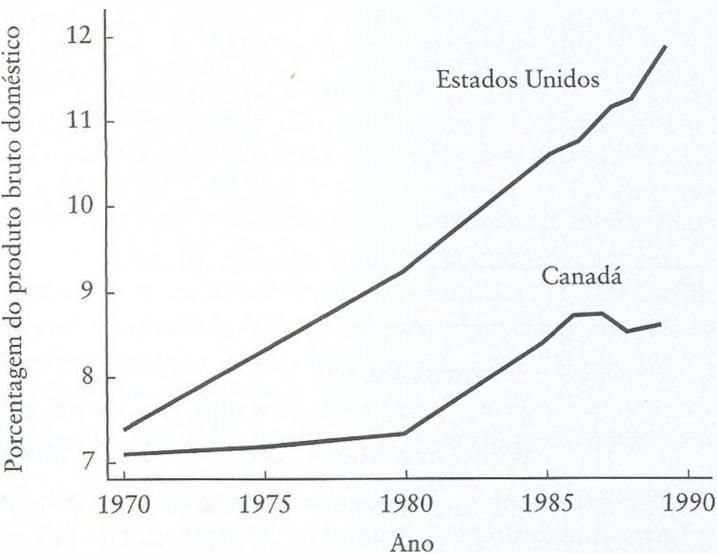
**FIGURA 2.12**

Gráfico de linha: gastos com cuidados com a saúde como uma porcentagem do produto bruto doméstico para os Estados Unidos e Canadá, 1970–1989.

2.4 Aplicações Adicionais

Suponha que queiramos reduzir o número de mortes infantis causadas por lesões. Primeiro, necessitamos entender a natureza do problema. A seguir, é apresentado um conjunto de dados que indica as causas de morte para 100 crianças entre as idades de cinco e nove anos vítimas fatais de lesões [12]. Os dados são nominais: 1 representa acidente por veículo moto-

rizado, 2 afogamento, 3 incêndio no lar, 4 homicídio e 5 designa outras causas, inclusive sufocamento, quedas e envenenamento. Depois de fornecidos esses dados, que podemos concluir com relação às mortes infantis por lesão?

1	5	3	1	2	4	1	3	1	5
2	1	1	5	3	1	2	1	4	1
4	1	3	1	5	1	2	1	1	2
5	1	1	5	1	5	3	1	2	1
2	3	1	1	2	1	5	1	5	1
1	2	5	1	1	1	3	4	1	1
1	1	2	1	1	2	1	1	2	3
3	3	1	5	2	3	5	1	3	4
1	1	2	4	5	4	1	5	1	5
5	1	1	5	1	1	5	1	1	5

Embora o conjunto inteiro de resultados esteja disponível, é extremamente difícil fazer qualquer tipo de afirmação sobre esses dados. No entanto, se desejássemos resumir as observações, poderíamos começar construindo a distribuição de freqüências. Para os dados nominais e ordinais, uma distribuição de freqüências é uma tabela constituída de uma lista de categorias ou classes junto com as contagens numéricas que correspondem a cada uma delas. Para construir uma distribuição de freqüências para o conjunto de dados mostrados acima, começaríamos por listar as várias causas de morte; contaríamos então o número de crianças que morreram vítimas de cada uma dessas causas. As observações estão exibidas no formato de distribuição de freqüências na Tabela 2.9. Ao usarmos essa tabela, podemos ver que 48 dessas 100 mortes por lesão resultaram de acidentes de veículos motorizados, 14 foram causadas por afogamento, 12 por incêndios no lar, sete por homicídio e 19 por outras causas.

Tal como os dados nominais e ordinais, os discretos e contínuos podem também ser exibidos na forma de uma distribuição de freqüências. Para tanto, precisamos subdividir o intervalo de valores dos resultados em uma série de intervalos distintos não-sobrepostos. Os números de observações situadas dentro de cada par de limites são contados e arranjados em uma tabela. Suponha que estamos interessados em estudar as consequências do baixo peso ao nascer entre os recém-nascidos nos Estados Unidos. Para colocar a magnitude do problema no contexto, examinamos primeiro a distribuição dos pesos ao nascer de todos os bebês em 1986 [13]. Separamos essas observações em intervalos de igual largura; as freqüências correspondentes são exibidas na Tabela 2.10, que nos fornece mais informações sobre a distribuição de pesos ao nascer do que uma lista de 3.751.275 medidas. Podemos ver que a maioria das observações se encontra entre 2.000 e 4.499 gramas; relativamente poucas medidas estão fora desse intervalo. Os intervalos 3.000-3.499 e 3.500-3.999 gramas contêm as maiores quantidades de valores.

TABELA 2.9

Mortes por lesão de 100 crianças entre as idades de 5 e 9 anos, Estados Unidos, 1980–1985.

Causa	Número de Mortes
Veículo a motor	48
Afogamento	14
Incêndio no lar	12
Homicídio	7
Outros	19
Total	100

Depois de verificarmos as contagens reais, poderíamos também estar interessados em encontrar a freqüência relativa associada a cada intervalo na tabela. A freqüência relativa é a porcentagem do número total de observações que se encontra dentro de um intervalo. As freqüências relativas para os pesos ao nascer exibidos na Tabela 2.10 — calculadas dividindo-se o número de valores no intervalo pelo número total de medidas na tabela e multiplicando-se por 100 — são exibidas na Tabela 2.11. As tabelas indicam que $36,7 + 29,5 = 66,2\%$ dos pesos ao nascer estão entre 3.000 e 3.999 gramas e $4,3 + 15,9 + 36,7 + 29,5 + 9,2 = 95,6\%$ estão entre 2.000 e 4.499 gramas. Somente 2,5% das crianças nascidas em 1986 pesavam menos que 2.000 gramas.

Além das tabelas, podemos também usar gráficos para resumir e exibir um conjunto de dados. Por exemplo, poderíamos ilustrar os dados nominais da Tabela 2.9 usando o gráfico de barras na Figura 2.13. As categorias dentro das quais as observações se situam são colocadas no eixo horizontal; as barras verticais representam a freqüência de observações em cada classe. O gráfico enfatiza que uma grande proporção de mortes por lesões infantis resulta de acidentes por veículos motorizados.

Um gráfico de barras empilhadas pode ser usado para transmitir maior volume de informação em um único quadro. Nesse tipo de gráfico, as barras que representam a freqüência das observações em dois ou mais subgrupos diferentes são colocadas uma em cima das outras. Como exemplo, a Figura 2.14 exibe as taxas de mortalidade por 1.000 nascimentos (o número de mortes para cada 1.000 nascimentos) na França para quatro categorias de bebês — aqueles que eram natimortos, aqueles que morreram menos de uma semana depois

TABELA 2.10

Freqüências absolutas de pesos ao nascer para 3.751.275 bebês nascidos nos Estados Unidos, 1986.

Peso ao Nascer (gramas)	Número de Bebês
0–499	4.843
500–999	17.487
1.000–1.499	23.139
1.500–1.999	49.112
2.000–2.499	160.919
2.500–2.999	597.738
3.000–3.499	1.376.008
3.500–3.999	1.106.634
4.000–4.499	344.390
4.500–4.999	62.769
5.000–5.500	8.236
Total	3.751.275

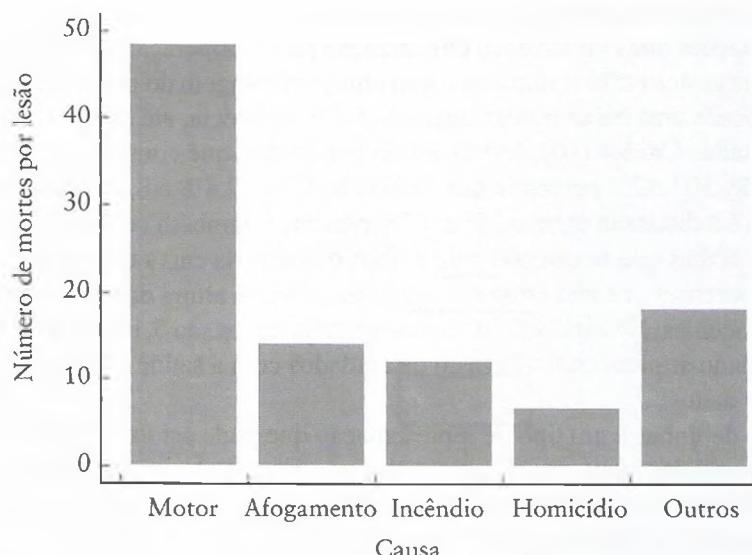
TABELA 2.11

Freqüências relativas de pesos ao nascer para 3.751.275 bebês nascidos nos Estados Unidos, 1986.

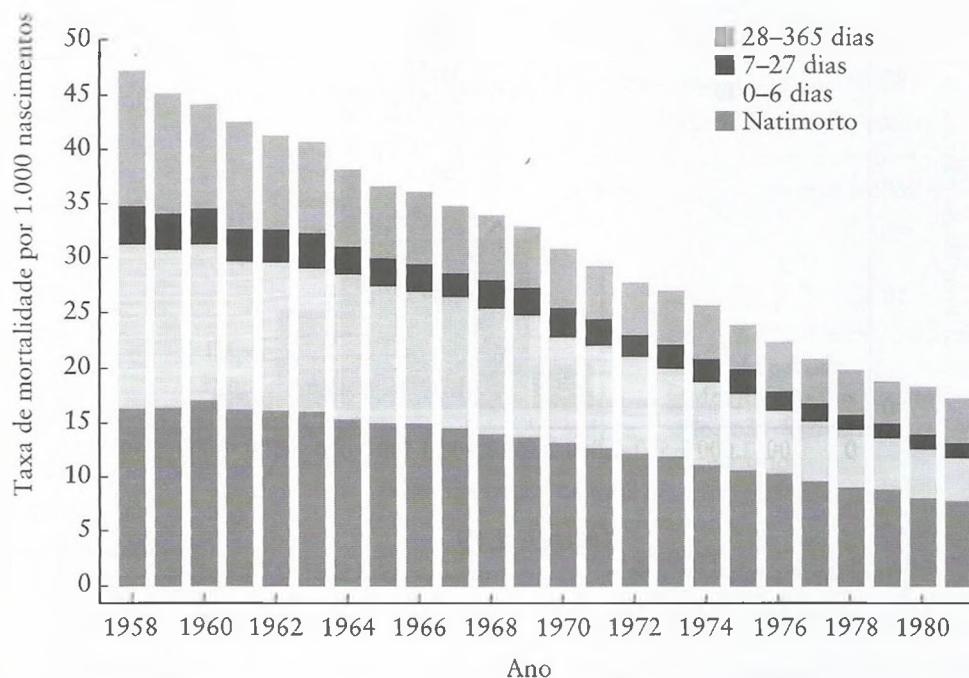
Peso ao Nascer (gramas)	Freqüência Relativa (%)
0–499	0,1
500–999	0,5
1.000–1.499	0,6
1.500–1.999	1,3
2.000–2.499	4,3
2.500–2.999	15,9
3.000–3.499	36,7
3.500–3.999	29,5
4.000–4.499	9,2
4.500–4.999	1,7
5.000–5.500	0,2
Total	100,0

de nascer, aqueles que morreram entre 7 e 27 dias depois de nascer e aqueles que sobreviveram por mais de 27 dias, mas menos que um ano [14]. Como cada uma dessas taxas diminui no tempo, o gráfico é capaz de produzir uma afirmação poderosa sobre a mortalidade infantil global.

Das várias representações gráficas que podem ser usadas para exibir os dados discretos e contínuos, o histograma talvez seja o mais comum. Tal como um gráfico de barras, o histograma é uma representação pictográfica de uma distribuição de freqüências. O eixo horizontal exibe os limites verdadeiros dos intervalos dentro dos quais as observações são clas-

**FIGURA 2.13**

Morte por lesão de 100 crianças entre as idades de 5 e 9 anos, Estados Unidos, 1980–1985.

**FIGURA 2.14**

Mortalidade infantil e perinatal na França, 1978–1981.

sificadas; o eixo vertical mostra a freqüência ou freqüência relativa das observações dentro de cada um dos intervalos. Como exemplo, a Figura 2.15 é um histograma dos dados do peso ao nascer resumidos na Tabela 2.10. Ao olharmos para o gráfico, podemos ver que os dados são assimétricos à esquerda.

Um box plot é outro tipo de gráfico freqüentemente usado para os dados discretos e contínuos. O gráfico exibe um resumo das observações usando um eixo vertical ou horizon-

tal. Suponha que estejamos interessados em comparar os gastos de cuidados com a saúde em 1989 para as 24 nações que constituem a Organização para Cooperação Econômica e Desenvolvimento. Esses gastos estão resumidos como uma porcentagem do produto bruto nacional na Figura 2.16, desde uma baixa porcentagem de 5,1% na Grécia, até uma alta porcentagem de 11,8% nos Estados Unidos [10]. As três linhas horizontais que constituem a caixa central indicam que o 25º, 50º e 75º percentis dos dados são 6,7%, 7,4% e 8,3%, respectivamente. A altura da caixa é a distância entre o 25º e o 75º percentil, também conhecidos como quartis dos dados. As linhas que se estendem de ambos os lados da caixa central marcam as observações mais extremas que não estão mais do que 1,5 vez a altura da caixa além dos quartis ou valores adjacentes. Na Figura 2.16, os valores adjacentes são 5,1% e 8,8%. Os Estados Unidos têm um dado atípico, com um gasto de cuidados com a saúde que não é comum entre o restante dos dados.

Um gráfico de linhas é um tipo de representação que pode ser usado para ilustrar a relação entre duas medidas contínuas. Cada um dos pontos na linha representa um par de valores; as próprias linhas nos permitem traçar a mudança na quantidade no eixo y que corresponde a uma mudança ao longo do eixo x. A Figura 2.17, tal como a Figura 2.1, mostra os dados relacionados ao consumo de cigarros nos Estados Unidos. Note que o gráfico de linha mostra mais detalhes do que o de barras correspondente.

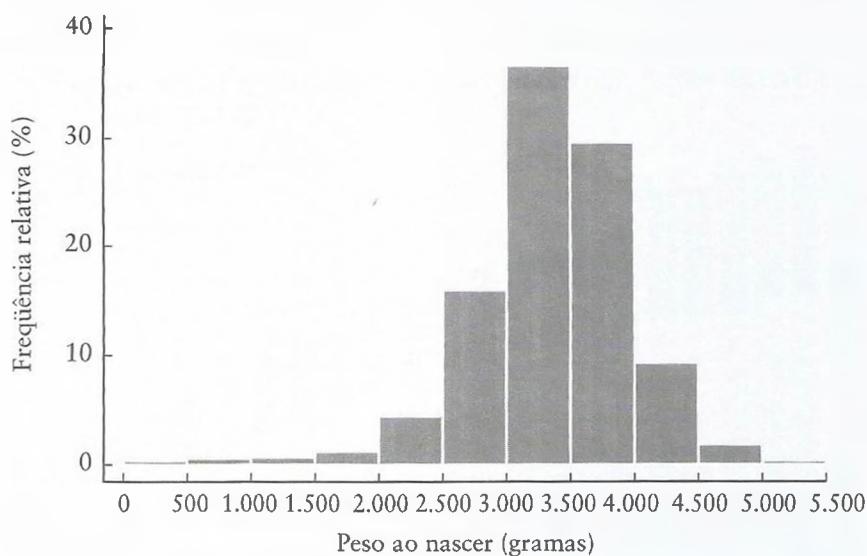
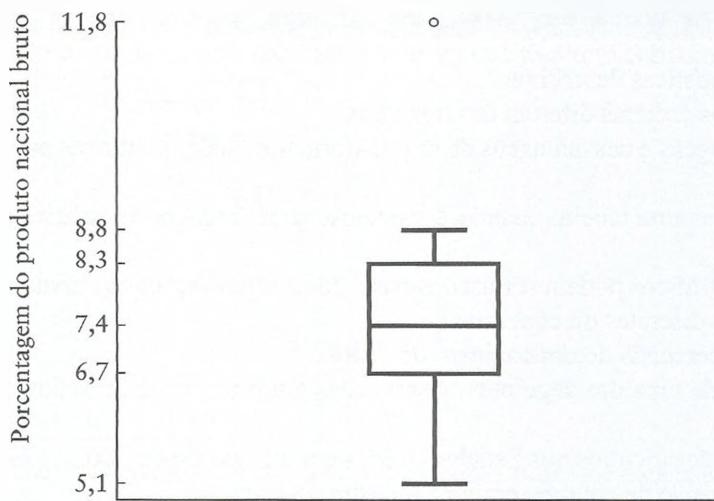


FIGURA 2.15
Frequências relativas dos pesos ao nascer para 3.751.275 bebês nascidos nos Estados Unidos, 1986.

No decorrer dos anos, o uso de computadores em estatística aumentou drasticamente. Como resultado, muitos dos demorados cálculos de antigamente podem agora ser realizados com muito mais eficiência usando-se um pacote estatístico, que consiste de uma série de programas que foram concebidos para analisar dados numéricos. Uma variedade de pacotes está disponível; geralmente, eles diferem com relação aos comandos que usa e aos formatos das saídas que produz.

Um pacote estatístico que tanto é poderoso quanto relativamente fácil de se usar é o chamado *Stata*, um programa interativo que nos auxilia a gerenciar, exibir e analisar dados. As observações ou medidas são colocadas em colunas; a cada coluna é atribuído um nome de variável. Usamos esses nomes para executar procedimentos analíticos específicos ou comandos. Quando for apropriado, reproduzimos a saída do Stata para ilustrar o que o computador

**FIGURA 2.16**

Gastos de cuidados com a saúde como uma porcentagem do produto nacional bruto para 24 nações, 1989.

**FIGURA 2.17**

Consumo de cigarros por pessoa com idade de 18 anos ou mais, Estados Unidos, 1900–1990.

é capaz de fazer. Se alguns leitores quiserem usar outro pacote estatístico, incorporamos também os resultados do Minitab e do SAS.

Computadores são particularmente úteis para se construir gráficos. De fato, as Figuras de 2.13 até 2.17 foram geradas pelo Stata. Para criar a Figura 2.17, salvamos os anos entre 1900 e 1990 sob o nome de variável *ano* e salvamos os valores correspondentes de consumo *per capita* de cigarros sob o nome *cigarro* (os nomes das variáveis no Stata não podem ter mais que oito letras). O Stata plotou os pontos que representam cada par de valores. Os pontos foram conectados e os rótulos foram adicionados com o uso de comandos apropriados.

2.5 Exercícios de Revisão

1. O que são estatísticas descritivas?
2. Como os dados ordinais diferem dos nominais?
3. Quais as vantagens e desvantagens de se transformar medidas contínuas em discretas ou ordinais?
4. Ao se construir uma tabela, quando é vantajoso usar freqüências relativas em vez de absolutas?
5. Que tipos de gráficos podem ser usados para exibir observações nominais ou ordinais? E observações discretas ou contínuas?
6. O que são os percentis de um conjunto de dados?
7. Declare se cada uma das seguintes observações é um exemplo de dados discretos ou contínuos:
 - (a) O número de suicídios nos Estados Unidos em um ano específico.
 - (b) A concentração de chumbo em uma amostra de água.
 - (c) A duração de tempo que um paciente de câncer sobrevive depois do diagnóstico.
 - (d) O número de abortos prévios que uma mãe grávida teve.
8. Nas páginas seguintes são listados os gastos com os cuidados com a saúde *per capita* em 1989 para 23 das 24 nações que constituem a Organização para Cooperação Econômica e Desenvolvimento [10]. (Os gastos *per capita* para a Turquia não estão disponíveis.)

Nação	Gastos per Capita (US\$)
Austrália	1.032
Áustria	1.093
Bélgica	980
Inglatera	836
Canadá	1.683
Dinamarca	912
Finlândia	1.067
França	1.274
Alemanha	1.232
Grécia	371
Islândia	1.353
Irlanda	658
Itália	1.050
Japão	1.035
Luxemburgo	1.193
Países Baixos	1.135
Nova Zelândia	820
Noruega	1.234
Portugal	464
Espanha	644
Suécia	1.361
Suíça	1.376
Estados Unidos	2.354

- (a) Ordene esses países de acordo com os gastos *per capita* com a saúde.
- (b) Construa um histograma para os valores dos gastos *per capita*.
- (c) Descreva a forma do histograma.

9. A tabela abaixo categoriza 10.614.000 visitas ao consultório de especialistas de doenças cardiovasculares nos Estados Unidos por duração de cada visita [15]. Uma duração de 0 (zero) minuto implica que o paciente não teve contato direto com o especialista.

Duração (minutos)	Número de visitas (milhares)
0	390
1–5	227
6–10	1.023
11–15	3.390
16–30	4.431
31–60	968
61+	185
Total	10.614

Pode-se fazer a afirmação de que as visitas a consultórios de especialistas de doenças cardiovasculares têm duração mais freqüente entre 16 e 30 minutos. Você concorda com essa afirmação? Por quê? Justifique.

10. A distribuição de freqüências a seguir exibe os números de casos pediátricos de Aids registrados nos Estados Unidos entre 1983 e 1989 [9].

Ano	Número de casos
1983	122
1984	250
1985	455
1986	848
1987	1.412
1988	2.811
1989	3.098

Construa um gráfico de barras que mostre o número de casos por ano. O que o gráfico lhe conta sobre a Aids pediátrica nesse período de tempo?

11. Abaixo estão relacionados os números de pessoas que foram executadas nos Estados Unidos em cada ano, desde a decisão de 1976 da Suprema Corte que permitiu a realização da pena de morte [16].

Ano	Número de execuções	Ano	Número de execuções
1976	0	1986	18
1977	1	1987	25
1978	0	1988	11
1979	2	1989	16
1980	0	1990	23
1981	1	1991	14
1982	2	1992	31
1983	5	1993	38
1984	21	1994	28
1985	18		

Use esses dados para criar um gráfico de barras de execuções por ano. Como o número de execuções variou desde 1976?

12. Um estudo foi conduzido para examinar o sexo e as diferenças raciais entre os indivíduos a partir de 65 anos de idade que sofreram fratura dos quadris entre 1984 e 1987 [17]. Os dados que resumem todas as altas de hospital registradas pelo programa Medicare aparecem abaixo.

Grupo de Idade	Homens Brancos	Homens Negros	Mulheres Brancas	Mulheres Negras
65–74	36.473	2.295	103.105	3.425
75–84	62.513	2.902	233.047	6.819
85–94	40.975	1.659	189.459	5.968
95	4.088	208	18.247	934

- (a) Com base nesses dados, construa um gráfico de barras empilhadas que mostre o número de altas de hospital que acompanham a fratura de quadril por grupo de idade. (Cada barra deve consistir de quatro seções separadas que representam homens brancos, homens negros, mulheres brancas e mulheres negras.)
- (b) Como o número total de fraturas de quadril varia com a idade?
- (c) Com base no gráfico, o que você conclui sobre a relação entre o sexo e a fratura de quadril?
13. Em uma investigação dos fatores de risco para as doenças cardiovasculares, os níveis séricos de cotinina — produto metabólico da nicotina — foram registrados para um grupo de fumantes e um grupo de não-fumantes[18]. As distribuições de freqüências correspondentes são mostradas abaixo.

Nível de cotinina (ng/ml)	Fumantes	Não-Fumantes
0–13	78	3.300
14–49	133	72
50–99	142	23
100–149	206	15
150–199	197	7
200–249	220	8
250–299	151	9
300+	412	11
Total	1.539	3.445

- (a) É correto comparar as distribuições dos níveis de cotinina para fumantes e não-fumantes, com base nas freqüências absolutas em cada intervalo? Por quê?
- (b) Calcule as freqüências relativas das leituras dos níveis séricos de cotinina para cada grupo.
- (c) Construa um par de polígonos de freqüência.
- (d) Descreva a forma de cada polígono. O que você pode dizer sobre a distribuição de níveis de cotinina registrados em cada grupo?
- (e) Para todos os indivíduos nesse estudo, o status do fumo é auto-registrado. Você acha que algum dos indivíduos pode estar mal classificado? Por quê?

14. As freqüências relativas das concentrações de chumbo no sangue para dois grupos de trabalhadores no Canadá — um examinado em 1979 e outro em 1987 — são exibidas abaixo [19].

Chumbo no sangue ($\mu\text{g}/\text{dl}$)	1979 (%)	1987(%)
<20	11,5	37,8
20–29	12,1	14,7
30–39	13,9	13,1
40–49	15,4	15,3
50–59	16,5	10,5
60–69	12,8	6,8
70–79	8,4	1,4
≥ 80	9,4	0,4

(a) Em que ano os trabalhadores tendem a ter níveis mais baixos de chumbo no sangue?

(b) Calcule as freqüências relativas acumuladas para cada um dos grupos de trabalhadores. Use esses dados para construir um par de polígonos de freqüência acumulada.

(c) Para que grupo a distribuição de níveis de chumbo no sangue é estocasticamente maior?

15. Os números registrados de nascidos vivos nos Estados Unidos para cada mês no período de janeiro de 1991 a dezembro de 1992 são apresentados abaixo [20].

Mês 1991	Número (milhares)	Mês 1992	Número (milhares)
janeiro	325	janeiro	334
fevereiro	312	fevereiro	304
março	346	março	360
abril	340	abril	330
maio	355	maio	361
junho	342	junho	333
julho	358	julho	352
agosto	346	agosto	350
setembro	365	setembro	357
outubro	355	outubro	345
novembro	324	novembro	332
dezembro	342	dezembro	325

(a) Construa um gráfico de linha que exiba o número registrado de nascidos vivos no tempo.

(b) Com base nesse período de dois anos, você acha que o número de nascidos vivos segue um padrão sazonal nos Estados Unidos?

16. Uma distribuição de freqüências para os níveis séricos de zinco de 462 homens entre as idades de 15 e 17 anos é exibida a seguir [21]. As observações estão armazenadas no disco anexo no conjunto de dados *serzinc* (Apêndice B, Tabela B.1). As 462 medidas séricas de zinco, que foram registradas em microgramas por decilitro, estão salvas sob a variável de nome *zinc*.

Estados Unidos, idade 15-17 anos

Nível Sérico de Zinco (mg/dl)	Número de Homens
50–59	6
60–69	35
70–79	110
80–89	116
90–99	91
100–109	63
110–119	30
120–129	5
130–139	2
140–149	2
150–159	2

- (a) Calcule a freqüência relativa associada a cada um dos intervalos na tabela. O que você pode concluir sobre essa distribuição de níveis séricos de zinco?
- (b) Elabore um histograma dos dados. As observações devem ser divididas nos 11 intervalos de igual largura especificados na distribuição de freqüências anterior, de 50-59 a 150-159 $\mu\text{g}/\text{dl}$.
- (c) Descreva a forma do histograma.
17. As porcentagens de bebês com baixo peso ao nascer, em vários países do mundo, estão contidas no conjunto de dados unicef [22] (Apêndice B, Tabela B.2). As próprias medidas estão salvas sob a variável de nome lowbwt.
- (a) Construa um box plot para as porcentagens de bebês com baixo peso ao nascer.
- (b) Os dados parecem assimétricos? Em caso positivo, são assimétricos à direita ou à esquerda?
- (c) Os dados contêm alguma observação atípica?
18. Os números de residentes assistidos por enfermagem doméstica com pelo menos 65 anos por 1.000 habitantes para cada estado nos Estados Unidos estão contidos no conjunto de dados nurshome [23] (Apêndice B, Tabela B.3). Os nomes dos estados estão salvos sob o nome da variável state e o número de residentes assistidos por enfermagem caseira por 1.000 habitantes sob a variável de nome resident.
- (a) Que estado tem o menor número de residentes assistidos por enfermagem doméstica por 1.000 habitantes com 65 anos e acima? Que estado tem o maior número? Que fator poderia influenciar a substancial variabilidade entre os diferentes estados?
- (b) Construa um box plot para o número de residentes assistidos por enfermagem doméstica por 1.000 habitantes.
- (c) As observações são simétricas ou assimétricas? Existe algum estado que poderia ser considerado atípico?
- (d) Exiba o número de residentes assistidos por enfermagem doméstica por 1.000 habitantes usando um histograma. Você acha este gráfico mais ou menos informativo do que o box plot?
19. As concentrações declaradas de alcatrão e de nicotina para 35 marcas de cigarros canadenses estão armazenadas em um conjunto de dados chamado cigaret [24] (Apêndice B, Tabela B.4). As concentrações de alcatrão por cigarro em miligramas estão salvos sob a variável de nome tar e as correspondentes concentrações de nicotina sob o nome nicotine.
- (a) Produza um gráfico de dispersão unidimensional das concentrações declaradas de alcatrão por cigarro. Esteja certo de identificar as posições as quais duas ou mais medidas tenham os mesmos valores e portanto se sobreponem.
- (b) Descreva a distribuição de valores.
- (c) Construa um gráfico de dispersão bidimensional da concentração de alcatrão *versus* a concentração de nicotina. Nomeie os eixos apropriadamente.
- (d) Parece existir uma relação entre essas duas quantidades?
20. As taxas de nascimento por mulheres não-casadas nos Estados Unidos de 1940–1992 estão salvos no conjunto de dados brate [25] (Apêndice B, B.5). Os anos estão salvos sob a variável de nome year e os números de nascidos vivos por 1.000 mulheres não-casadas entre 15 e 44 anos estão salvos sob o nome birthrt.
- (a) Crie um gráfico de linha que exiba as taxas de nascimento no tempo para mulheres não-casadas.
- (b) Muitas pessoas acreditam que o grande número de crianças nascidas de mães não-casadas é um problema relativamente recente em nossa sociedade. Depois de ver o gráfico de linha, você concorda?

3

Medidas-Resumo Numéricas

No capítulo anterior, estudamos as tabelas e os gráficos como métodos para se organizar, resumir visualmente e exibir um conjunto de dados. Embora essas técnicas sejam extremamente úteis, não permitem fazer afirmações concisas e quantitativas que caracterizem a distribuição dos valores como um todo. Para fazê-lo, contamos com as *medidas-resumo numéricas*. Juntos, os vários tipos de estatísticas descritivas podem fornecer um grande volume de informações de um conjunto de observações.

3.1 Medidas de Tendência Central

A característica de um conjunto de dados mais comumente investigada é o seu centro ou o ponto ao redor do qual as observações tendem a se agrupar. Suponha que estejamos interessados em examinar o efeito da inalação de ozônio e dióxido de enxofre por adolescentes que sofrem de asma. As medidas iniciais do volume expiratório forçado em um segundo para 13 indivíduos envolvidos nesse estudo [1] estão listadas na Tabela 3.1. Lembre-se de que FEV_i é o volume de ar que pode ser expelido dos pulmões depois de um segundo de esforço constante. Anteriormente à investigação dos efeitos dos poluentes na função pulmonar, poderíamos determinar o valor típico do FEV_i antes da exposição dos indivíduos desse grupo.

3.1.1 Média

A medida de tendência central mais freqüentemente usada é a *média aritmética*, calculada com a soma de todas as observações de um conjunto de dados e divisão do resultado pelo número total de medidas. Na Tabela 3.1, por exemplo, temos 13 observações. Se x é usado para representar FEV_i, $x_1 = 2,30$ representa a primeira observação da série; $x_2 = 2,15$ a segunda; até $x_{13} = 3,38$. Em geral, x_i refere-se a uma medida FEV_i simples, na qual i pode tomar qualquer valor de 1 até n , o número total de observações no grupo. A média das observações no conjunto de dados — representada por \bar{x} ou x -barra — é

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Note que usamos alguma notação matemática. A letra grega maiúscula sigma, Σ , é o símbolo para a somatória. A expressão $\sum_{i=1}^n x_i$ significa que temos de somar os valores de todas as observações no grupo, de x_1 até x_n . Quando Σ aparece no texto, os limites da somatória estão colocados ao seu lado; quando é exibido separadamente em uma equação, os limites

TABELA 3.1

Volumes expiratórios forçados em um segundo para 13 adolescentes que sofrem de asma

Indivíduo	FEV ₁ (litros)
1	2,30
2	2,15
3	3,50
4	2,60
5	2,75
6	2,82
7	4,05
8	2,25
9	2,68
10	3,00
11	4,02
12	2,85
13	3,38

estão acima e abaixo dele. Ambas as representações denotam exatamente a mesma coisa. Em casos nos quais está clara a intenção de somar todas as observações em um conjunto de dados, os limites podem ser totalmente abandonados. Para os dados FEV₁,

$$\begin{aligned}\bar{x} &= \frac{1}{13} \sum_{i=1}^{13} x_i \\ &= \left(\frac{1}{13} \right) (2,30 + 2,15 + 3,50 + 2,60 + 2,75 + 2,82 + 4,05 \\ &\quad + 2,25 + 2,68 + 3,00 + 4,02 + 2,85 + 3,38) \\ &= \frac{38,35}{13} \\ &= 2,95 \text{ litros.}\end{aligned}$$

A média pode ser usada como uma medida-resumo tanto para as medidas discretas como para as contínuas. Geralmente, no entanto, não é apropriada para os dados nominais nem para os ordinais. Lembre-se de que, para esses tipos de dados, os números são meramente rótulos, de modo que mesmo se representássemos os tipos sanguíneos O, A, B e AB pelos números 1, 2, 3 e 4, a média do tipo sanguíneo de 1,8 não tem significado. Uma exceção a essa regra é aplicada quando temos dados dicotômicos e os dois resultados possíveis são representados pelos valores 0 e 1. Nessa situação, a média das observações é igual à proporção de 1s no conjunto de dados. Por exemplo, suponha que queiramos conhecer a proporção de adolescentes asmáticos do sexo masculino no estudo previamente mencionado. Listados na Tabela 3.2 estão os dados dicotômicos relevantes; o valor 1 representa um homem e 0 designa uma mulher. Se calcularmos a média dessas observações, encontraremos

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ &= \left(\frac{1}{13} \right) (0 + 1 + 1 + 0 + 0 + 1 + 1 + 1 + 0 + 1 + 1 + 1 + 0) \\ &= \frac{8}{13} \\ &= 0.615.\end{aligned}$$

TABELA 3.2

Indicadores de gênero
para 13 adolescentes asmáticos

Indivíduo	Sexo
1	0
2	1
3	1
4	0
5	0
6	1
7	1
8	1
9	0
10	1
11	1
12	1
13	0

Conseqüentemente, 61,5% dos indivíduos estudados são homens.

O método pelo qual a média é calculada considera a magnitude de cada observação em um conjunto de dados. O que acontece quando uma observação tem um valor muito diferente dos outros? Suponha, por exemplo, que tenhamos registrado os dados da Tabela 3.1 em um disquete de computador e que ele tenha sido accidentalmente submetido ao raio X no aeroporto; a medida de FEV₁ do indivíduo 11 está agora registrada como 40,2 em vez de 4,02. A média de FEV₁ de todos os 13 indivíduos seria calculada como

$$\begin{aligned}\bar{x} &= \left(\frac{1}{13} \right) (2,30 + 2,15 + 3,50 + 2,60 + 2,75 + 2,82 + 4,05 \\ &\quad + 2,25 + 2,68 + 3,00 + 40,2 + 2,85 + 3,38) \\ &= \frac{74,53}{13} \\ &= 5,73 \text{ litros},\end{aligned}$$

que é quase duas vezes maior do que antes. Obviamente, a média é extremamente sensível aos valores não-usuais. Nesse exemplo em particular, teríamos justamente questionado uma medida de FEV₁ de 40,2 litros e corrigido o erro ou isolado essa observação. Entretanto, o erro poderia não ser tão óbvio ou a observação não-usual poderia não ser absolutamente um erro. Como nossa intenção é caracterizar um grupo inteiro de indivíduos, poderíamos usar uma medida-resumo que não seja tão sensível a cada uma das observações.

3.1.2 Mediana

Uma medida de tendência central que não é sensível ao valor de cada medida é a mediana, que pode ser usada como uma medida-resumo para as observações ordinais, assim como para os dados discretos e contínuos. A *mediana* é definida como o 50º percentil de um conjunto de medidas; se uma lista de observações é ordenada da menor até a maior, metade dos valores são maiores ou iguais à mediana, enquanto a outra metade é menor ou igual a ela. Conseqüentemente, se um conjunto de dados contém um total de n observações, no qual n é ímpar, a mediana é o valor do meio ou a $[(n + 1)/2]$ —ésima medida; se n for par, a mediana é usualmente tomada como a média dos dois valores mais centrais do intervalo, a

$(n/2)$ -ésima e $[(n/2) + 1]$ -ésima observações. Se ordenássemos as 13 medidas FEV₁ listadas na Tabela 3.1, resultaria a seguinte seqüência:

2,15, 2,25, 2,30, 2,60, 2,68, 2,75, 2,82, 2,85, 3,00, 3,38, 3,50, 4,02, 4,05.

Como há um número ímpar de observações na lista, a mediana é a $(13+1)/2 = 7^{\text{a}}$ observação ou 2,82. Sete das observações são menores ou iguais a 2,82 e sete são maiores ou iguais a 2,82.

O cálculo da mediana leva em consideração somente a ordenação e a magnitude relativa das observações em um conjunto de dados. Na situação em que a FEV₁ do indivíduo 11 foi registrada como 40,2 em vez de 4,02, a ordenação das medidas mudaria somente um pouco:

2,15, 2,25, 2,30, 2,60, 2,68, 2,75, 2,82, 2,85, 3,00, 3,38, 3,50, 4,05, 40,2.

Como resultado, a FEV₁ mediana permaneceria em 2,82 litros. Diferentemente da média, diz-se que a mediana é *robusta*, ou seja, muito menos sensível aos pontos não-usuais dos dados.

3.1.3 Moda

Uma terceira medida de tendência central é a moda, que pode ser usada como medida-resumo para todos os tipos de dados. A *moda* de um conjunto de dados é a observação que ocorre mais freqüentemente. Os dados FEV₁ na Tabela 3.1 não têm uma única moda, pois cada um dos valores ocorre somente uma vez. A moda para os dados dicotômicos da Tabela 3.2 é 1. Esse valor aparece oito vezes, enquanto 0 aparece somente cinco.

A melhor medida de tendência central para um específico conjunto de dados depende freqüentemente do modo pelo qual os valores estão distribuídos. Se são simétricos e *unimodais* — significando que, se desenhássemos um histograma ou um polígono de freqüência, haveria somente um pico, como na distribuição aplanaada mostrada na Figura 3.1(a) — então a média, a mediana e a moda deveriam ser aproximadamente as mesmas. Se a distribuição de valores é simétrica, mas *bimodal*, de modo que o polígono de freqüência correspon-

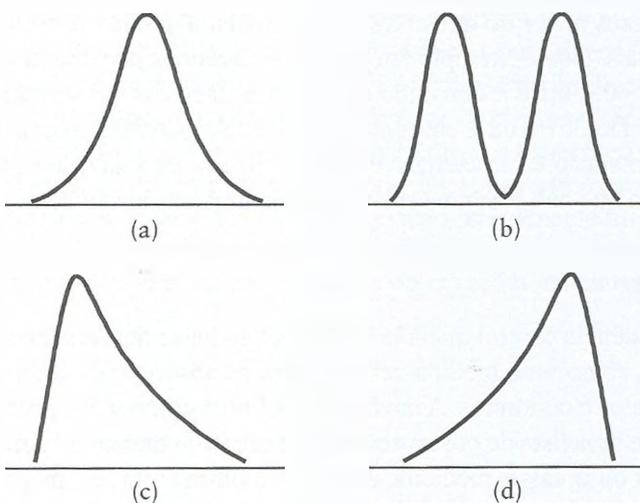


FIGURA 3.1
Possíveis distribuições dos valores de dados.

dente tivesse dois picos, como na Figura 3.1(b), então a média e a mediana seriam, mais uma vez, aproximadamente as mesmas. Observe, no entanto, que esse valor comum se encontraria entre os dois picos e seria, portanto, uma medida improvável de ocorrer. Uma distribuição bimodal indica freqüentemente que a população da qual os valores são tomados consiste realmente de dois subgrupos distintos que diferem na característica medida; nessa situação, poderia ser melhor registrar duas modas em vez da média e da mediana ou tratar os dois subgrupos separadamente. Os dados na Figura 3.1(c) são assimétricos à direita e os da Figura 3.1(d) à esquerda. Quando os dados não são simétricos, a mediana é freqüentemente a melhor medida de tendência central. Por ser sensível às observações extremas, a média é puxada em direção dos valores atípicos e, consequentemente, poderia terminar excessivamente inflada ou reduzida em excesso. Note-se que, quando os dados são assimétricos à direita, a média se encontra à direita da mediana e, quando são assimétricos à esquerda, a média se encontra à esquerda da mediana.

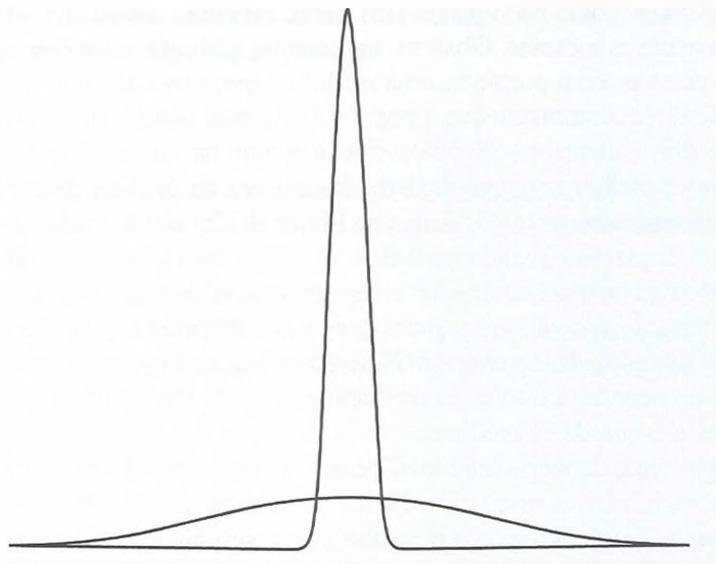
Independentemente da medida de tendência central usada em uma situação particular, pode ser enganoso assumir que esse valor seja representativo de todas as observações do grupo. Um exemplo que ilustra esse ponto foi incluído no episódio de 17 de novembro de 1991 do popular programa de notícias “60 Minutes”. O programa continha um segmento sobre dieta e mortalidade que contrastava as experiências francesas e americanas. Embora a dieta francesa fosse extremamente alta em gordura e colesterol, a França tem uma taxa muito mais baixa de doenças do coração do que os Estados Unidos. Essa diferença paradoxal foi atribuída ao hábito francês de beber vinho — tinto, em particular — nas refeições. Estudos têm sugerido que o consumo moderado de álcool pode reduzir o risco de doenças do coração. Enquanto o consumo de vinho *per capita* na França é um dos mais altos do mundo, o dos Estados Unidos é um dos mais baixos; o programa sugeriu fortemente que os franceses bebem uma quantidade moderada de vinho a cada dia, talvez dois ou três copos. No entanto, a realidade pode ser bastante diferente. De acordo com uma pesquisa da indústria de vinho conduzida em 1990, mais da metade dos adultos franceses absolutamente nunca bebem vinho [2]. Dos que o fazem, somente 28% dos homens e 11% das mulheres o bebem diariamente. Obviamente a distribuição é muito mais variável do que sugeria o “valor típico”. Lembre-se que, quando resumimos um conjunto de dados, informações são sempre perdidas. Assim, embora seja útil saber onde o centro de um conjunto de dados se encontra, usualmente essa informação não é suficiente para caracterizar uma distribuição inteira de medidas.

Como outro exemplo, em cada um dos valores das duas distribuições muito diferentes mostradas na Figura 3.2, a média, a mediana e a moda são iguais. Para saber o quanto nossa medida de tendência central realmente é boa, necessitamos de alguma idéia sobre a variação entre as medidas. Todas as observações tendem a ser bastante similares e consequentemente se encontram próximas do centro ou estão dispersas por um amplo intervalo de valores?

3.2 Medidas de Dispersão

3.2.1 Amplitude

Um número que pode ser usado para descrever a variabilidade de um conjunto de dados é conhecido como amplitude. Define-se *amplitude* de um grupo de medidas como a diferença entre a maior observação e a menor. Embora a amplitude seja fácil de calcular, sua utilidade é limitada, pois considera somente os valores extremos de um conjunto de dados e não a maioria das observações. Em consequência, tal como a média, é altamente sensível aos valores excepcionalmente grandes ou pequenos. A amplitude para os dados FEV₁ na Tabela 3.1 é $4.05 - 2.15 = 1.90$ litro. Se a FEV₁ do indivíduo 11 fosse registrada como 40,2 ao invés de

**FIGURA 3.2**

Duas distribuições com médias, medianas e modas idênticas.

4,02, a amplitude seria $40,2 - 2,15 = 38,05$ litros, um valor 20 vezes maior. As amplitudes dos valores para a concentração anual de dióxido de enxofre no ar de diversas cidades ao redor do mundo são apresentadas na Figura 3.3 [3].

3.2.2. Intervalo Interquartil

Uma segunda medida de variabilidade — que não é tão facilmente influenciada por valores extremos — é chamada de intervalo interquartil, calculado subtraindo-se o 25º percentil dos dados do 75º percentil; consequentemente, ele engloba 50% do meio das observações. (Lembre-se de que o 25º percentil e o 75º percentil do conjunto de dados são chamados de quartis). Para os dados FEV₁ na Tabela 3.1, o 75º percentil é 3,38. Note-se que três observações são maiores do que esse valor e nove são menores. Analogamente, o 25º percentil é 2,60. Consequentemente, o intervalo interquartil é $3,38 - 2,60 = 0,78$ litro.

Se um computador não estiver disponível, existem regras para se encontrar o k -ésimo percentil de um conjunto de dados, assim como para se encontrar a mediana. Nesse caso, a regra usada depende de o número de observações n ser par ou ímpar. Novamente começamos por ordenar as medidas desde a menor até a maior. Se $nk/100$ é um inteiro, o k -ésimo percentil dos dados é a média da $(nk/100)$ -ésima e $(nk/100 + 1)$ -ésima observações. Se $nk/100$ não for um inteiro, o k -ésimo percentil é a $(j + 1)$ -ésima medida, no qual j é o maior inteiro menor que $nk/100$. Para encontrar o 25º percentil das 13 medidas FEV₁, por exemplo, primeiramente notamos que $13(25)/100 = 3,25$ não é um inteiro. Em decorrência, o 25º percentil é a $3 + 1 = 4^{\text{a}}$ medida (como 3 é o maior inteiro menor que 3,25) ou 2,60 litros. Analogamente, $13(75)/100 = 9,75$ não é um inteiro e o 75º percentil é a $9 + 1 = 10^{\text{a}}$ maior medida ordenada crescentemente ou 3,38 litros. Os intervalos interquartis — assim como as médias, medianas e amplitudes — dos números de episódios de vários tipos de comportamento sexual praticados por homens homossexuais antes e depois de aprenderem sobre a Aids são apresentados na Figura 3.4 [4]. Note-se que as médias são maiores do que as medianas em todos os casos, indicando que os dados são assimétricos e que existem diversos valores extraordinariamente altos que levam as médias a serem infladas. A diferença entre as médias e as medianas é menos evidente depois que os homens aprenderam sobre a Aids; a educação sobre o vírus parece que teve um efeito restritivo no comportamento sexual, especialmente nos casos extremos.

O que está mostrado é a amplitude dos valores anuais em locais individuais e a média composta de 5 anos para a cidade.

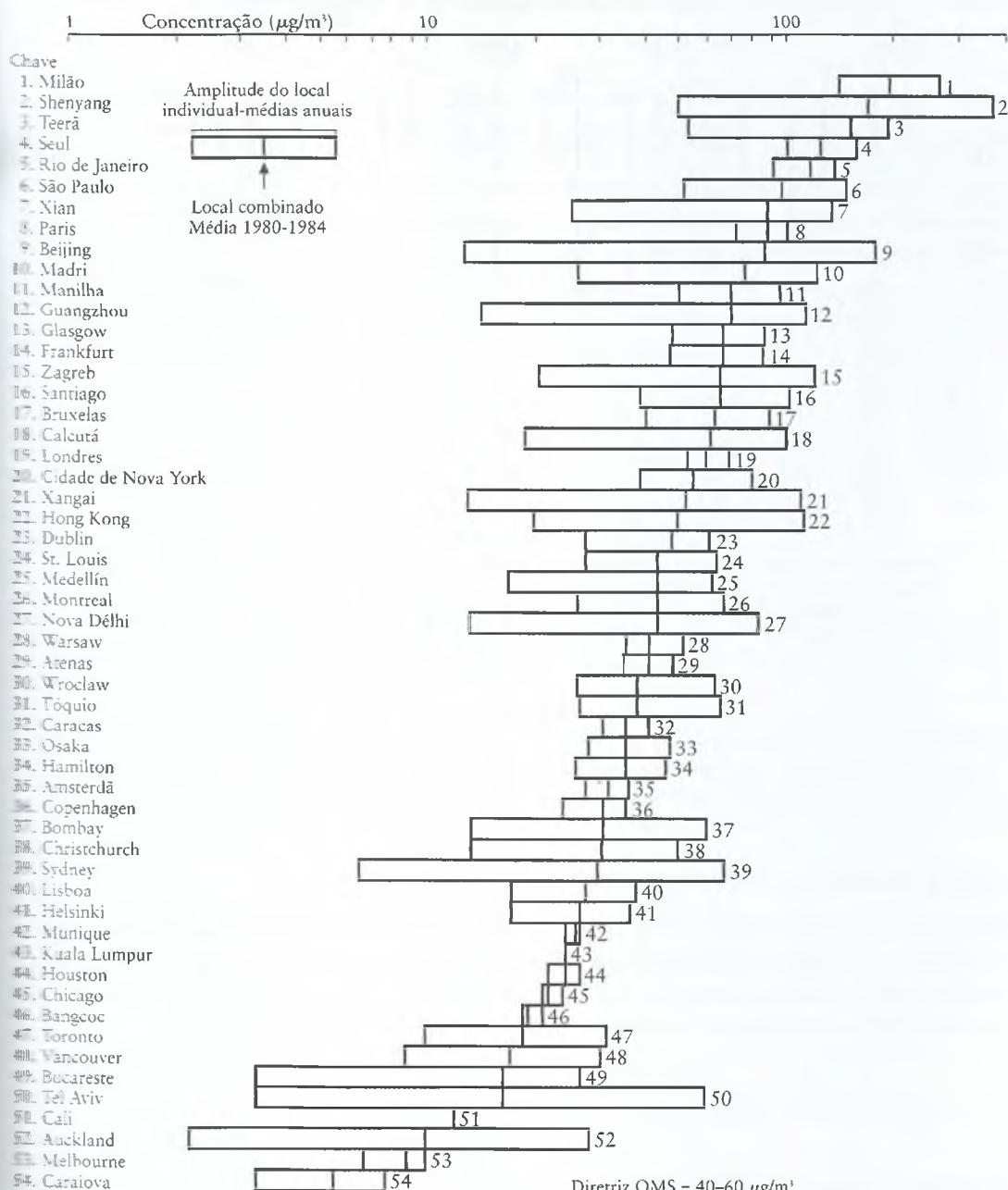
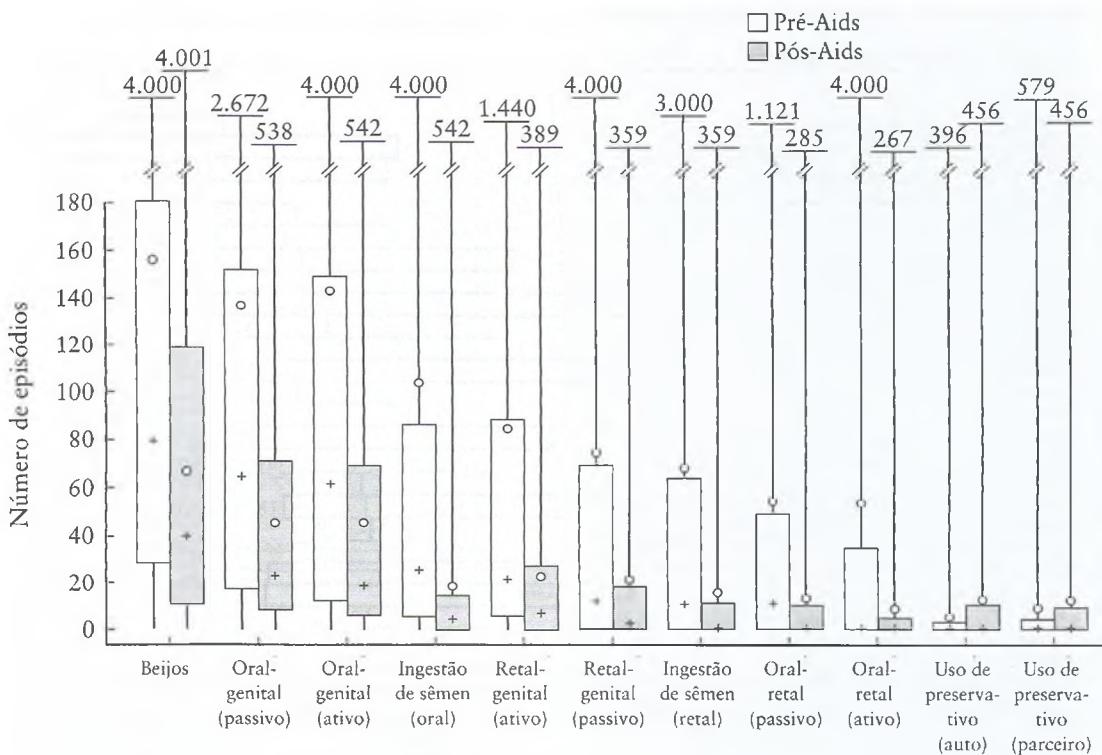


FIGURA 3.3
Resumo das médias de dióxido de enxofre, 1980–1984.

**FIGURA 3.4**

A freqüência anual mediana (+) e a freqüência anual média (o), o intervalo interquartil (caixas) e a amplitude do comprometimento em atos específicos durante o ano antes de se ouvir sobre a AIDS e depois de se ouvir sobre a Aids.

3.2.3 Variância e Desvio-Padrão

Outra medida de dispersão comumente usada para um conjunto de dados é conhecida como a variância. A variância quantifica a variabilidade ou o espalhamento ao redor da média das medidas. Para quantificar essa variabilidade, poderíamos simplesmente tentar calcular a distância média das observações individuais a partir de \bar{x} , ou

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}).$$

No entanto, pode-se mostrar matematicamente que a expressão $\sum_{i=1}^n (x_i - \bar{x})$ é sempre igual a zero. Por definição, a soma dos desvios da média de todas as observações menores que \bar{x} é igual à soma dos desvios de todas as observações maiores que \bar{x} ; como consequência, essas duas somas cancelam-se mutuamente. Para eliminar esse problema, poderíamos, ao contrário, tirar a média dos valores absolutos dos desvios da média, todos positivos. Enquanto não há nada conceitualmente errado nessa abordagem, a medida-resumo resultante não tem certas propriedades estatísticas importantes e é raramente vista na literatura. Um procedimento mais amplamente usado é elevar ao quadrado os desvios da média — lembre-se de que um valor absoluto elevado ao quadrado é sempre positivo — e então encontrar a média dessas distâncias elevadas ao quadrado. Essa medida-resumo é a variância das observações.

Mais explicitamente, a variância é calculada ao se subtrair a média de um conjunto de valores de cada uma das observações, elevar ao quadrado esses desvios, somá-los e dividir

A soma pelo número de observações do conjunto de dados menos 1. Ao representar a variância por s^2 :

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2.$$

A razão para se dividir por $n-1$ em vez de n será discutida mais tarde, no Capítulo 9; podemos, não obstante, imaginar a variância como uma média de desvios elevados ao quadrado. Para as 13 medidas FEV, apresentadas na Tabela 3.1, a média é 2,95 litros e os desvios elevados ao quadrado a partir da média estão dados abaixo.

Indivíduo	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	2,30	-0,65	0,4225
2	2,15	-0,80	0,6400
3	3,50	0,55	0,3025
4	2,60	-0,35	0,1225
5	2,75	-0,20	0,0400
6	2,82	-0,13	0,0169
7	4,05	1,10	1,2100
8	2,25	-0,70	0,4900
9	2,68	-0,27	0,0729
10	3,00	0,05	0,0025
11	4,02	1,07	1,1449
12	2,85	-0,10	0,0100
13	3,38	0,43	0,1849
Total	38,35	0,00	4,6596

Logo, a variância é

$$\begin{aligned} s^2 &= \frac{1}{(13-1)} \sum_{i=1}^{13} (x_i - 2,95)^2 \\ &= \frac{4,6596}{12} \\ &= 0,39 \text{ litros}^2. \end{aligned}$$

O *desvio-padrão* de um conjunto de dados é a raiz quadrada positiva da variância. Assim, para as 13 medidas FEV, acima, o desvio-padrão é igual a

$$\begin{aligned} s &= \sqrt{s^2} \\ &= \sqrt{0,39 \text{ litros}^2} \\ &= 0,62 \text{ litros}. \end{aligned}$$

Na prática, o desvio-padrão é usado mais freqüentemente do que a variância. A razão disso é que o desvio-padrão tem a mesma unidade de medida que a média, em vez da unidade elevada ao quadrado. Em uma comparação de dois grupos de dados, o grupo com o menor desvio-padrão tem as observações mais homogêneas; o grupo com o maior desvio-padrão exibe maior variabilidade. A magnitude real do desvio-padrão depende dos valores do conjunto de dados – o que é grande para um grupo de dados pode ser pequeno para um outro. Além disso, como o desvio-padrão tem unidade de medida, não tem sentido comparar desvios-pa-

drão para duas quantidades não-relacionadas. Juntos, a média e o desvio-padrão de um conjunto de dados podem ser usados para resumir as características da distribuição inteira de valores. Veremos como isso funciona na Seção 3.4.

3.2.4 Coeficiente de Variação

Ainda que não faça sentido comparar desvios-padrão, é possível comparar a variabilidade entre dois ou mais conjuntos de dados que representam quantidades variadas com diferentes unidades de medida, ao usarmos uma medida-resumo numérica conhecida como o coeficiente de variação. O *coeficiente de variação* relaciona o desvio-padrão de um conjunto de valores à sua média; ele é a razão entre s e multiplicada por 100 e é, portanto, uma medida de variabilidade relativa. Como o desvio-padrão e a média partilham as mesmas unidades de medida, as unidades cancelam-se e deixam o coeficiente de variação adimensional. O coeficiente de variação para os dados FEV é

$$\begin{aligned} CV &= \frac{s}{\bar{x}} \times 100\% \\ &= \frac{0,62}{2,95} \times 100\% \\ &= 21,0\%. \end{aligned}$$

É difícil avaliar se esse valor é grande ou pequeno; o coeficiente de variação é mais útil para se comparar dois ou mais conjuntos de dados. Como é independente das unidades de medida, pode ser usado para avaliar a variação relativa entre quaisquer dois conjuntos de observações. Embora o coeficiente de variação seja ainda usado como uma medida-resumo em alguns círculos, suas propriedades estatísticas não são muito boas. Como resultado, seu uso está diminuindo e deve ser desaconselhado.

3.3 Dados Agrupados

Se desejarmos contar o dinheiro que temos em nossos bolsos, existem duas maneiras para fazê-lo. A primeira é somar consecutivamente os valores das moedas, conforme as pegamos. A segunda é agrupar as moedas por denominação e em seguida multiplicarmos o valor de cada uma pelo número de moedas daquela denominação e, finalmente, somar esses valores. Por exemplo: se temos três moedas de um centavo, quatro de cinco, duas de dez e uma de 25, temos um total de

$$\begin{aligned} 3(1) + 4(5) + 2(10) + 1(25) &= 3 + 20 + 20 + 25 \\ &= 68 \text{ centavos}. \end{aligned}$$

O mesmo procedimento pode ser usado para somar qualquer conjunto de observações. Por exemplo, considere os dados na Tabela 3.3 [5]. Entre os pacientes com doença falciforme — uma forma hereditária de anemia — as transfusões de sangue regulares são freqüentes para evitar derrames recorrentes após evento cerebrovascular inicial. A terapia de transfusão a longo prazo, no entanto, tem riscos associados próprios e nem sempre é aconselhável. A Tabela 3.3 lista as durações de terapia para dez pacientes inscritos em um estudo que investiga os efeitos da interrupção das transfusões de sangue. Poderíamos estar interessados em determinar a média desses valores.

3.3.1 Média de Dados Agrupados

A técnica-padrão para se encontrar a média das observações é adicionar os valores e dividir por $n = 10$. Nesse caso, obtemos

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ &= \left(\frac{1}{10}\right)(12 + 11 + 12 + 6 + 11 + 11 + 8 + 5 + 5 + 5) \\ &= \frac{86}{10} \\ &= 8,6 \text{ anos.}\end{aligned}$$

TABELA 3.3

Duração da terapia de transfusão para dez pacientes com doença falciforme.

Indivíduo	Duração (anos)
1	12
2	11
3	12
4	6
5	11
6	11
7	8
8	5
9	5
10	5

Alternativamente, poderíamos obter a soma das medidas ao se agrupar primeiro as observações com iguais valores; note-se que existem três 5s, um 6, um 8, três 11s e dois 12s. Logo,

$$\begin{aligned}\sum_{i=1}^{10} x_i &= 3(5) + 1(6) + 1(8) + 3(11) + 2(12) \\ &= 15 + 6 + 8 + 33 + 24 \\ &= 86,\end{aligned}$$

e

$$\begin{aligned}\bar{x} &= \frac{86}{10} \\ &= 8,6 \text{ anos.}\end{aligned}$$

Obtemos a mesma média, independentemente do método usado.

A técnica de agrupar medidas de valores iguais antes de calcular a média tem uma vantagem distinta sobre o método-padrão: esse procedimento pode ser aplicado para dados que tenham sido resumidos na forma de uma distribuição de freqüências, os quais são chamados freqüentemente de *dados agrupados*. Mesmo que as observações originais não estejam mais

disponíveis — ou talvez nunca estarão, se os dados foram coletados no formato agrupado desde o início — poderíamos ainda estar interessados em calcular as medidas-resumo numéricas para os dados. No entanto, surge um obstáculo que é o nosso desconhecimento dos valores das observações individuais; apesar disso, somos capazes de determinar o número de medidas situadas em cada um dos intervalos especificados. Essa informação pode ser usada para calcular uma *média de dados agrupados*.

Para calcular a média de um conjunto de dados arranjados como uma distribuição de freqüências, começamos por assumir que todos os valores situados em um intervalo particular são iguais ao seu ponto médio. Lembre-se dos dados séricos de colesterol para o grupo de 25 a 34 anos que foram apresentados na Tabela 2.6 [6], os quais estão reproduzidos na Tabela 3.4. O primeiro intervalo contém os valores que variam de 80 até 119 mg/100 ml, com um ponto médio de 99,5. Assumimos, em consequência, que todas as 13 medidas dentro dele tomam o valor 99,5 mg/100 ml. Analogamente, assumimos que as 150 observações no segundo intervalo, 120 até 159 mg/100 ml, tomam o valor de 139,5 mg/100 ml. Por fazermos essas suposições, nossos cálculos são somente aproximados. Além disso, os resultados mudariam se agrupássemos os dados de modo diferente.

Para encontrarmos a média de dados agrupados, primeiro somamos as medidas, multiplicando o ponto médio de cada intervalo pela freqüência correspondente, e somando esses produtos; então dividimos a soma pelo número total de valores. Logo,

$$\bar{x} = \frac{\sum_{i=1}^8 m_i f_i}{\sum_{i=1}^8 f_i}$$

TABELA 3.4

Freqüências absolutas de níveis séricos de colesterol para homens dos Estados Unidos, com idade entre 25 e 34 anos, 1976-1980.

Nível de Colesterol (mg/100 ml)	Número de Homens
80–119	13
120–159	150
160–199	442
200–239	299
240–279	115
280–319	34
320–359	9
360–399	5
Total	1.067

onde k é o número de intervalos na tabela, m_i é o ponto médio do i -ésimo intervalo e f_i é a freqüência associada com o i -ésimo intervalo. Note-se que a soma das freqüências, $\sum_{i=1}^8 f_i$, é igual ao número total de observações, n . Para os dados na Tabela 3.4,

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^8 m_i f_i}{\sum_{i=1}^8 f_i} \\ &= \left(\frac{1}{1.067} \right) [99,5(13) + 139,5(150) + 179,5(442) + 219,5(299) \\ &\quad + 259,5(115) + 299,5(34) + 339,5(9) + 379,5(5)]\end{aligned}$$

$$\frac{212.166,5}{1.067} \\ 198,8 \text{ mg/100 ml.}$$

A média de dados agrupados é realmente uma média ponderada dos pontos médios dos intervalos; cada ponto médio é ponderado pela freqüência de observações dentro do intervalo.

3.3.2 Variância de Dados Agrupados

Depois de calcularmos a média de um conjunto de dados agrupados, poderíamos também encontrar sua variância ou seu desvio-padrão. Novamente, assumimos que todas as observações situadas em um intervalo particular são iguais ao ponto médio do intervalo, m_i . A variância dos dados agrupados é

$$s^2 = \frac{\sum_{i=1}^k (m_i - \bar{x})^2 f_i}{\left[\sum_{i=1}^k f_i \right] - 1},$$

em que todos os termos estão definidos, como para a média. Portanto, a variância dos dados agrupados na Tabela 3.4 é,

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^8 (m_i - 198,8)^2 f_i}{\left[\sum_{i=1}^8 f_i \right] - 1} \\ &= \left(\frac{1}{1.067 - 1} \right) [(-99,3)^2(13) + (-59,3)^2(150) + (-19,3)^2(442) \\ &\quad + (20,7)^2(299) + (60,7)^2(115) + (100,7)^2(34) \\ &\quad + (140,7)^2(9) + (180,7)^2(5)] \\ &= \frac{2.058.342,8}{1066} \\ &= 1.930,9 \text{ (mg/100 ml)}^2. \end{aligned}$$

Lembre-se de que o desvio-padrão é a raiz quadrada da variância; consequentemente, o desvio-padrão dos dados agrupados de níveis séricos de colesterol é

$$\begin{aligned} s &= \sqrt{1.930,9 \text{ (mg/100 ml)}^2} \\ &= 43,9 \text{ mg/100 ml.} \end{aligned}$$

3.4 Desigualdade de Chebychev

Uma vez que a média e o desvio-padrão de um conjunto de dados tenham sido calculados, esses dois números podem ser usados para resumir o todo das características da distribuição de valores. A média nos mostra onde as observações estão centralizadas; o desvio-padrão nos dá uma idéia da quantidade de dispersão ao redor do centro. Juntos, podem ser usados para se construir um intervalo que contenha uma proporção especificada de observações no conjunto de dados.

Diz-se freqüentemente que a média mais ou menos dois desvios-padrões engloba a maioria dos dados. Se conhecermos algo mais sobre a forma da distribuição dos valores,

essa afirmação pode ser feita mais precisamente. Quando os dados são simétricos e unimodais, por exemplo, podemos dizer que aproximadamente 67% das observações se encontram no intervalo $\bar{x} \pm 1s$, cerca de 95% no intervalo $\bar{x} \pm 2s$ e quase todas as observações no intervalo $\bar{x} \pm 3s$. Essa afirmação é conhecida como *regra empírica*. Veremos a regra empírica novamente no Capítulo 7, quando falaremos sobre as distribuições teóricas de valores de dados.

Infelizmente, a regra empírica é uma aproximação que se aplica somente quando os dados são simétricos e unimodais. Se não forem, a *desigualdade de Chebychev* pode ser usada em seu lugar para resumir a distribuição de valores. A desigualdade de Chebychev é menos específica do que a regra empírica, mas é verdadeira para qualquer conjunto de observações, independentemente de qual seja a sua forma. Ela nos permite dizer que para qualquer número k maior ou igual a 1, pelo menos $[1 - (1/k)^2]$ das medidas no conjunto de dados se encontram a até k desvios-padrão de suas médias [7]. Dado que $k = 2$, por exemplo, pelo menos

$$\begin{aligned} 1 - \left(\frac{1}{2}\right)^2 &= 1 - \left(\frac{1}{4}\right) \\ &= \frac{3}{4} \end{aligned}$$

dos valores se encontram a até dois desvios-padrão da média. Equivalentemente, poderíamos dizer que o intervalo $\bar{x} \pm 2s$ engloba pelo menos 75% das observações no grupo. Essa afirmação é verdadeira, independentemente de quais sejam os valores de \bar{x} de s . Analogamente, se $k = 3$, pelo menos

$$\begin{aligned} 1 - \left(\frac{1}{3}\right)^2 &= 1 - \left(\frac{1}{9}\right) \\ &= \frac{8}{9} \end{aligned}$$

das observações se encontram a até três desvios-padrão da média; portanto, $\bar{x} \pm 3s$ contém pelo menos 88,9% das medidas.

A desigualdade de Chebychev é uma afirmação mais conservadora do que a regra empírica. Ela se aplica à média e ao desvio-padrão de qualquer distribuição de valores, independentemente de qual seja sua forma. Retornando aos dados de FEV, na Tabela 3.1, podemos dizer que o intervalo

$$2,95 \pm (2 \times 0,62)$$

ou

$$(1,71, 4,19)$$

engloba pelo menos 75% das observações, enquanto o intervalo

$$2,95 \pm (3 \times 0,62)$$

ou

$$(1,09, 4,81)$$

contém pelo menos 88,9%. De fato, ambos os intervalos contêm todas as 13 medidas. Analogamente, para os dados de níveis séricos de colesterol na Tabela 3.4, podemos estabelecer o intervalo

$$198,8 \pm (2 \times 43,9)$$

que

$$(111,0, 286,6)$$

contém pelo menos 75% dos valores, enquanto o intervalo

$$198,8 \pm (3 \times 43,9)$$

que

$$(67,1, 330,5)$$

contém pelo menos 88,9%. Assim, embora conservadora, a desigualdade de Chebychev nos permite usar a média e o desvio-padrão para **qualquer** conjunto de dados — justamente dois números — para descrever o grupo inteiro.

3.5 Aplicações Adicionais

Em um estudo que investiga as causas de morte entre pessoas com asma severa, os dados foram registrados para dez pacientes que chegaram ao hospital em estado de parada respiratória e inconscientes. A Tabela 3.5 lista os batimentos cardíacos para os dez pacientes na internação do hospital [8]. Como podemos caracterizar esse conjunto de observações?

Para iniciar, poderíamos obter um batimento cardíaco típico para os dez indivíduos. A medida de tendência central mais comumente utilizada é a média. Para encontrar a média desses dados, simplesmente somamos todas as observações e dividimos a soma por $n = 10$. Assim, para os dados da Tabela 3.5,

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ &= \left(\frac{1}{10}\right)(167 + 150 + 125 + 120 + 150 + 150 + 40 \\ &\quad + 136 + 120 + 150) \\ &= \frac{1.308}{10} \\ &= 130,8 \text{ batidas por minuto.}\end{aligned}$$

O batimento cardíaco médio na internação do hospital é de 130,8 batidas por minuto.

Nesse conjunto de dados, o batimento cardíaco do paciente 7 é consideravelmente mais baixo do que o dos outros pacientes. O que aconteceria se essa observação fosse removida do grupo? Nesse caso

$$\begin{aligned}\bar{x} &= \left(\frac{1}{9}\right)(167 + 150 + 125 + 120 + 150 + 150 + 136 + 120 + 150) \\ &= \frac{1.268}{9} \\ &= 140,9 \text{ batidas por minuto.}\end{aligned}$$

TABELA 3.5

Batimentos cardíacos para dez pacientes asmáticos em estado de parada respiratória.

Paciente	Batimento Cardíaco
1	167
2	150
3	125
4	120
5	150
6	150
7	40
8	136
9	120
10	150

A média aumentou em aproximadamente dez batidas por minuto. Essa mudança demonstra a influência que uma simples observação não-usual pode ter sobre a média.

Uma segunda medida de tendência central é a mediana ou o 50º percentil do conjunto de dados. Ao ordenarmos as medidas desde a menor até a maior, temos

$$40, 120, 120, 125, 136, 150, 150, 150, 150, 167.$$

Como há um número par de observações, a mediana é tomada como a média dos dois valores mais centrais. Nesse caso, esses valores são a $10/2 = 5^{\text{a}}$ e a $(10/2) + 1 = 6^{\text{a}}$ observações. Conseqüentemente, a mediana dos dados é $(136 + 150)/2 = 143$ batidas por minuto, um número um tanto maior do que a média. Cinco observações são menores do que a mediana e cinco são maiores.

O cálculo da mediana leva em conta a ordenação e as grandezas relativas das observações. Se removemos novamente o paciente 7, a ordenação dos batimentos cardíacos seria

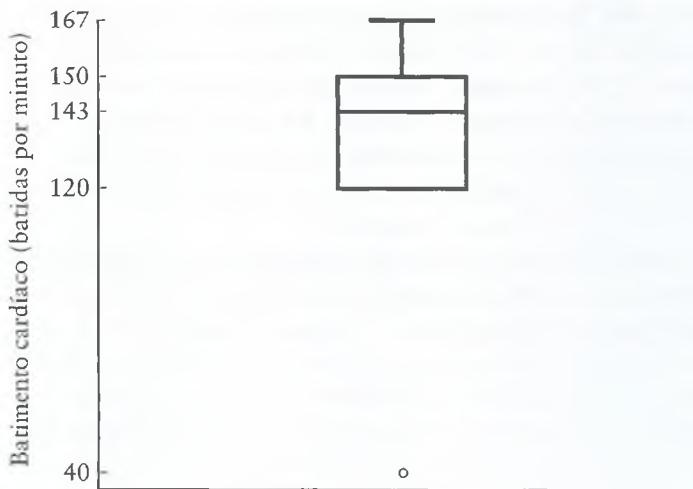
$$120, 120, 125, 136, 150, 150, 150, 150, 167.$$

Existem nove observações na lista; a mediana é a $[(9+1)/2] = 5^{\text{a}}$ medida ou 150 batidas por minuto. Embora a mediana aumente um pouco quando o paciente 7 é removido, ela não muda muito como o faz a média.

A moda de um conjunto de dados é a observação que ocorre mais freqüentemente. Para as medidas na Tabela 3.5, a moda é de 150 batidas por minuto, único valor que ocorre quatro vezes.

Ao encontrarmos o centro de um conjunto de dados, freqüentemente queremos também estimar a variabilidade entre as observações, o que nos permite quantificar em que grau o resumo é representativo do grupo. Uma medida de dispersão que pode ser usada é a amplitude a menor medida dos dados, que é a diferença entre a maior e a menor medidas. Para os batimentos cardíacos da Tabela 3.5, a amplitude é de $167 - 40 = 127$ batidas por minuto. Como a amplitude considera somente as observações mais extremas de um conjunto de dados, ela é altamente sensível aos pontos atípicos. Se removêssemos o paciente 7 do grupo, a amplitude dos dados seria somente de $167 - 120 = 47$ batidas por minuto.

O intervalo interquartil de um conjunto de dados é definido como o 75º percentil menos o 25º percentil. Se construíssemos um box plot usando os dados da Tabela 3.5 — como na Figura 3.5 — o intervalo interquartil seria a altura da caixa central. (Note-se que para esse particular conjunto de medidas, o valor adjacente mais baixo é igual ao 25º percentil.) Para encontrar o 25º percentil dos dados, notamos que $nk/100 = 10(25)/100 = 2,5$ não é um inteiro.

**FIGURA 3.5**

Batimentos cardíacos para dez pacientes asmáticos em estado de parada respiratória.

Portanto, o 25º percentil é a $2 + 1 = 3^{\text{a}}$ medida ordenada crescentemente, ou 120 batidas por minuto. Analogamente, $10(75)/100 = 7,5$ não é um inteiro e o 75º percentil é a $7 + 1 = 8^{\text{a}}$ medida ordenada crescentemente ou 150 batidas por minuto. Subtraindo-se esses valores, o intervalo interquartil para os dados de batimentos cardíacos é $150 - 120 = 30$ batidas por minuto; esse é o intervalo central que contém 50% das observações. O intervalo interquartil é freqüentemente usado com a mediana para descrever uma distribuição de valores.

As medidas de dispersão mais comumente usadas para um conjunto de valores de dados são a variância e o desvio-padrão. A variância quantifica a variabilidade dos dados ao redor da média e é calculada pela subtração da média de cada uma das medidas, elevação ao quadrado desses desvios, soma e divisão pelo número total de observações menos 1. A variância dos batimentos cardíacos na Tabela 3.5 é

$$\begin{aligned}
 s^2 &= \frac{1}{(10 - 1)} \sum_{i=1}^{10} (x_i - 130,8)^2 \\
 &= \left(\frac{1}{9}\right)[(36,2)^2 + (19,2)^2 + (-5,8)^2 + (-10,8)^2 + (19,2)^2 \\
 &\quad + (19,2)^2 + (-90,8)^2 + (5,2)^2 + (-10,8)^2 + (19,2)^2] \\
 &= \frac{11.323,6}{9} \\
 &= 1.258,2 \text{ (batidas por minuto)}^2.
 \end{aligned}$$

O desvio-padrão é a raiz quadrada positiva da variância. É usado mais freqüentemente na prática, porque tem a mesma unidade de medida que a média. Para as dez medidas de batimentos cardíacos, o desvio-padrão é

$$\begin{aligned}
 s &= \sqrt{1.258,2 \text{ (batidas por minuto)}^2} \\
 &= 35,5 \text{ batidas por minuto}.
 \end{aligned}$$

O desvio-padrão é usado tipicamente com a média para se descrever um conjunto de valores.

Agora que temos certa familiaridade com as medidas-resumo numéricas, considere a distribuição de freqüências dos pesos ao nascer da Tabela 3.6. Esses dados foram apresentados primeiro na Tabela 2.9 [9]. Podemos adicionalmente resumir esses dados e fazer uma afirmação concisa sobre sua distribuição? Embora não conheçamos os valores reais das 3.751.275 medidas dos pesos ao nascer, conhecemos os números de observações situadas em cada intervalo. Podemos, portanto, aplicar as técnicas para dados agrupados para obtermos as medidas-resumo numéricas para essas observações.

Para encontrar a média de dados agrupados, começamos por assumir que todas as observações em um intervalo particular são iguais ao seu ponto médio. Por exemplo: assumimos que as 4.843 medidas do primeiro intervalo tenham o valor de 249,5 gramas e que as 17.487 medidas do segundo intervalo tenham o valor de 749,5 gramas. Multiplicamos, então, cada ponto médio pela freqüência correspondente no intervalo, somamos esses produtos e dividimos a soma pelo número total de observações. Para os dados na Tabela 3.6,

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{11} m_i f_i}{\sum_{i=1}^{11} f_i} \\ &= \left(\frac{1}{3.751.275} \right) [(249,5)(4.843) + (749,5)(17.487) + (1.249,5)(23.139) \\ &\quad + (1.749,5)(49.112) + (2.249,5)(160.919) \\ &\quad + (2.749,5)(597.738) + (3.249,5)(1.376.008) \\ &\quad + (3.749,5)(1.106.634) + (4.249,5)(344.390) \\ &\quad + (4.749,5)(62.769) + (5.249,5)(8.236)] \\ &= \frac{12.560.121.114,5}{3.751.275} \\ &= 3.348,2 \text{ gramas.}\end{aligned}$$

TABELA 3.6

Freqüências absolutas dos pesos ao nascer,
Estados Unidos, 1986.

Peso ao Nascer (gramas)	Número de Bebês
0–499	4.843
500–999	17.487
1.000–1.499	23.139
1.500–1.999	49.112
2.000–2.499	160.919
2.500–2.999	597.738
3.000–3.499	1.376.008
3.500–3.999	1.106.634
4.000–4.499	344.390
4.500–4.999	62.769
5.000–5.499	8.236
Total	3.751.275

A média de dados agrupados é uma média ponderada dos pontos médios dos intervalos.

Além de calcularmos uma medida de tendência central, podemos também calcular uma medida de dispersão para a distribuição de freqüências. A variância dos dados agrupados na Tabela 3.6 é

$$\begin{aligned}
 s^2 &= \frac{\sum_{i=1}^{11} (m_i - 3.348,2)^2 f_i}{\left[\sum_{i=1}^{11} f_i\right] - 1} \\
 &= \frac{1}{(3.751.275 - 1)} [(-3.098,7)^2(4.843) + (-2.598,7)^2(17.487) \\
 &\quad + (-2.098,7)^2(23.139) + (-1.598,7)^2(49.112) \\
 &\quad + (-1.098,7)^2(160.919) + (-598,7)^2(597.738) \\
 &\quad + (-98,7)^2(1.376.008) + (401,3)^2(1.106.634) \\
 &\quad + (901,3)^2(344.390) + (1.401,3)^2(62.769) \\
 &\quad + (1.901,3)^2(8.236)] \\
 &= \frac{1.423.951.273.348,3}{3.751.274} \\
 &= 379.591,4 \text{ gramas}^2.
 \end{aligned}$$

O desvio-padrão, que é a raiz quadrada da variância, é

$$\begin{aligned}
 s &= \sqrt{379.591,4 \text{ gramas}^2} \\
 &= 616,1 \text{ gramas}.
 \end{aligned}$$

Em vez de obtermos todas essas medidas-resumo numéricas manualmente, podemos usar o computador para fazer os cálculos. A Tabela 3.7 mostra o resultado relevante do Stata para os dados de batimentos cardíacos na Tabela 3.5. Os percentis selecionados dos dados estão no lado esquerdo da tabela. Ao usarmos esses valores, podemos determinar a mediana e o intervalo interquartil. A coluna do meio contém as quatro medidas menores e as quatro maiores que nos permitem calcular a amplitude. A informação do lado direito da tabela inclui o número de observações, a média dos dados, o desvio-padrão e a variância.

TABELA 3.7

Saída do Stata mostrando as medidas-resumo numéricas.

hrrate				
	Percentiles	Smallest		
1%	40	40		
5%	40	120		
10%	80	120	Obs	10
25%	120	125	Sum of Wgt.	10
50%	143		Mean	130.8
		Largest	Std. Dev.	35.4708
75%	150	150		
90%	158.5	150	Variance	1258.178
95%	167	150	Skewness	-1.772591
99%	167	167	Kurtosis	5.479789

A Tabela 3.8 mostra a saída correspondente do Minitab. Note-se que ele fornece o número de observações, a média, a mediana e o desvio-padrão das medidas. Os valores mínimo e máximo podem ser usados para calcular a amplitude e os valores rotulados Q1 e Q3 — os 25º e 75º percentis ou quartis — para calcular o intervalo interquartil. A seção da saída chamada de TRMEAN contém a *média aparada* em 5% dos dados e, para calculá-la, as observações são ordenadas. As 5% menores e as 5% maiores medidas são descartadas; das 90% restantes é tirada a média. Para os dados de batimentos cardíacos, existem dez observações e 5% de 10 é 0,5. Arredondando para 1 (para cima), a medida menor e a maior são removidas. A média é então calculada para as oito medidas restantes. Quando os dados são aparados, eliminam-se os valores atípicos potenciais, e assim esse tipo de média não é influenciado por valores excepcionalmente altos nem baixos na mesma medida em que a média não aparada não é influenciada.

TABELA 3.8

Saída do Minitab que exibe as medidas-resumo numéricas.

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
HRTRATE	10	130.8	143.0	137.6	35.5	11.2
	MIN	MAX	Q1	Q3		
HRTRATE	40.0	167.0	120.0	150.0		

3.6 Exercícios de Revisão

- Defina e compare média, mediana e moda como medidas de tendência central.
- Sob que condições é preferível o uso da média? E o da mediana? E o da moda?
- Defina e compare as três medidas de dispersão comumente usadas — a amplitude, o intervalo interquartil e o desvio-padrão.
- Para observações que foram classificadas na forma de uma distribuição de freqüências, de modo que as medidas originais não estejam mais disponíveis, é possível calcular as medidas-resumo numéricas? Explique brevemente. Por que poderiam informações pessoais — tal como os rendimentos anuais — ser coletadas dessa maneira?
- Como a desigualdade de Chebychev é útil para descrever um conjunto de observações? Quando a regra empírica pode ser usada em seu lugar?
- Um estudo foi conduzido para investigar o prognóstico a longo prazo de crianças que sofreram um episódio agudo de meningite bacteriana (inflamação das membranas que envolvem o cérebro e a medula espinhal). Abaixo estão listados os tempos para o ataque apoplético de 13 crianças que tomaram parte no estudo [10]. Em meses, as medidas foram:

0,10 0,25 0,50 4 12 12 24 24 31 36 42 55 96

- (a) Obtenha as seguintes medidas-resumo numéricas dos dados

- i. média
- ii. mediana
- iii. moda
- iv. amplitude
- v. intervalo interquartil
- vi. desvio-padrão

- (b) Mostre que $\sum_{i=1}^{13} (x_i - \bar{x})$ é igual a zero.

7. Em Massachusetts, oito indivíduos sofreram um episódio inexplicável de intoxicação por vitamina D que exigiu hospitalização; pensou-se que essas ocorrências extraordinárias pudessem resultar de uma excessiva suplementação de leite [11]. Os níveis de cálcio e albumina — um tipo de proteína — no sangue para cada indivíduo no momento da internação no hospital são mostrados abaixo:

Cálcio (mmol/l)	Albumina (g/l)
2,92	43
3,84	42
2,37	42
2,99	40
2,67	42
3,17	38
3,74	34
3,44	42

- (a) Obtenha a média, a mediana, o desvio-padrão e a amplitude dos níveis de cálcio registrados.
 (b) Calcule a média, a mediana, o desvio-padrão e a amplitude para os dados de níveis de albumina.
 (c) Para indivíduos saudáveis, o intervalo normal de valores de cálcio é de 2,12 até 2,74 mmol/l, enquanto o intervalo de níveis de albumina é de 32 até 55 g/l. Você acredita que os pacientes que sofreram intoxicação por vitamina D tinham níveis normais de cálcio e de albumina no sangue?
 8. Um estudo foi conduzido comparando mulheres adolescentes que sofriam de bulimia com mulheres adolescentes com composição corporal e níveis de atividade física similares. Abaixo estão listadas as medidas de entrada calórica diária, registradas em quilocalorias por quilograma, para as amostras de adolescentes de cada grupo [12].

Consumo calórico diário (kcal/kg)				
Bulímica			Saudável	
15,9	18,9	25,1	20,7	30,6
16,0	19,6	25,2	22,4	33,2
16,5	21,5	25,6	23,1	33,7
17,0	21,6	28,0	23,8	36,6
17,6	22,9	28,7	24,5	37,1
18,1	23,6	29,2	25,3	37,4
18,4	24,1	30,9	25,7	40,8
18,9	24,5	30,6		

- (a) Obtenha o consumo calórico diário mediano tanto para as adolescentes bulímicas como para as saudáveis.
 (b) Calcule o intervalo interquartil para cada grupo.
 (c) Um valor típico do consumo calórico diário é maior para os indivíduos que sofrem de bulimia ou para adolescentes saudáveis? Que grupo tem maior variabilidade nas medidas?
 9. As Figuras 3.6 e 3.7 exibem as taxas de mortalidade infantil para 111 nações em três continentes: África, Ásia e Europa [13]. A taxa de mortalidade infantil para um país é o número de mortes de crianças com menos de um ano em um determinado ano dividido pelo número total de nascidos vivos naquele ano. A Figura 3.6 fornece os histogramas que ilustram a distribuição das taxas de mortalidade infantil para cada um dos continentes. A Figura 3.7 exibe os mesmos dados com o uso dos gráficos de dispersão unidimensionais e box plots.

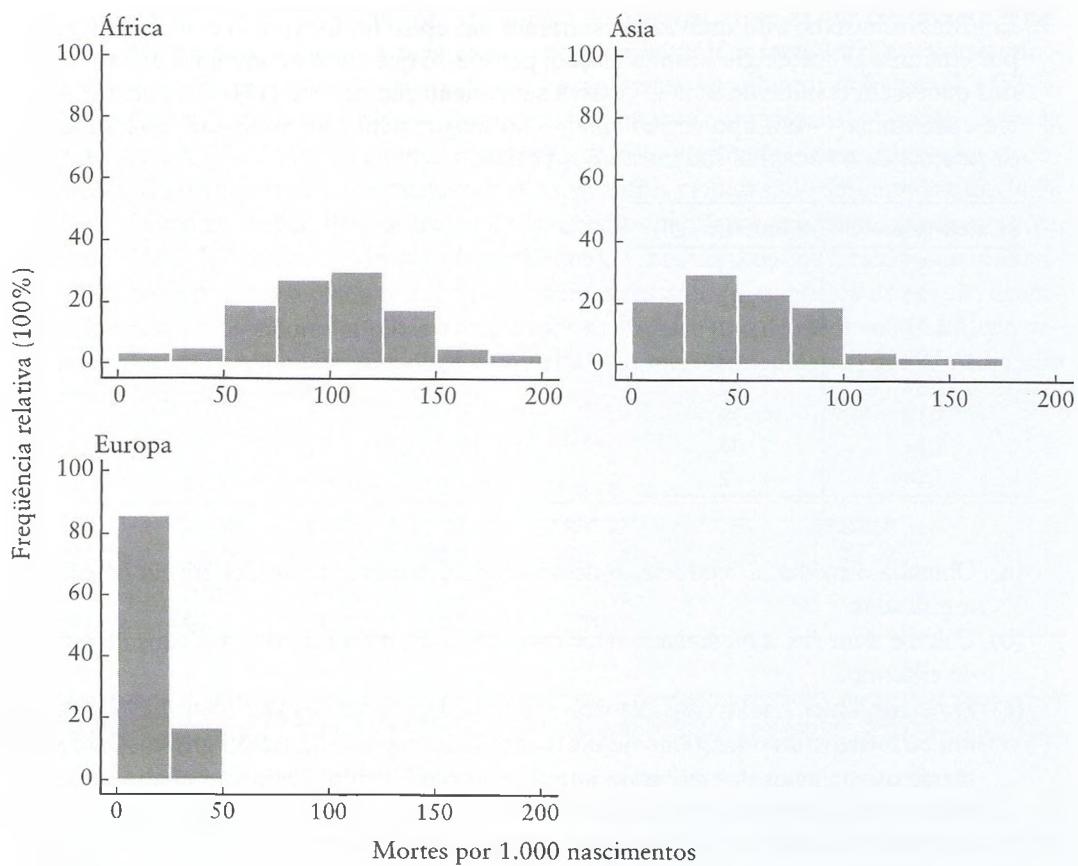


FIGURA 3.6
Histogramas de taxas de mortalidade infantil para a África, Ásia e Europa, 1992.

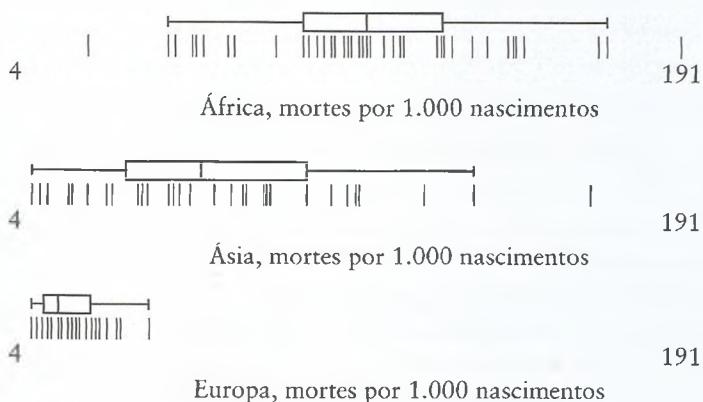


FIGURA 3.7
Gráficos de dispersão unidimensionais e box plots das taxas de mortalidade infantil para a África, Ásia e Europa, 1992.

- (a) Sem fazer qualquer cálculo, que continente você espera ter a menor média? E a maior mediana? E o menor desvio-padrão? Explique seu raciocínio.
- (b) Para a África, você esperaria que as taxas de mortalidade infantil média e mediana fossem aproximadamente iguais? Você esperaria que a média e a mediana fossem iguais para a Ásia? Por quê?
10. Abaixo está um par de distribuições de freqüências que contêm os níveis séricos de cotinina para um grupo de fumantes de cigarro e um grupo de não-fumantes [14]. Essas medidas foram registradas como parte de um estudo que investigou vários fatores de risco para doenças cardiovasculares.
- | Nível de Cotinina
(ng/ml) | Fumantes | Não-Fumantes |
|------------------------------|----------|--------------|
| 0–13 | 78 | 3.300 |
| 14–49 | 133 | 72 |
| 50–99 | 142 | 23 |
| 100–149 | 206 | 15 |
| 150–199 | 197 | 7 |
| 200–249 | 220 | 8 |
| 250–299 | 151 | 9 |
| 300+ | 412 | 11 |
| Total | 1.539 | 3.445 |
- (a) Calcule a média e o desvio-padrão dos dados agrupados das medidas do nível sérico de cotinina tanto para fumantes como para não-fumantes. Para o último intervalo — 300+ ng/ml — considere o ponto médio do intervalo como 340 ng/ml.
- (b) Em que intervalo se situa o nível sérico de cotinina mediano dos fumantes? E dos não-fumantes?
- (c) Compare as distribuições de níveis séricos de cotinina para os fumantes e os não-fumantes.
11. Os níveis séricos de zinco para 462 homens entre as idades de 15 e 17 anos estão armazenados em seu disco em um conjunto de dados chamado *serzinc* (Apêndice B, Tabela B.1); as medidas séricas de zinco em microgramas por decilitro estão salvas sob a variável de nome *zinc* [15].
- (a) Obtenha a média, a mediana, o desvio-padrão, a amplitude e o intervalo interquartil dos dados.
- (b) Use a desigualdade de Chebychev para descrever a distribuição de valores.
- (c) Que porcentagem dos valores você esperaria encontrar dentro de dois desvios-padrão da média? E dentro de três desvios-padrão da média? Que porcentagem das 462 medidas realmente ficam dentro desses intervalos?
- (d) A regra empírica resume melhor esses níveis séricos de zinco do que a desigualdade de Chebychev? Explique.
12. As porcentagens de baixos pesos ao nascer de bebês — definidos como os que pesam menos que 2.500 gramas — para diversas nações estão salvas sob a variável de nome *lowbwt* no conjunto de dados *unicef* [13] (Apêndice B, Tabela B.2).
- (a) Calcule a média e a mediana dessas observações.
- (a) Calcule a média aparada a 5%.
- (a) Para esse conjunto de dados, quais desses números você preferiria como medida de tendência central? Explique.

13. As concentrações declaradas de nicotina para 35 marcas de cigarros canadenses estão salvas sob a variável de nome `nicotine`, no conjunto de dados `cigarett` [16] (Apêndice B, Tabela B.4).
- Encontre as concentrações de nicotina média e mediana.
 - Produza um histograma das medidas de nicotina. Descreva a forma das distribuições de valores.
 - Que número você acha que fornece o melhor resultado de tendência central para essas concentrações, a média ou a mediana? Por quê?
14. Abaixo está uma distribuição de freqüências que contém um resumo das pressões sanguíneas sistólicas em repouso para uma amostra de 35 pacientes com doença isquêmica do coração ou supressão do fluxo de sangue para o coração [17].

Pressão Sangüínea (mm Hg)	Número de Pacientes
115–124	4
125–134	5
135–144	5
145–154	7
155–164	5
165–174	4
175–184	5
Total	35

- Calcule a média e o desvio-padrão dos dados agrupados.
 - As 35 medidas da pressão sangüínea sistólica estão armazenadas em seu disco em um arquivo chamado `ischemic` (Apêndice B, Tabela B.6); os valores estão salvos sob a variável de nome `sbp`. Calcule a média não-agrupada e o desvio-padrão não-agrupado desses dados.
 - As medidas-resumo numéricas agrupadas e não-agrupadas têm os mesmos valores? Por quê?
15. O conjunto de dados `lowbwt` contém a informação registrada para uma amostra de 100 bebês com baixos pesos ao nascer — que pesam menos do que 1.500 gramas — nascidos em dois hospitais-escola em Boston, Massachusetts [18] (Apêndice B, Tabela B.7). As medidas de pressão sangüínea sistólica estão salvas sob a variável de nome `sbp`. A variável aleatória dicotômica `sex` designa o sexo de cada criança, com 1 representando um menino e 0 uma menina.
- Construa um par de box plots para as medidas de pressão sangüínea sistólica — uma para meninos e uma para meninas. Compare as duas distribuições de valores.
 - Calcule a média e o desvio-padrão das medidas de pressão sangüínea sistólica para meninos e meninas. Que grupo tem a maior média? E o maior desvio-padrão?
 - Calcule o coeficiente de variação que corresponde a cada sexo. Há alguma evidência de que a variabilidade na pressão sangüínea sistólica difere para meninos e meninas?

6

Probabilidade

Nos capítulos anteriores, estudamos como as estatísticas descritivas podem ser usadas para organizar e resumir um conjunto de dados. No entanto, além de descrever um grupo de observações, podemos também investigar como a informação contida na amostra pode ser usada para inferir as características da população da qual foi retirada. Antes de fazê-lo, precisamos estabelecer os fundamentos. A base para a inferência estatística é a teoria da probabilidade. No Capítulo 5, usamos o termo *probabilidade* como sinônimo de *proporção*. Antes de darmos uma definição mais precisa da probabilidade, precisamos explicar o conceito de um evento.

5.1 Operações sobre Eventos e Probabilidade

Evento é o elemento básico para o qual a probabilidade pode ser aplicada. É o resultado de uma observação ou de um experimento ou a descrição de algum resultado potencial. Por exemplo, poderíamos considerar o evento de uma mulher de 30 anos viver para ver o seu 70º aniversário ou o evento da mesma mulher ser diagnosticada com câncer cervical, antes que atinja 40 anos. Outro evento poderia ser uma usina de energia nuclear causar uma desintegração nos próximos dez anos. Um evento pode ocorrer ou não. No estudo de probabilidade, os eventos são representados por letras maiúsculas, tais como A , B e C .

Diversas operações podem ser aplicadas aos eventos. A *intersecção* de dois eventos A e B , representada por $A \cap B$, é definida como o evento “tanto A como B ”. Por exemplo, seja A representante do evento de uma mulher de 30 anos estar viva em seu 70º aniversário e B o evento de seu marido de 30 anos estar ainda vivo aos 70 anos. A intersecção de A e B seria o evento de tanto a mulher de 30 anos como seu marido estarem vivos com 70 anos.

A *união* de A e B , representada por $A \cup B$, é o evento “ou A ou B ou ambos A e B ”. No exemplo mencionado acima, a união de A e B seria o evento de ou a mulher de 30 anos ou seu marido de 30 anos viverem até 70 anos ou que ambos vivam até os 70 anos.

O *complemento* de um evento A , indicado por A^c ou \bar{A} , é o evento “não A ”. Consequentemente, A^c é o evento de que uma mulher de 30 anos morra antes de atingir os 70 anos.

Essas três operações — a intersecção, a união e o complemento — podem ser usadas para descrever mesmo as situações mais complicadas, em termos de eventos simples. Para auxiliar a tornar essa idéia mais concreta, uma figura chamada *diagrama de Venn* é um dispositivo útil para representar a relação entre os eventos. Na Figura 6.1, por exemplo, a área de cada caixa representa todos os resultados que poderiam possivelmente ocorrer. Dentro das caixas, os círculos rotulados A representam o subconjunto de resultados para os quais uma

mulher de 30 anos viva até os 70 anos e aqueles de B indicam os resultados nos quais seu marido de 30 anos sobrevive até os 70 anos. A intersecção de A e B é representada pela área na qual os dois círculos se sobrepõem e está sombreada na Figura 6.1(a). A união de A e B está sombreada na Figura 6.1(b) e é a área em que está ou A ou B ou ambos. O complemento de A , como mostrado na Figura 6.1(c), é tudo que está dentro da caixa e que se encontra fora de A .

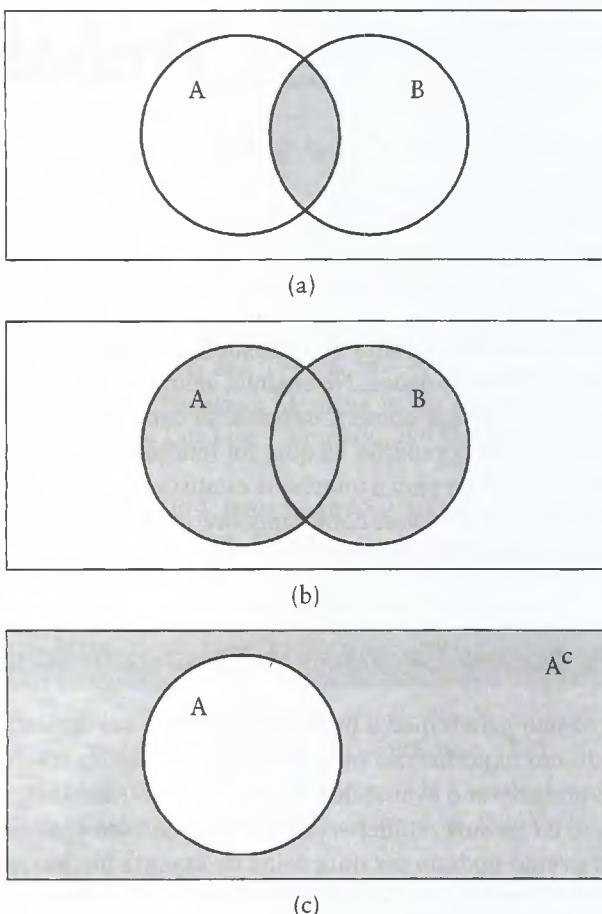


FIGURA 6.1
Diagramas de Venn representando as operações sobre eventos.

Estamos agora prontos para discutir o conceito de probabilidade. Como um sistema matemático, a teoria da probabilidade é bem definida. Como desejamos aplicar essa teoria, no entanto, necessitamos de uma definição prática de trabalho. Muitas definições de probabilidade têm sido propostas; a aqui apresentada é chamada de *definição freqüentista* e estabelece:

Se um experimento é repetido n vezes sob condições essencialmente idênticas e se o evento A ocorre m vezes, então, conforme n aumenta, a razão m/n se aproxima de um limite fixado, que é a probabilidade de A .

$$P(A) = \frac{m}{n}.$$

Em outras palavras, a *probabilidade* de um evento A é sua freqüência relativa de ocorrência — ou a proporção das vezes que o evento ocorre — em um número grande de repe-

tidas tentativas, sob condições virtualmente idênticas. A natureza prática dessa definição faz com que ela seja um tanto vaga, embora funcione bastante bem.

Como uma aplicação da definição freqüentista, podemos determinar a probabilidade de que um bebê recém-nascido esteja vivo em seu primeiro aniversário. Consulte, na Tabela 5.1, a tábua de vida de 1992 para a população dos Estados Unidos [1]. Entre os 100.000 indivíduos nascidos nessa coorte — consideramos esses bebês como os “experimentos” — o evento de sobreviver ao primeiro ano de vida ocorre 99.149 vezes. Portanto,

$$\begin{aligned} P(\text{de uma criança sobreviver ao seu primeiro ano de vida}) &= \frac{99.149}{100.000} \\ &= 0,99149. \end{aligned}$$

Assumimos que 100.000 repetições seja um número grande o suficiente para satisfazer à definição freqüentista de probabilidade.

O valor numérico de uma probabilidade se encontra entre 0 e 1. Se um evento particular acontece com certeza, ele ocorre em cada uma das n tentativas e tem a probabilidade $n/n = 1$. Seja novamente A representante do evento de uma mulher de 30 anos viver até os 70 anos. Nesse caso,

$$\begin{aligned} P(A \cup A^c) &= P(\text{ou } A \text{ ou } A^c \text{ ou ambos}) \\ &= P(\text{uma mulher de 30 anos viver até atingir os 70 anos ou} \\ &\quad \text{não viver até os 70 anos}) \\ &= 1, \end{aligned}$$

como é certo que a mulher viverá ou morrerá. Na Figura 6.1(c), A e A^c juntos preenchem a caixa inteira. Além disso, note que é impossível que A e A^c ocorram simultaneamente. Se um evento nunca pode acontecer, ele tem a probabilidade $0/n = 0$; por isso,

$$\begin{aligned} P(A \cap A^c) &= P(A \text{ e } A^c) \\ &= P(\text{uma mulher de 30 anos viver até atingir os 70 anos e} \\ &\quad \text{não viver até os 70 anos}) \\ &= 0. \end{aligned}$$

Um evento que nunca pode ocorrer é chamado *evento nulo* e é representado pelo símbolo ϕ . Logo, $A \cap A^c = \phi$. A maioria dos eventos tem probabilidades entre 0 e 1.

Ao usarmos a definição freqüentista de probabilidade de um evento A , podemos calcular de maneira direta a probabilidade de um evento complementar A^c . Se um experimento é repetido n vezes sob condições essencialmente idênticas e o evento A ocorre m vezes, o evento A^c , ou não- A , precisa ocorrer $n - m$ vezes. Portanto, para n grande,

$$\begin{aligned} P(A^c) &= \frac{n - m}{n} \\ &= 1 - \frac{m}{n} \\ &= 1 - P(A). \end{aligned}$$

A probabilidade de que um recém-nascido não sobreviva ao primeiro ano de vida é de 1 menos a probabilidade de que ele o faça ou

$$1 - 0,99149 = 0,00851.$$

Dois eventos A e B que não podem ocorrer simultaneamente são denominados *mutuamente exclusivos* ou *disjuntos*. Se A é o evento em que o peso ao nascer esteja abaixo de 2.000 gramas e B o evento de que esteja entre 2.000 e 2.499 gramas, por exemplo, os eventos A e B são mutuamente exclusivos. Uma criança não pode estar nos dois grupos de peso ao mesmo tempo. Por definição, $A \cap B = \emptyset$ e $P(A \cap B) = 0$. Na Figura 6.2, os círculos não-sobrepostos representam eventos mutuamente exclusivos.

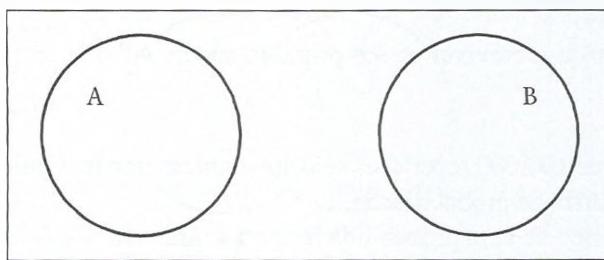


FIGURA 6.2

Diagramas de Venn representando dois eventos mutuamente exclusivos.

Quando dois eventos são mutuamente exclusivos, a *regra aditiva de probabilidade* estabelece que a probabilidade de que um dos eventos ocorra é igual à soma das probabilidades dos eventos individuais; mais explicitamente,

$$P(A \cup B) = P(A) + P(B).$$

Suponha que saibamos que a probabilidade de o peso ao nascer ser menor que 2.000 gramas, seja 0,025, e a probabilidade de estar entre 2.000 e 2.499 seja 0,043. A probabilidade de que um desses dois eventos ocorra, ou que a criança pese menos que 2.500 gramas é

$$\begin{aligned} P(A \cup B) &= 0,025 + 0,043 \\ &= 0,068. \end{aligned}$$

A regra aditiva pode ser estendida para o caso de três ou mais eventos mutuamente exclusivos. Se A_1, A_2, \dots, A_n são n eventos tais que $A_1 \cap A_2 = \emptyset, A_1 \cap A_3 = \emptyset, A_2 \cap A_3 = \emptyset$ e assim por diante para todos os pares possíveis, então

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

Se os eventos A e B não são mutuamente exclusivos, como na Figura 6.1(b), a regra aditiva não mais se aplica. Seja A o evento de que o peso ao nascer esteja abaixo de 2.000 gramas e B o evento de que esteja abaixo de 2.500 gramas. Desde que os dois eventos possam ocorrer simultaneamente — considere uma criança cujo peso ao nascer seja 1.850 gramas — existirá uma área na qual se sobreporão. Se simplesmente somássemos as probabilidades dos eventos individuais, essa área de sobreposição seria contada duas vezes. Portanto, quando dois eventos não são mutuamente exclusivos, a probabilidade de que um dos eventos ocorra é igual à soma das probabilidades individuais, menos a probabilidade de suas intersecções:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

6.2 Probabilidade Condicional

Freqüentemente estamos interessados em determinar a probabilidade de que um evento B ocorra, dado que nós já conhecemos o resultado de outro evento A . A ocorrência prévia de A faz a probabilidade de B se modificar? Por exemplo, em vez de encontrarmos a probabilidade de que uma pessoa viva até 65 anos, podemos querer saber a probabilidade de o indivíduo sobreviver pelos próximos cinco anos, uma vez já tenha atingido os 60 anos. Nesse caso, lidamos com uma *probabilidade condicional*. A notação $P(B | A)$ é usada para representar a probabilidade do evento B , dado que o evento A tenha já ocorrido.

A *regra multiplicativa de probabilidade* estabelece que a probabilidade de que dois eventos A e B ocorram é igual à probabilidade de A multiplicada pela probabilidade de B , dado que A já tenha ocorrido. Isso pode ser expresso como

$$P(A \cap B) = P(A)P(B | A).$$

Desde que é arbitrário qual evento chamamos de A e qual chamamos de B , podemos também escrever

$$P(A \cap B) = P(B)P(A | B).$$

Ao dividirmos ambos os lados da primeira equação por $P(A)$, obtemos a fórmula para uma probabilidade condicional como

$$P(B | A) = \frac{P(A \cap B)}{P(A)},$$

dado que $P(A) \neq 0$. Analogamente, temos

$$P(A | B) = \frac{P(A \cap B)}{P(B)},$$

desde que $P(B) \neq 0$.

Se A é o evento que um indivíduo esteja vivo aos 60 anos e B é o evento que sobrevive até 65 anos, então $A \cap B$ é o evento que a pessoa esteja viva aos 60 anos e também aos 65 anos. Se alguém está vivo aos 65, estava também aos 60 anos, portanto, $A \cap B$ é simplesmente o evento que o indivíduo sobreviva ao seu 65º aniversário. De acordo com a tábua de vida de 1992 para a população dos Estados Unidos,

$$P(A) = P(\text{que um indivíduo viva para atingir os 60 anos})$$

$$\begin{aligned} &= \frac{85.993}{100.000} \\ &= 0,85993. \end{aligned}$$

Em outras palavras, o evento A ocorre 85.993 vezes dentre 100.000 ensaios. Analogamente,

$$P(A \cap B) = P(\text{que um indivíduo viva para atingir os 65 anos de idade})$$

$$\begin{aligned} &= \frac{80.145}{100.000} \\ &= 0,80145. \end{aligned}$$

Portanto,

$$\begin{aligned}
 P(B | A) &= P(\text{que um indivíduo viva até os 65 anos} | \text{que esteja vivo aos 60}) \\
 &= \frac{P(A \cap B)}{P(A)} \\
 &= \frac{0,80145}{0,85993} \\
 &= 0,9320.
 \end{aligned}$$

Um modo equivalente de se calcular essa probabilidade seria iniciar com 85.993 pessoas vivas com idade de 60 anos e notar que o evento de sobreviver até 65 anos ocorre 80.145 vezes nesses 85.993 ensaios. Por isso,

$$\begin{aligned}
 P(B | A) &= \frac{80.145}{85.993} \\
 &= 0,9320.
 \end{aligned}$$

Se uma pessoa vive até os 60 anos, sua chance de sobreviver à idade de 65 anos é maior do que ao nascer.

Quando estamos preocupados com dois eventos tais que o resultado de um não tenha efeito na ocorrência ou não-ocorrência do outro, diz-se que os eventos são *independentes*. Se A e B são eventos independentes,

$$P(A | B) = P(A)$$

e

$$P(B | A) = P(B).$$

Nesse caso em especial, a regra multiplicativa de probabilidade pode ser escrita

$$P(A \cap B) = P(A) P(B).$$

É importante notar que os termos *independente* e *mutuamente exclusivos* não significam a mesma coisa. Se A e B são independentes e o evento A ocorre, o resultado de B não é afetado. O evento B tanto poderia ocorrer como não ocorrer e $P(B | A) = P(B)$. Se A e B são mutuamente exclusivos, no entanto, e o evento A ocorre, o evento B não pode ocorrer. Por definição, $P(B | A) = 0$.

6.3 Teorema de Bayes

O Capítulo 4 inclui uma representação de dados coletados no Levantamento Nacional de Entrevistas de Saúde de 1980–1981 [2]. Os dados relacionam-se com as debilidades auditivas devido a lesões registradas por indivíduos de 17 anos de idade e mais velhos. As 163.157 pessoas incluídas no estudo foram subdivididas em três categorias mutuamente exclusivas: os atualmente empregados, os atualmente desempregados e os que estão fora da força de trabalho.

Status	População	Debilidades
Atualmente empregados	98.917	552
Atualmente desempregados	7.462	27
Fora da força de trabalho	56.778	368
Total	163.157	947

Seja E_1 o evento que um indivíduo incluído no levantamento esteja atualmente empregado, E_2 o evento que esteja atualmente desempregado e E_3 o evento que o indivíduo esteja fora da força de trabalho. Se assumimos que esses números são grandes o bastante para satisfazer à definição freqüentista de probabilidade, então, a partir dos dados fornecidos, encontramos

$$\begin{aligned} P(E_1) &= \frac{98.917}{163.157} \\ &= 0,6063, \end{aligned}$$

$$\begin{aligned} P(E_2) &= \frac{7.462}{163.157} \\ &= 0,0457, \end{aligned}$$

e

$$\begin{aligned} P(E_3) &= \frac{56.778}{163.157} \\ &= 0,3480. \end{aligned}$$

Se S é o evento que o indivíduo em estudo esteja atualmente empregado ou atualmente desempregado ou que não esteja na força de trabalho

$$S = E_1 \cup E_2 \cup E_3.$$

Como as três categorias são mutuamente exclusivas, a regra aditiva de probabilidade pode ser aplicada:

$$\begin{aligned} P(S) &= P(E_1 \cup E_2 \cup E_3) \\ &= P(E_1) + P(E_2) + P(E_3) \\ &= 0,6063 + 0,0457 + 0,3480 \\ &= 1,0000. \end{aligned}$$

Quando as probabilidades de eventos mutuamente exclusivos somam 1, diz-se que os eventos são *exaustivos*: nesse caso, não existem outros resultados possíveis. Portanto, cada pessoa incluída no levantamento precisa estar dentro de um dos três grupos.

Seja agora H o evento de que um indivíduo tenha uma debilidade auditiva devido a lesões. No total,

$$\begin{aligned} P(H) &= \frac{947}{163.157} \\ &= 0,0058. \end{aligned}$$

Ao se verificar cada um dos subgrupos de *status* de emprego separadamente,

$$P(H | E_1) = P(\text{que um indivíduo tenha uma debilidade auditiva} | \text{que esteja atualmente empregado})$$

$$= \frac{552}{98.917} \\ = 0,0056,$$

$$P(H | E_2) = P(\text{que um indivíduo tenha uma debilidade auditiva} | \text{que esteja atualmente desempregado})$$

$$= \frac{27}{7.462} \\ = 0,0036,$$

e

$$P(H | E_3) = P(\text{que um indivíduo tenha uma debilidade auditiva} | \text{que esteja fora da força de trabalho})$$

$$= \frac{368}{56.778} \\ = 0,0065.$$

A probabilidade de se ter uma debilidade auditiva é menor entre os atualmente desempregados e maior entre os que estão fora da força de trabalho.

Note que H , o evento que um indivíduo tenha uma debilidade auditiva devido a lesões, pode ser expresso como a união de três eventos mutuamente exclusivos: $E_1 \cap H$, o evento de que um indivíduo esteja atualmente empregado e tenha uma debilidade auditiva; $E_2 \cap H$, o evento de que ele esteja atualmente desempregado e tenha uma debilidade auditiva e $E_3 \cap H$, o evento que o indivíduo esteja fora da força de trabalho e tenha uma debilidade auditiva. Assim,

$$H = (E_1 \cap H) \cup (E_2 \cap H) \cup (E_3 \cap H).$$

Todo indivíduo que tenha debilidade auditiva pode ser colocado em uma e somente uma dessas três categorias. Como as categorias são mutuamente exclusivas, podemos aplicar a regra aditiva; logo,

$$P(H) = P[(E_1 \cap H) \cup (E_2 \cap H) \cup (E_3 \cap H)] \\ = P(E_1 \cap H) + P(E_2 \cap H) + P(E_3 \cap H).$$

Isso é algumas vezes chamado de *lei da probabilidade total*.

Agora, ao aplicarmos a regra multiplicativa para cada termo do lado direito da equação separadamente e inserir as probabilidades previamente calculadas,

$$P(H) = P(E_1 \cap H) + P(E_2 \cap H) + P(E_3 \cap H) \\ = P(E_1) P(H | E_1) + P(E_2) P(H | E_2) + P(E_3) P(H | E_3) \\ = 0,0034 + 0,0002 + 0,0023 \\ = 0,0059.$$

Esses cálculos estão resumidos na tabela a seguir, na qual i , o subscrito do evento E , toma valores de 1 até 3.

Evento E_i	$P(E_i)$	$P(H E_i)$	$P(E_i)P(H E_i)$
E_1	0,6063	0,0056	0,0034
E_2	0,0457	0,0036	0,0002
E_3	0,3480	0,0065	0,0023
$P(H)$			0,0059

Se ignorarmos os erros de arredondamento nesses cálculos, o valor 0,0059 é o número que originalmente geramos como a probabilidade de que um indivíduo tenha uma debilidade auditiva devido a lesões,

$$\begin{aligned} P(H) &= \frac{947}{163.157} \\ &= 0,0058. \end{aligned}$$

O método de cálculo mais complicado, usando a expressão

$$P(H) = P(E_1)P(H|E_1) + P(E_2)P(H|E_2) + P(E_3)P(H|E_3),$$

pode ser útil, quando não podemos calcular $P(H)$ diretamente.

Suponha agora que mudamos nossa perspectiva e tentamos encontrar $P(E_1|H)$, a probabilidade de que um indivíduo esteja atualmente empregado dado que tenha uma debilidade auditiva. A regra multiplicativa de probabilidade estabelece que

$$P(E_1 \cap H) = P(H)P(E_1|H);$$

Portanto,

$$P(E_1|H) = \frac{P(E_1 \cap H)}{P(H)}.$$

Ao aplicarmos a regra multiplicativa ao numerador do lado direito da equação, temos

$$P(E_1|H) = \frac{P(E_1)P(H|E_1)}{P(H)}.$$

Ao usarmos a identidade que foi derivada acima,

$$P(H) = P(E_1)P(H|E_1) + P(E_2)P(H|E_2) + P(E_3)P(H|E_3),$$

resulta em

$$P(E_1|H) = \frac{P(E_1)P(H|E_1)}{P(E_1)P(H|E_1) + P(E_2)P(H|E_2) + P(E_3)P(H|E_3)}.$$

Essa expressão um tanto desencorajadora é conhecida como *teorema de Bayes*. Ao substituirmos os valores numéricos de todas as probabilidades

$$\begin{aligned} P(E_1 | H) &= \frac{(0,6063)(0,0056)}{(0,6063)(0,0056) + (0,0457)(0,0036) + (0,3480)(0,0065)} \\ &= 0,583. \end{aligned}$$

A probabilidade de um indivíduo empregado atualmente dado ter uma debilidade auditiva devido a lesões é aproximadamente 0,583. Nesse exemplo em particular, o resultado pode ser verificado diretamente ao olhar-se os dados originais. Entre as 947 pessoas com debilidades auditivas, 552 estão atualmente empregadas. Portanto,

$$\begin{aligned} P(E_1 | H) &= \frac{552}{947} \\ &= 0,583. \end{aligned}$$

O teorema de Bayes não é restrito às situações nas quais os indivíduos ficam em um de três subgrupos distintos. Se A_1, A_2, \dots e A_n são n eventos mutuamente exclusivos e exaustivos tais que

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= P(A_1) + P(A_2) + \dots + P(A_n) \\ &= 1, \end{aligned}$$

o teorema de Bayes estabelece que

$$P(A_i | B) = \frac{P(A_i) P(B | A_i)}{P(A_1) P(B | A_1) + \dots + P(A_n) P(B | A_n)}$$

para cada i , $1 \leq i \leq n$.

O teorema de Bayes é valioso porque nos permite recalcular uma probabilidade com base em algumas informações novas. No exemplo do Levantamento Nacional de Entrevis-tas de Saúde, sabemos que

$$\begin{aligned} P(E) &= P(\text{que um indivíduo esteja atualmente empregado}) \\ &= 0,6063. \end{aligned}$$

Se fornecermos um volume adicional de informações — o conhecimento de que um indivíduo em particular tem uma debilidade auditiva devido a lesões, por exemplo — modifica-se nossa avaliação da probabilidade de que ele esteja atualmente empregado? Observamos que sim. Usando o teorema de Bayes, encontramos que

$$\begin{aligned} P(E_1 | H) &= P(\text{que um indivíduo esteja atualmente empregado} | \text{que tenha uma debili-dade auditiva}) \\ &= 0,5832. \end{aligned}$$

Se falamos que alguém tem uma debilidade auditiva, a probabilidade de que esteja atualmen-te empregado diminui um pouco.

6.4 Testes de Diagnósticos

O teorema de Bayes é empregado freqüentemente na realização de testes de diagnósticos ou triagens. A *triagem* é a aplicação de um teste em indivíduos que não apresentam qualquer sintoma clínico para classificá-los com relação às probabilidades de terem em uma doença particular. Os que apresentam resultado positivo são considerados mais prováveis de terem a doença e normalmente são submetidos a procedimentos de diagnósticos adicionais ou a tratamentos. A triagem é freqüentemente mais utilizada por profissionais da área de saúde em situações nas quais a detecção prévia de doença contribua para prognósticos mais favoráveis ao indivíduo ou para a população em geral. O teorema de Bayes nos permite usar a probabilidade para avaliar incertezas associadas.

6.4.1 Sensibilidade e Especificidade

Suponha que estamos interessados em dois estados de saúde mutuamente exclusivos e exaustivos: D_1 é o evento em que um indivíduo tem uma doença particular e D_2 o evento em que ele não tenha a doença. Podemos usar a notação mais sucinta definida anteriormente — a saber, D e D^c — mas queremos enfatizar que a situação pode ser generalizada para incluir três ou mais eventos. Seja T^+ representante de um resultado positivo do teste de triagem. Queremos encontrar $P(D_1 | T^+)$, a probabilidade de que uma pessoa com um resultado de teste positivo realmente tenha a doença.

O câncer do colo do útero é uma doença cuja chance de refreamento é alta, desde que detectado no início. O Papanicolau é um procedimento de triagem amplamente aceito que pode detectar um câncer que seja ainda assintomático; tem sido creditado como o primeiro responsável pelo decréscimo da taxa de mortalidade por câncer do colo do útero nos anos recentes. Um teste de proficiência *in loco*, conduzido em 1972, 1973 e 1978, avaliou a competência dos técnicos que analisavam o Papanicolau para anormalidades [3]. Os técnicos de 306 laboratórios de citologia em 44 estados foram avaliados.

No total, 16,25% dos testes realizados em mulheres com câncer resultaram em falsos negativos. Um *falso negativo* ocorre quando o teste de uma mulher com câncer no colo do útero indica incorretamente que ela não o tem. Portanto, nesse estudo,

$$P(\text{teste negativo} | \text{câncer}) = 0,1625.$$

Os outros $100 - 16,25 = 83,75\%$ das mulheres que tinham câncer no colo do útero de fato apresentaram resultado positivo,

$$P(\text{teste positivo} | \text{câncer}) = 0,8375.$$

A probabilidade de um resultado positivo de teste dado que o indivíduo testado realmente tenha a doença é chamada *sensibilidade* de um teste. Nesse estudo, a sensibilidade do Papanicolau foi de 0,8375.

Nem todas as mulheres realmente testadas sofriam de câncer no colo do útero. De fato, 18,64% dos testes resultaram *falsos positivos*, implicando que

$$P(\text{teste positivo} | \text{sem câncer}) = 0,1864.$$

A *especificidade* de um teste é a probabilidade de que seu resultado seja negativo, dado que o indivíduo testado não tenha a doença. Nesse estudo, a especificidade do Papanicolau foi

$$\begin{aligned} P(\text{teste negativo} | \text{não câncer}) &= 1 - 0,1864 \\ &= 0,8136. \end{aligned}$$

6.4.2 Aplicações do Teorema de Bayes

Agora que examinamos a precisão do Papanicolau entre mulheres que têm câncer no colo do útero e mulheres que não o tem, podemos investigar a questão de fundamental preocupação para os indivíduos que estão sendo testados e para os profissionais da área de saúde envolvidos na triagem: qual é a probabilidade de que uma mulher com “Papanicolau” positivo para o câncer realmente tenha a doença? Seja D_1 representando o evento que uma mulher tenha câncer no colo do útero e D_2 o evento que ela não tenha. Seja, também, T^+ representando um Papanicolau positivo. Queremos calcular $P(D_1 | T^+)$. Ao usarmos o teorema de Bayes podemos escrever

$$\begin{aligned} P(D_1 | T^+) &= \frac{P(D_1 \cap T^+)}{P(T^+)} \\ &= \frac{P(D_1) P(T^+ | D_1)}{P(D_1) P(T^+ | D_1) + P(D_2) P(T^+ | D_2)}. \end{aligned}$$

Já sabemos que $P(T^+ | D_1) = 0,8375$ e $P(T^+ | D_2) = 0,1864$. Precisamos agora encontrar $P(D_1)$ e $P(D_2)$.

$P(D_1)$ é a probabilidade de que uma mulher sofra de câncer no colo do útero. Pode também ser interpretada como a proporção de mulheres que tem câncer do colo do útero em um determinado ponto no tempo ou a *prevalência* da doença. Uma fonte registra que a taxa de casos desse câncer entre as mulheres estudadas em 1983-1984 foi de 8,3 por 100.000 [4]. Usando esses dados,

$$P(D_1) = 0,000083.$$

$P(D_2)$ é a probabilidade de que uma mulher não tenha câncer no colo do útero. Como D_2 é o complemento de D_1 ,

$$\begin{aligned} P(D_2) &= 1 - P(D_1) \\ &= 1 - 0,000083 \\ &= 0,999917. \end{aligned}$$

Substituindo essas probabilidades no teorema de Bayes,

$$\begin{aligned} P(D_1 | T^+) &= \frac{0,000083 \times 0,8375}{(0,000083 \times 0,8375) + (0,999917 \times 0,1864)} \\ &= 0,000373. \end{aligned}$$

$P(D_1 | T^+)$, a probabilidade de doença dado um resultado positivo de teste, é chamado *valor preditivo* de um teste positivo. Aqui, ele nos mostra que, para cada 1.000.000 Papanicolau positivos, somente 373 representam casos verdadeiros de câncer no colo uterino.

O teorema de Bayes pode também ser usado para calcular o valor preditivo de um teste negativo. Se T^- representa um resultado negativo de teste, o valor preditivo negativo ou a probabilidade de não-doença dado um resultado negativo de teste é igual a

$$\begin{aligned} P(D_2 | T^-) &= \frac{P(D_2) P(T^- | D_2)}{P(D_2) P(T^- | D_2) + P(D_1) P(T^- | D_1)} \\ &= \frac{0,999917 \times 0,8136}{(0,999917 \times 0,8136) + (0,000083 \times 0,1625)} \\ &= 0,999983. \end{aligned}$$

Portanto, para cada 1.000.000 mulheres com Papanicolau negativos, 999.983 não sofrem da doença. A Figura 6.3 ilustra os resultados do processo inteiro do teste de diagnóstico. Note-se que todos os números foram arredondados para o inteiro mais próximo.

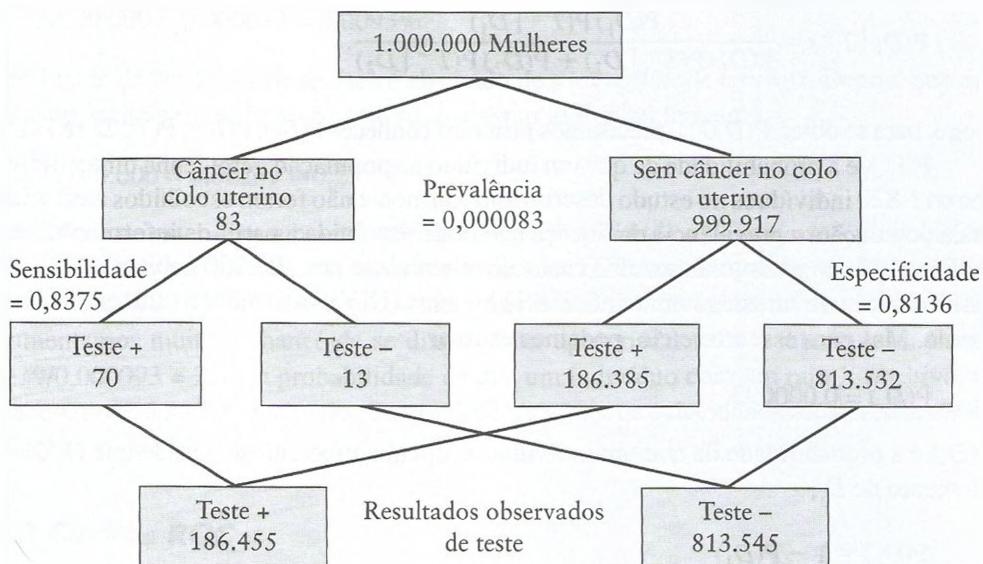


FIGURA 6.3

Desempenho do Papanicolau como teste de diagnóstico para o câncer de colo de útero.

Embora os Papanicolau sejam amplamente aceitos como teste de triagem para o câncer no colo uterino, sua alta taxa de precisão previamente assumida está sendo questionada. Diversos estudos estimam que a proporção de resultados falsos negativos esteja no intervalo de 20% a 40% ou seja mesmo tão grande quanto 89% [5, 6]. Encontra-se uma proporção de resultados falsos positivos tão alta quanto 86%. Alguns dos erros de laboratório são devidos às técnicas pobres de amostragem de células ou à inadequada preparação de espécimes; outros resultam da fatiga sofrida pelos técnicos de laboratório que precisam diariamente examinar grande número de lâminas.

Como um segundo exemplo da aplicação do teorema de Bayes em testes de diagnósticos, considere os seguintes dados: entre os 1.820 indivíduos em um estudo, 30 sofreram de tuberculose e 1.790 não [7]. Raios X do tórax foram administrados para todos; 73 tiveram raio X positivo — o que implica evidência significante de doença inflamatória — enquanto os resultados dos outros 1.747 foram negativos. Os dados para esse estudo são apresentados na tabela abaixo. Qual é a probabilidade de que um indivíduo selecionado aleatoriamente tenha tuberculose, uma vez que seu raio X seja positivo?

Raio X	Tuberculose		Total
	Não	Sim	
Negativo	1.739	8	1.747
Positivo	51	22	73
Total	1.790	30	1.820

Seja D_1 representante do evento de que um indivíduo sofra de tuberculose e D_2 do evento de que não sofra. Esses dois eventos são mutuamente exclusivos e exaustivos. Além dis-

so, T^+ representa um raio X positivo. Queremos encontrar $P(D_1 | T^+)$, a probabilidade de que um indivíduo que apresenta resultado positivo para a tuberculose tenha realmente a doença. Esse é o valor preditivo positivo do raio X. Usando o teorema de Bayes, podemos escrever

$$P(D_1 | T^+) = \frac{P(D_1) P(T^+ | D_1)}{P(D_1) P(T^+ | D_1) + P(D_2) P(T^+ | D_2)}.$$

Logo, para se obter $P(D_1 | T^+)$, precisamos primeiro conhecer $P(D_1)$, $P(D_2)$, $P(T^+ | D_1)$ e $P(T^+ | D_2)$.

$P(D_1)$ é a probabilidade de que um indivíduo na população geral tenha tuberculose. Como os 1.820 indivíduos no estudo descrito anteriormente não foram escolhidos aleatoriamente da população, a prevalência da doença não pode ser obtida a partir da informação na tabela. Em 1987, no entanto, houve 9,3 casos de tuberculose por 100.000 habitantes [8]. Com a disseminação do vírus de imunodeficiência humana (HIV), esse número aumentou drasticamente. Mas para esse exercício, podemos estimar

$$P(D_1) = 0,000093.$$

$P(D_2)$ é a probabilidade de que um indivíduo não tenha tuberculose. Desde que D_2 é o complemento de D_1 ,

$$\begin{aligned} P(D_2) &= 1 - P(D_1) \\ &= 1 - 0,000093 \\ &= 0,999907. \end{aligned}$$

$P(T^+ | D_1)$ é a probabilidade de um raio X positivo dado que um indivíduo tenha tuberculose — a sensibilidade do teste. Nesse estudo, a sensibilidade do raio X é

$$\begin{aligned} P(T^+ | D_1) &= \frac{22}{30} \\ &= 0,7333. \end{aligned}$$

$P(T^+ | D_2)$, a probabilidade de um raio X positivo dado que uma pessoa não tenha tuberculose, é o complemento da especificidade. Portanto,

$$\begin{aligned} P(T^+ | D_2) &= 1 - P(T^- | D_2) \\ &= 1 - \frac{1.739}{1.790} \\ &= 1 - 0,9715 \\ &= 0,0285. \end{aligned}$$

Usando toda essa informação, podemos agora calcular a probabilidade de que um indivíduo sofra de tuberculose, dado que tenha um raio X positivo; essa probabilidade é

$$\begin{aligned} P(D_1 | T^+) &= \frac{P(D_1) P(T^+ | D_1)}{P(D_1) P(T^+ | D_1) + P(D_2) P(T^+ | D_2)} \\ &= \frac{(0,000093)(0,7333)}{(0,000093)(0,7333) + (0,999907)(0,0285)} \\ &= 0,00239 \end{aligned}$$

Para cada 100.000 raios X positivos, somente 239 assinalam casos verdadeiros de tuberculose.

Note-se que, antes que um raio X seja aplicado, um indivíduo aleatoriamente selecionado da população dos Estados Unidos tem uma

$$9,3/100.000 = 0,000093 = 0,0093\%$$

probabilidade de ter tuberculose. Isso é chamado de *probabilidade a priori*. Depois que um raio X é aplicado e o resultado é positivo, o mesmo indivíduo tem uma

$$239/100.000 = 0,00239 = 0,239\%$$

probabilidade de estar com tuberculose. Essa é a *probabilidade a posteriori*. A probabilidade a posteriori leva em conta uma nova quantidade de informações — o resultado positivo do teste. Embora $99.761/100.000$ pessoas com raio X positivo realmente não tenham a doença, aumentamos muito a chance de se diagnosticar apropriadamente a tuberculose. Como $0,00239/0,000093 = 25,7$, a probabilidade de que um indivíduo com um raio X positivo tenha tuberculose é 25,7 vezes maior do que a de uma pessoa selecionada aleatoriamente da população.

6.4.3 Curvas ROC

A diagnose é um processo imperfeito. Teoricamente, é preferível ter um teste tanto altamente sensível como específico. Na realidade, entretanto, tal procedimento usualmente não é possível. Muitos testes estão realmente baseados em uma medida clínica que pode assumir uma série de valores; nesse caso, há um compromisso inerente entre a sensibilidade e a especificidade.

Considere a Tabela 6.1 que exibe os dados de um programa de transplante de rim no qual aloenxertos renais foram realizados [9]. O nível sérico de creatinina — composto químico encontrado no sangue e medido em miligramas percentuais — foi usado como ferramenta de diagnóstico para se detectar a rejeição potencial do transplante. Um nível aumentado de creatinina é freqüentemente associado com falha orgânica subsequente.

Se usarmos um nível maior que 2,9 mg% como indicador de rejeição iminente, o teste tem uma sensibilidade de 0,303 e uma especificidade de 0,909. Para aumentar a sensibilidade, podemos baixar o ponto de corte arbitrário que distingue um resultado de teste positivo de um negativo; se usamos 1,2 mg%, por exemplo, uma proporção muito maior dos resultados seria designada positiva. Nesse caso, dificilmente falharíamos em identificar um paciente que rejeitasse o órgão. Ao mesmo tempo, aumentaríamos a probabilidade de um falso resultado positivo por esse meio, o que diminui a especificidade. Ao aumentarmos a especificidade, rigorosamente sempre classificariam mal uma pessoa que não rejeitasse o órgão e diminuiríamos a sensibilidade. Geralmente, um teste de sensibilidade é mais útil quando a falha para se detectar uma doença o mais cedo possível tem consequências perigosas; um teste específico é importante em situações nas quais um resultado falso positivo é prejudicial.

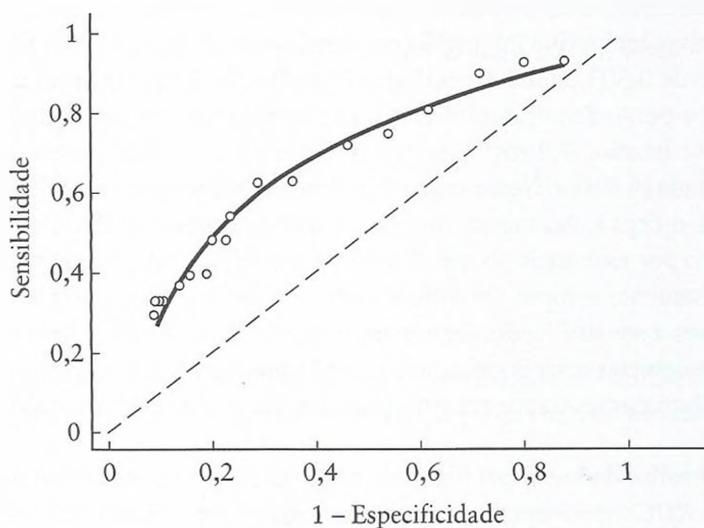
A relação entre sensibilidade e especificidade pode ser ilustrada usando-se um gráfico conhecido como *curva ROC* (*receiver operator characteristic curve*). Uma curva ROC é um gráfico de linha que plota a probabilidade de um resultado positivo verdadeiro — ou a sensibilidade do teste — versus a probabilidade de um resultado falso positivo para uma série de diferentes pontos de corte. Esses gráficos foram usados inicialmente no campo das comunicações. Como exemplo, a Figura 6.4 exibe uma curva ROC para os dados mostrados na Tabela 6.1. Quando um teste de diagnóstico existente é avaliado, esse tipo de gráfico pode

TABELA 6.1

Sensibilidade e especificidade do nível sérico de creatinina para prever a rejeição do transplante.

Creatinina Sérica (mg %)	Sensibilidade	Especificidade
1,2	0,939	0,123
1,3	0,939	0,203
1,4	0,909	0,281
1,5	0,818	0,380
1,6	0,758	0,461
1,7	0,727	0,535
1,8	0,636	0,649
1,9	0,636	0,711
2,0	0,545	0,766
2,1	0,485	0,773
2,2	0,485	0,803
2,3	0,394	0,811
2,4	0,394	0,843
2,5	0,364	0,870
2,6	0,333	0,891
2,7	0,333	0,894
2,8	0,333	0,896
2,9	0,303	0,909

ser usado como auxílio da avaliação da utilidade do teste e para determinar o ponto de corte mais apropriado. A linha tracejada na Figura 6.4 corresponde a um teste que dá resultados positivos e negativos somente ao acaso, e, portanto, não tem valor inerente. Quanto mais perto a linha está do canto superior esquerdo do gráfico, mais preciso é o teste. Além disso, o ponto que se encontra mais próximo desse canto é normalmente escolhido como o corte que maximiza simultaneamente tanto a sensibilidade como a especificidade.

**FIGURA 6.4**

Curva ROC para os níveis séricos de creatinina como um previsor de rejeição de transplante.

6.4.4 Cálculos de Prevalência

Além de serem usados em aplicações que envolvem o teorema de Bayes, os testes de diagnósticos ou triagens podem ser utilizados também para calcular a prevalência de doenças em uma população especificada. Por exemplo, o Departamento de Saúde do Estado de Nova York iniciou um programa para fazer a triagem para o HIV de todos os bebês nascidos em um período de 28 meses. Como os anticorpos maternais atravessam a placenta, a presença de anticorpos em um bebê sinaliza a infecção na mãe. Por serem os testes realizados anonimamente, nenhuma verificação dos resultados é possível. Os resultados da triagem registradas em todo o estado estão apresentados na Tabela 6.2 [10].

TABELA 6.2

Porcentagem positiva de HIV para bebês, por região, para o Estado de Nova York, dezembro de 1987 — março de 1990.

Região	Porcentagem Positivo	Número Positivo	Total Testado
Estado de Nova York não incluída a Cidade de Nova York	601	346.522	0,17
Arredores de NY	329	120.422	0,27
Vale do Médio Hudson	71	29.450	0,24
Interior Urbano	119	88.088	0,14
Interior Rural	82	108.562	0,08
Cidade de Nova York	3.650	294.062	1,24
Manhattan	799	50.364	1,59
Bronx	998	58.003	1,72
Brooklyn	1.352	104.613	1,29
Queens	424	67.474	0,63
Staten Island	77	13.608	0,57

Seja n^+ o número de recém-nascidos que apresentou resultado positivo e n o número total de bebês que passou pela triagem. Em cada região de Nova York, a prevalência sérica de HIV — ou $P(H)$, onde H é o evento de que uma mãe esteja infectada com o vírus — é calculada como n^+/n . Em Manhattan, por exemplo, 50.364 bebês foram testados e 799 dos resultados foram positivos. Nesse bairro,

$$\begin{aligned} \frac{n^+}{n} &= \frac{799}{50.364} \\ &= 0,0159. \end{aligned}$$

No entanto, há um problema aqui: a quantidade n^+/n realmente representa $P(T^+)$, a probabilidade de um resultado positivo de teste. Se o teste de triagem fosse perfeito, $P(H)$ e $P(T^+)$ seriam idênticos. Porém, o teste não é infalível; são possíveis tanto os resultados falsos positivos como os falsos negativos. De fato, ao se aplicar a lei da probabilidade total seguida da regra multiplicativa, a probabilidade verdadeira de um teste positivo é

$$\begin{aligned} P(T^+) &= P(T^+ \cap H) + P(T^+ \cap H^C) \\ &= P(T^+ | H)P(H) + P(T^+ | H^C)P(H^C) \\ &= P(T^+ | H)P(H) + [1 - P(T^- | H^C)][1 - P(H)]. \end{aligned}$$

Note-se que um resultado de teste positivo pode ocorrer de dois modos diferentes: a mãe está infectada com HIV ou não. Em adição à prevalência de infecção, essa equação incorpora tanto a sensibilidade como a especificidade do teste de diagnóstico.

Se n^+ / n é a probabilidade de um resultado de teste positivo, como calcularmos a prevalência do HIV? Ao usarmos a expressão para $P(T^+)$, podemos resolver para a quantidade verdadeira de interesse. Depois de alguma manipulação algébrica, verificamos que

$$\begin{aligned} P(H) &= \frac{P(T^+) - P(T^+ | H^C)}{P(T^+ | H) - P(T^+ | H^C)} \\ &= \frac{(n^+ / n) - P(T^+ | H^C)}{P(T^+ | H) - P(T^+ | H^C)}. \end{aligned}$$

Desde que a prevalência da infecção do HIV seja também uma probabilidade, seu valor deverá estar entre 0 e 1. Vamos examinar a expressão para $P(H)$. Para qualquer teste de triagem de valor,

$$P(T^+ | H) > P(T^+ | H^C);$$

a probabilidade de um resultado de teste positivo entre indivíduos infectados com HIV é maior do que a probabilidade entre os não infectados. Como resultado, o denominador da razão é positivo. Para $P(H)$ ser maior do que 0, exige-se que o numerador seja também positivo. Conseqüentemente, precisamos ter

$$\begin{aligned} \frac{n^+}{n} &> P(T^+ | H^C) \\ &= 1 - P(T^- | H^C). \end{aligned}$$

A proporção de testes com resultados positivos na população inteira precisa ser maior do que a de resultados positivos entre os que não estão infectados com HIV. Note-se que a especificidade do teste de triagem desempenha um papel crítico no cálculo de prevalência; se ela é muito baixa, pode não ser detectada pelo teste com especificidade inadequada.

Retornemos aos dados da Tabela 6.2. Não conhecemos a sensibilidade nem a especificidade do procedimento de diagnóstico usado, embora possamos assegurar que o teste não foi perfeito. Suponha, no entanto, que a sensibilidade do teste de triagem seja 0,99 e que sua especificidade seja 0,998; esses valores representam a extremidade mais alta do intervalo de valores possíveis. Lembre-se, também, de que a probabilidade de um resultado de teste positivo em Manhattan é 0,0159. Como resultado, a prevalência da infecção de HIV nesse bairro seria calculada como

$$\begin{aligned} P(H) &= \frac{0,0159 - (1 - 0,998)}{0,99 - (1 - 0,998)} \\ &= 0,0141, \end{aligned}$$

que é mais baixa do que a probabilidade de um resultado de teste positivo. Para o interior urbano de Nova York,

$$\begin{aligned} \frac{n^+}{n} &= \frac{119}{88.088} \\ &= 0,0014, \end{aligned}$$

e

$$\begin{aligned} P(H) &= \frac{0,0014 - (1 - 0,998)}{0,99 - (1 - 0,998)} \\ &= -0,0006. \end{aligned}$$

Mesmo com uma especificidade tão alta quanto 0,998, verificamos que a prevalência é negativa. Obviamente, esse é um resultado sem sentido porque, provavelmente, o procedimento do teste não foi preciso o suficiente para medir a prevalência muito baixa de HIV nessa região.

6.5 O Risco Relativo e a Razão de Chances

O conceito de risco relativo freqüentemente é útil quando queremos comparar as probabilidades de doença em dois diferentes grupos ou situações. O *risco relativo* — abreviado RR — é a chance que um membro de um grupo, ao receber alguma exposição, tem de desenvolver a doença relativamente à chance que um membro de outro grupo não-exposto terá de desenvolvê-la. Ele é definido como a probabilidade de doença no grupo exposto dividida pela probabilidade de doença no grupo não-exposto ou

$$RR = \frac{P(\text{doença} | \text{exposto})}{P(\text{doença} | \text{não-exposto})}$$

Considere um estudo que examina os fatores de risco para o câncer de mama entre as mulheres que participaram do primeiro Levantamento Nacional de Exame de Nutrição e de Saúde [11]. Em um *estudo de coorte* como esse, a exposição é medida no momento da investigação. Grupos de indivíduos com exposição e sem exposição — indivíduos sem exposição são freqüentemente chamados de *controle* — são acompanhados para se verificar casos de doenças. Nesse estudo de câncer de mama, uma mulher é considerada como “exposta” se deu à luz pela primeira vez com 25 anos ou mais velha. Em uma amostra de 4.540 mulheres que deram à luz seus primeiros filhos antes de 25 anos, 65 desenvolveram o câncer de mama. Das 1.628 mulheres que deram à luz com 25 anos ou mais velhas, 31 foram diagnosticadas com câncer de mama. Se assumimos que os números são suficientemente grandes para satisfazerem à definição freqüentista de probabilidade, o risco relativo de desenvolverem o câncer de mama é

$$\begin{aligned} RR &= \frac{P(\text{doença} | \text{exposto})}{P(\text{doença} | \text{não-exposto})} \\ &= \frac{31/1.628}{65/4.540} \\ &= 1,33. \end{aligned}$$

Um risco relativo de 1,33 implica que as mulheres que deram à luz pela primeira vez em idade mais avançada têm 33% mais probabilidade de desenvolver o câncer de mama do que as que deram à luz mais jovens. No Capítulo 15, explicaremos como determinar se essa é uma diferença importante.

Em geral, um risco relativo de 1,0 indica que as probabilidades de doença nos grupos exposto e não-exposto são idênticas; consequentemente, não existe uma associação entre a exposição e a doença. Um risco relativo maior do que 1,0 implica que há risco aumentado

da doença nos indivíduos expostos, enquanto um valor menor que 1,0 sugere que há um risco diminuído de que os indivíduos expostos desenvolvam a doença.

Note-se que o valor do risco relativo é independente das magnitudes das probabilidades relevantes; somente a razão dessas probabilidades é importante. Isso é especialmente útil quando estamos preocupados com eventos improváveis. Nos Estados Unidos, por exemplo, a probabilidade de um homem acima de 35 anos morrer de câncer de pulmão é de 0,002679 para os atualmente fumantes e de 0,000154 para os não-fumantes [12]. O risco relativo de morte para fumantes *versus* não-fumantes, no entanto, é

$$\begin{aligned} RR &= \frac{0,002679}{0,000154} \\ &= 17,4. \end{aligned}$$

Analogamente, a probabilidade de uma mulher com mais de 35 anos morrer de câncer no pulmão é de 0,001304 para as atualmente fumantes e de 0,000121 para as não-fumantes; o risco relativo é

$$\begin{aligned} RR &= \frac{0,001304}{0,000121} \\ &= 10,8. \end{aligned}$$

Embora lidemos com eventos de probabilidades muito baixas, o risco relativo nos permite ver que o fumo tem grande efeito na probabilidade de que um indivíduo em particular morra de câncer no pulmão.

Outra medida de probabilidades relativas de doenças comumente usada é a *razão de chances* ou *chance relativa* (OR, *odds ratio*, em inglês). Se um evento se realiza com probabilidade p , a chance em favor do evento é $p/(1-p)$ para 1. Se $p = 1/2$, por exemplo, a chance é $(1/2)/(1/2) = 1$ para 1. Nesse caso, o evento é igualmente provável de ocorrer como de não ocorrer. Se $p = 2/3$, a chance do evento é $(2/3)/(1/3) = 2$ para 1, a probabilidade de que o evento ocorra é duas vezes a probabilidade de que não ocorra. Analogamente, se para cada 100.000 indivíduos existem 9,3 casos de tuberculose, a chance de uma pessoa aleatoriamente selecionada ter a doença é

$$\frac{(9,3/100.000)}{(99.990,7/100.000)} = 0,00009301 \text{ para 1.}$$

Inversamente, sabemos que, se a chance em favor de um evento é a para b , a probabilidade de que o evento ocorra é $a/(a+b)$. A razão de chances é definida como a chance de doença entre os indivíduos expostos dividida pela chance de doença entre os indivíduos não-expostos ou

$$OR = \frac{P(\text{doença} | \text{exposto})/[1 - P(\text{doença} | \text{exposto})]}{P(\text{doença} | \text{não-exposto})/[1 - P(\text{doença} | \text{não-exposto})]}$$

Ela pode também ser definida como a chance de exposição entre os indivíduos doentes dividida pela chance de exposição entre os que não estão doentes ou

$$OR = \frac{P(\text{exposição} | \text{doentes})/[1 - P(\text{exposição} | \text{doentes})]}{P(\text{exposição} | \text{não-doentes})/[1 - P(\text{exposição} | \text{não-doentes})]}$$

Matematicamente, pode-se mostrar que essas duas definições para a chance relativa são equivalentes.

Considere os seguintes dados, tomados de outro estudo de fatores de risco para o câncer de mama. Esse é um estudo de caso-controle que examina os efeitos do uso de contraceptivos orais [13]. Em um *estudo de caso-controle*, os investigadores iniciam pela identificação dos grupos de indivíduos com a doença (os casos) e sem a doença (os controles) e retornam no tempo para determinar se a exposição em questão estava presente ou ausente para cada indivíduo. Entre as 989 mulheres que tinham câncer de mama no estudo, 273 usaram previamente contraceptivos orais e 716 não. Das 9.901 mulheres que não tiveram câncer de mama, 2.641 usaram contraceptivos orais e 7.260 não. Em um estudo de caso-controle, as proporções de indivíduos com e sem a doença são escolhidas pelo investigador; portanto, as probabilidades de doença nos grupos expostos e não-expostos não podem ser determinadas. No entanto, podemos calcular a probabilidade de exposição para ambos os casos e controles. Conseqüentemente, ao se usar a segunda definição para a razão de chances,

$$\begin{aligned} \text{OR} &= \frac{P(\text{exposição} | \text{doentes})/[1 - P(\text{exposição} | \text{doentes})]}{P(\text{exposição} | \text{não-doentes})/[1 - P(\text{exposição} | \text{não-doentes})]} \\ &= \frac{(273/989)/(1 - 273/989)}{(2.641/9.901)/(1 - 2.641/9.901)} \\ &= \frac{(273/989)/(716/989)}{(2.641/9.901)/(7.260/9.901)} \\ &= \frac{273/716}{2.641/7.260} \\ &= 1,05. \end{aligned}$$

Esses dados implicam que as mulheres que usaram contraceptivos orais têm chance de desenvolver o câncer de mama que é somente 1,05 vez a chance das que não usaram. Novamente, interpretaremos esse resultado no Capítulo 15. Tal como com o risco relativo, no entanto, uma razão de chances de 1,0 indica que a exposição não tem efeitos na probabilidade de doença.

O risco relativo e a razão de chances são duas medidas diferentes que tentam explicar o mesmo fenômeno. Embora o risco relativo possa parecer mais intuitivo, a razão de chances tem melhores propriedades estatísticas, as quais serão explicadas mais tarde, no texto. Em qualquer evento, para doenças raras, a razão de chances é uma grande aproximação do risco relativo. Para se verificar isso, se

$$P(\text{doença} | \text{exposto}) \approx 0$$

e

$$P(\text{doença} | \text{não-exposto}) \approx 0,$$

então

$$1 - P(\text{doença} | \text{exposto}) \approx 1$$

e

$$1 - P(\text{doença} | \text{não-exposto}) \approx 1.$$

Logo,

$$\begin{aligned}
 OR &= \frac{P(\text{doença} | \text{exposto})/[1 - P(\text{doença} | \text{exposto})]}{P(\text{doença} | \text{não-exposto})/[1 - P(\text{doença} | \text{não-exposto})]} \\
 &= \frac{P(\text{doença} | \text{exposto})/1}{P(\text{doença} | \text{não-exposto})/1} \\
 &= \frac{P(\text{doença} | \text{exposto})/1}{P(\text{doença} | \text{não-exposto})/1} \\
 &= RR.
 \end{aligned}$$

Quando usamos razões de chances e riscos relativos, devemos tomar cuidado para colarmos a informação obtida dentro do contexto. Como mencionamos anteriormente, os valores numéricos dessas medidas não refletem as magnitudes das probabilidades usadas para calculá-las. Para ilustrar esse ponto, um terceiro estudo de câncer de mama — que investiga os efeitos do uso de hormônio em mulheres pós-menopausa — concluiu que as mulheres que usaram a terapia de hormônio durante cinco a nove anos têm uma chance de desenvolver câncer de mama invasivo 1,46 vez a chance de as mulheres que nunca usaram hormônios [14]. Pode parecer um aumento bastante significativo ao risco. No entanto, o efeito desse aumento depende da probabilidade básica da doença para as mulheres que não foram expostas à terapia de hormônios. Tem sido registrado que uma mulher de 60 anos tem uma chance de 3,59% de desenvolver câncer de mama nos próximos dez anos [15]. Nesse caso, teríamos

$$OR = 1,46$$

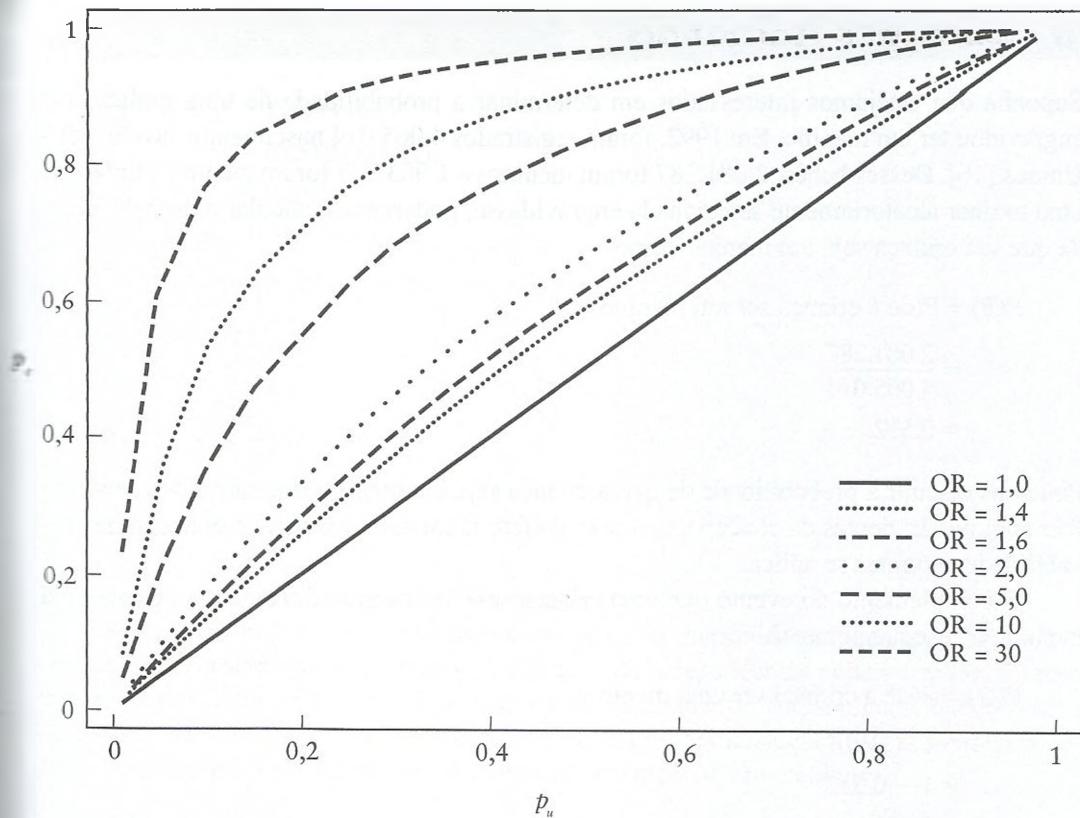
$$\begin{aligned}
 &= \frac{P(\text{câncer} | \text{uso de hormônio})/[1 - P(\text{câncer} | \text{uso de hormônio})]}{P(\text{câncer} | \text{sem uso})/[1 - P(\text{câncer} | \text{sem uso})]} \\
 &= \frac{P(\text{câncer} | \text{uso de hormônio})/[1 - P(\text{câncer} | \text{uso de hormônio})]}{0,0359/(1 - 0,0359)}
 \end{aligned}$$

e assim que

$$P(\text{câncer} | \text{uso de hormônio}) = 0,0516.$$

Embora a razão de chances de 1,46 seja relativamente alta, a mudança na probabilidade real da doença de 3,59% para 5,16% não é alarmante.

A Figura 6.5 ilustra a relação entre a probabilidade de um resultado e a razão de chances. No gráfico, $p_{\text{não-exp}}$ representa a probabilidade básica de doença no grupo não-exposto e p_{exp} é a probabilidade mais alta de doença no grupo que sofreu alguma exposição. Se a razão de chances é igual a 1,0, $p_{\text{não-exp}}$ e p_{exp} devem ser as mesmas, independentemente do valor de $p_{\text{não-exp}}$. Por outro lado, se a razão de chances é igual a 2,0, nossa interpretação depende do valor de $p_{\text{não-exp}}$. Por exemplo, se a probabilidade de doença no grupo não-exposto é 0,05, no grupo exposto é 0,095, um aumento de 90%. Se a probabilidade de doença entre os não-expostos é 0,50, no entanto, a probabilidade entre os expostos é 0,667, um aumento de somente 33%. A Tabela 6.3 mostra a relação entre $p_{\text{não-exp}}$, p_{exp} , a razão de chances e o risco relativo para uma variedade de probabilidades básicas.

**FIGURA 6.5**

Relação entre as probabilidades de um evento em um grupo exposto e em um não-exposto e as razões de chances.

TABELA 6.3

Relações entre as probabilidades de um evento em um grupo exposto e um não-exposto, razão de chances e o risco relativo.

Razão de Chances	p_u	p_e	Risco Relativo
1,2	0,01	0,012	1,20
1,2	0,05	0,059	1,19
1,2	0,25	0,286	1,14
1,2	0,50	0,545	1,09
1,4	0,01	0,014	1,39
1,4	0,05	0,069	1,37
1,4	0,25	0,318	1,27
1,4	0,50	0,583	1,17
2,0	0,01	0,020	1,98
2,0	0,05	0,095	1,90
2,0	0,25	0,400	1,60
2,0	0,50	0,667	1,33

6.6 Aplicações Adicionais

Suponha que estejamos interessados em determinar a probabilidade de uma mulher que engravidou ter um menino. Em 1992, foram registrados 4.065.014 nascimentos nos Estados Unidos [16]. Desses bebês, 2.081.287 foram meninos e 1.983.727 foram meninas. Então, se uma mulher aleatoriamente selecionada engravidasse, poderíamos calcular a probabilidade de que sua criança seja um menino como

$$\begin{aligned} P(B) &= P(\text{de a criança ser um menino}) \\ &= \frac{2.081.287}{4.065.014} \\ &= 0,512. \end{aligned}$$

Podemos discutir a probabilidade de que a criança seja um menino somente antes de a mulher engravidar; depois da concepção, o sexo do feto já foi determinado e o conceito de probabilidade não mais se aplica.

O complemento do evento que uma criança seja um menino é o evento que seja uma menina. Conseqüentemente,

$$\begin{aligned} P(G) &= P(\text{de a criança ser uma menina}) \\ &= 1 - P(B) \\ &= 1 - 0,512 \\ &\approx 0,488. \end{aligned}$$

Como a criança é menino ou menina, esses dois eventos são mutuamente exclusivos. Nesse caso, a regra aditiva de probabilidade estabelece que a probabilidade de qualquer um dos eventos ocorrer é igual à soma das probabilidades dos eventos individuais; assim,

$$\begin{aligned} P(B \cup G) &= P(B) + P(G) \\ &= 0,512 + 0,488 \\ &= 1,000. \end{aligned}$$

A soma das probabilidades desses dois eventos é 1, o que indica que são exaustivos. A criança só pode ser um menino ou uma menina; não existem outros resultados possíveis.

Suponha agora que selecionemos aleatoriamente duas mulheres de uma população e que ambas engravidem. Qual a probabilidade de que ambas as crianças sejam meninos? Sabemos que os dois eventos são independentes; o sexo da criança da primeira mulher não tem efeito no sexo da criança da segunda. Então, ao se usar a regra multiplicativa de probabilidade para eventos independentes e representar o evento de que ambas as crianças sejam meninos por $B_1 \cap B_2$,

$$\begin{aligned} P(B_1 \cap B_2) &= P(B_1)P(B_2) \\ &= (0,512)(0,512) \\ &= 0,262. \end{aligned}$$

Há três outros eventos possíveis: $B_1 \cap G_2$, a criança da primeira mulher será um menino e a da segunda mulher, uma menina; $G_1 \cap B_2$, a primeira mulher terá uma menina e a segun-

da um menino; $G_1 \cap G_2$, ambas as crianças serão meninas. As probabilidades desses eventos são

$$\begin{aligned} P(B_1 \cap G_2) &= P(B_1)P(G_2) \\ &= (0,512)(0,488) \\ &= 0,250, \end{aligned}$$

$$\begin{aligned} P(G_1 \cap B_2) &= P(G_1)P(B_2) \\ &= (0,488)(0,512) \\ &= 0,250, \end{aligned}$$

e

$$\begin{aligned} P(G_1 \cap G_2) &= P(G_1)P(G_2) \\ &= (0,488)(0,488) \\ &= 0,238. \end{aligned}$$

Note-se que essas quatro probabilidades somam 1.

Se escolhermos três mulheres da população e elas engravidarem, qual a probabilidade de que as três crianças sejam meninas? O conceito de independência pode ser estendido para três ou mais eventos diferentes; nesse caso, o sexo da criança de uma mulher não afeta o sexo de qualquer outra criança. A regra multiplicativa de probabilidade para eventos independentes estabelece que a probabilidade de que as três crianças sejam meninas é

$$\begin{aligned} P(G_1 \cap G_2 \cap G_3) &= P(G_1)P(G_2)P(G_3) \\ &= (0,488)(0,488)(0,488) \\ &= 0,116. \end{aligned}$$

Retornando ao exemplo em que selecionamos somente duas mulheres, qual a probabilidade de ambas as crianças serem meninos, dado que pelo menos uma criança seja um menino? A probabilidade de que um evento em particular ocorrerá desde que outro já tenha ocorrido é conhecida como probabilidade condicional. Representando o evento de que pelo menos uma criança seja um menino por A e aplicando a fórmula para uma probabilidade condicional,

$$\begin{aligned} P(B_1 \cap B_2 | A) &= P(\text{ambas as crianças sejam meninos} | \text{pelo menos uma seja um menino}) \\ &= \frac{P[(B_1 \cap B_2) \cap A]}{P(A)} \\ &= \frac{P(B_1 \cap B_2)}{P(A)}. \end{aligned}$$

O evento de que ambas as crianças sejam meninos e pelo menos uma seja menino é simplesmente o evento de que ambas as crianças sejam meninos. Já sabemos que $P(B_1 \cap B_2) = 0,262$. Qual é a $P(A)$, probabilidade de que seja pelo menos um menino? Note-se que esse evento pode ocorrer de três modos diferentes — ambas as crianças serão meninos, a primeira será um menino e a segunda uma menina ou a primeira será uma menina e a segunda um menino. Desde que esses três eventos são mutuamente exclusivos, aplicamos a regra aditiva para encontrar

$$\begin{aligned}
 P(A) &= P[(B_1 \cap B_2) \cup (B_1 \cap G_2) \cup (G_1 \cap B_2)] \\
 &= P(B_1 \cap B_2) + P(B_1 \cap G_2) + P(G_1 \cap B_2) \\
 &= 0,262 + 0,250 + 0,250 \\
 &= 0,762.
 \end{aligned}$$

Logo,

$$\begin{aligned}
 P(B_1 \cap B_2 | A) &= \frac{P(B_1 \cap B_2)}{P(A)} \\
 &= \frac{0,262}{0,762} \\
 &= 0,344.
 \end{aligned}$$

Se sabemos que pelo menos uma criança é um menino, a probabilidade de que ambas as crianças sejam meninos aumenta de 0,262 para 0,344.

À primeira vista, esse resultado pode parecer não-intuitivo. Estamos falando que uma criança é um menino; consequentemente, poderíamos esperar que a probabilidade de que a outra criança seja um menino seja simplesmente $P(B) = 0,512$. Em vez disso, calculamos a probabilidade como sendo 0,344. O ponto importante é que não especificamos qual das duas crianças seria um menino. Nesse exemplo, a ordem é importante. Quando tratamos com probabilidades, a resposta aparentemente óbvia nem sempre é a correta; cada problema precisa ser considerado cuidadosamente.

As probabilidades condicionais entram em ação freqüentemente quando trabalhamos com tábuas de vida. Suponha que gostaríamos de conhecer a probabilidade de uma pessoa atingir 80 anos de idade sendo que agora ela tem 40 anos. Seja A representante do evento em que o indivíduo tem 40 anos e B do evento em que ele atinja os 80 anos. Usando a Tabela 5.1, a tábua de vida de 1992 para a população dos Estados Unidos

$$\begin{aligned}
 P(A) &= P(\text{de uma pessoa atingir os 40 anos}) \\
 &= \frac{95.527}{100.000} \\
 &= 0,95527,
 \end{aligned}$$

e

$$\begin{aligned}
 P(A \cap B) &= P(\text{de uma pessoa atingir os 40 anos e os 80 anos}) \\
 &= P(\text{de uma pessoa atingir os 80 anos}) \\
 &= \frac{48.460}{100.000} \\
 &= 0,48460.
 \end{aligned}$$

Logo,

$$\begin{aligned}
 P(B | A) &= P(\text{de uma pessoa atingir os 80 anos} | \text{que tenha agora 40 anos}) \\
 &= \frac{P(A \cap B)}{P(A)} \\
 &= \frac{0,48460}{0,95527} \\
 &= 0,5073.
 \end{aligned}$$

Se um indivíduo tem 40 anos, sua chance de sobreviver até os 80 é maior do que era ao nascer.

Se A_1 e A_2 são eventos mutuamente exclusivos e exaustivos, de modo que

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) \\ &= 1, \end{aligned}$$

o teorema de Bayes estabelece que

$$P(A_1 | B) = \frac{P(A_1) P(B | A_1)}{P(A_1) P(B | A_1) + P(A_2) P(B | A_2)}.$$

O teorema de Bayes é importante em testes de diagnósticos, porque relaciona o valor preditivo de um teste com sua sensibilidade e com sua especificidade, assim como a prevalência da doença na população testada.

Considere os seguintes dados, tomados de um estudo que investiga a precisão de três marcas de testes caseiros de gravidez [17]. Seja A_1 representante do evento que uma mulher esteja grávida, A_2 do evento de que ela não esteja grávida e T^+ do resultado do teste de gravidez caseiro positivo. A sensibilidade média para as três marcas de kits de teste é 80%; portanto,

$$P(T^+ | A_1) = 0,80.$$

Conseqüentemente, a probabilidade de um resultado falso negativo é

$$\begin{aligned} P(T^- | A_1) &= 1 - 0,80 \\ &= 0,20 \end{aligned}$$

A especificidade dos testes de gravidez caseiros é de 68%; então,

$$P(T^- | A_2) = 0,68$$

e a probabilidade de um falso positivo é

$$\begin{aligned} P(T^+ | A_2) &= 1 - 0,68 \\ &= 0,32. \end{aligned}$$

Qual a probabilidade de que uma mulher com um resultado positivo de um kit de teste caseiro esteja realmente grávida?

Suponha que na população testada, $P(A_1) = 0,60$; isto é, 60% das mulheres que usam os testes caseiros de gravidez estejam realmente grávidas. Desde que A_2 é o complemento de A_1 , a probabilidade de que uma mulher não esteja grávida é

$$\begin{aligned} P(A_2) &= 1 - P(A_1) \\ &= 1 - 0,60 \\ &= 0,40. \end{aligned}$$

Usando o teorema de Bayes, o valor preditivo de um teste positivo é

$$\begin{aligned} P(A_1 | T^+) &= \frac{P(A_1) P(T^+ | A_1)}{P(A_1) P(T^+ | A_1) + P(A_2) P(T^+ | A_2)} \\ &= \frac{(0,60)(0,80)}{(0,60)(0,80) + (0,40)(0,32)} \\ &= 0,79. \end{aligned}$$

Logo, o resultado positivo de um kit de teste caseiro aumenta a probabilidade de que uma mulher nessa população esteja grávida de 0,60 para 0,79.

Qual a probabilidade de que uma mulher não esteja grávida se o resultado de seu teste caseiro for negativo? Novamente aplicando o teorema de Bayes, o valor preditivo de um teste negativo é

$$\begin{aligned} P(A_2 | T^-) &= \frac{P(A_2) P(T^- | A_2)}{P(A_2) P(T^- | A_2) + P(A_1) P(T^- | A_1)} \\ &= \frac{(0,40)(0,68)}{(0,40)(0,68) + (0,60)(0,20)} \\ &= 0,69. \end{aligned}$$

Um resultado negativo do kit de teste caseiro aumenta a probabilidade de que uma mulher não esteja grávida de 0,40 para 0,69.

Quando desejamos comparar as probabilidades de um evento específico em dois diferentes grupos, o conceito de risco relativo é freqüentemente utilizado. Considere a Figura 6.6. Esse gráfico de barras ilustra os riscos de câncer de pulmão entre mulheres que têm fumado 21 cigarros por dia ou mais em relação às mulheres que nunca fumaram [12,18]. Para o grupo de pessoas que parou de fumar nos últimos dois anos, por exemplo,

$$\begin{aligned} RR &= \frac{P(\text{câncer do pulmão} | \text{parou nos últimos dois anos})}{P(\text{câncer do pulmão} | \text{não-fumantes})} \\ &= 32,4. \end{aligned}$$

É surpreendente que esse risco relativo seja maior do que o risco correspondente para os atualmente fumantes; no entanto, há muitas pessoas nesse grupo que foram forçadas a parar por causa de suas doenças. No decorrer do tempo, o risco de câncer no pulmão mesmo entre os fumantes inveterados decresce gradualmente depois que um indivíduo pára de fumar.

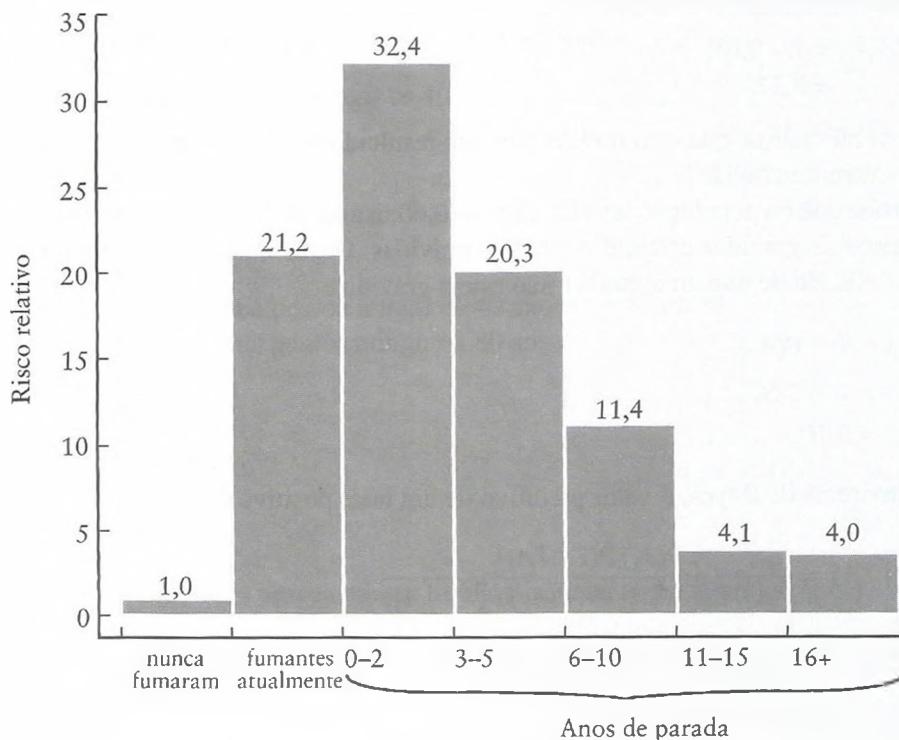


FIGURA 6.6

Riscos relativos de câncer de pulmão em mulheres ex-fumantes, 21 cigarros por dia ou mais.

A razão de chances é outra medida freqüentemente usada para comparar as probabilidades de um evento em dois diferentes grupos. Ao contrário do risco relativo, que compara as probabilidades diretamente, a razão de chances (como seu nome sugere) relaciona as chances do evento em duas populações. Para mulheres que fumaram 21 cigarros por dia ou mais, mas que pararam nos últimos dois anos, a chance de desenvolver câncer do pulmão relativa à chance das mulheres que nunca fumaram seria calculada como

$$\text{OR} = \frac{P(\text{câncer do pulmão} | \text{parou})/[1 - P(\text{câncer do pulmão} | \text{parou})]}{P(\text{câncer do pulmão} | \text{não-fumante})/[1 - P(\text{câncer do pulmão} | \text{não-fumante})]}$$

Para doenças raras como o câncer do pulmão, a razão de chances é muito próxima do risco relativo.

6.7 Exercícios de Revisão

- Qual é a definição freqüentista de probabilidade?
- Quais são as três operações básicas que podem ser realizadas nos eventos?
- Explique a diferença entre eventos mutuamente exclusivos e independentes.
- Qual é o valor do teorema de Bayes? Como ele se aplica aos testes de diagnósticos?
- O que aconteceria se você tentasse aumentar a sensibilidade de um teste de diagnóstico?
- Como as probabilidades de doença em dois diferentes grupos podem ser comparadas?
- Seja A representante do evento de que um indivíduo em particular esteja exposto a altos níveis de monóxido de carbono e B do evento de que ele esteja exposto a altos níveis de dióxido de nitrogênio.
 - Qual é o evento $A \cap B$?
 - Qual é o evento $A \cup B$?
 - Qual é o complemento de A ?
 - Os eventos A e B são mutuamente exclusivos?
- Para bebês americanos descendentes de mexicanos nascidos no Arizona em 1986 e 1987, a probabilidade de que sua idade gestacional seja menor que 37 semanas é 0,142 e a probabilidade de que seu peso ao nascer seja menor do que 2.500 gramas é 0,051 [19] e a probabilidade de que esses eventos ocorram simultaneamente é 0,031.
 - Seja A o evento de que a idade gestacional do bebê seja menor do que 37 semanas e B o evento de que seu peso ao nascer seja menor do que 2.500 gramas. Construa um diagrama de Venn para ilustrar a relação entre os eventos A e B .
 - A e B são independentes?
 - Para um recém-nascido americano descendente de mexicanos aleatoriamente selecionado, qual é a probabilidade de que A ou B ou ambos ocorram?
 - Qual é a probabilidade de que o evento A ocorra dado que se sabe que o evento B ocorreu?
- Considere as seguintes estatísticas de natalidade para a população dos Estados Unidos em 1992 [16]. De acordo com esses dados, as probabilidades de que uma mulher aleatoriamente selecionada que deu à luz em 1992 estivesse em cada um dos seguintes grupos de idade estão como segue:

Idade	Probabilidade
< 15	0,003
15–19	0,124
20–24	0,263
25–29	0,290
30–34	0,220
35–39	0,085
40–44	0,014
45–49	0,001
Total	1,000

- (a) Qual é a probabilidade de que uma mulher que deu à luz em 1992 tenha 24 anos ou seja mais jovem?
- (b) Qual é a probabilidade de que ela tenha 40 anos ou seja mais velha?
- (c) Se a mãe de uma criança em particular tem menos que 30 anos, qual é a probabilidade de que ela ainda não tenha 20 anos?
- (d) Se a mãe tiver 35 anos ou for mais velha, qual a probabilidade de que tenha menos de 40 anos?
10. As probabilidades associadas com a principal fonte de pagamento esperada para gastos hospitalares nos Estados Unidos em 1990 estão listadas abaixo [20].
- | Principal Fonte de Pagamento | Probabilidade |
|---------------------------------|---------------|
| Seguro privado | 0,387 |
| Medicare | 0,345 |
| Medicaid | 0,116 |
| Outros programas governamentais | 0,033 |
| Autopagamentos | 0,058 |
| Outros/ Sem custos | 0,028 |
| Não-definido | 0,033 |
| Total | 1,000 |
- (a) Qual a probabilidade de que a principal fonte de pagamento para uma despesa hospitalar seja o seguro privado de um paciente?
- (b) Qual a probabilidade de que a principal fonte de pagamento seja o Medicare, o Medicaid ou outro programa governamental?
- (c) Se a principal fonte de pagamento for um programa governamental, qual a probabilidade de ser o Medicare?
11. Observando a população dos Estados Unidos em 1993, a probabilidade de que um adulto entre 45 e 64 anos não tenha cobertura de seguro saúde de algum tipo é 0,123 [21].
- (a) Suponha que você selecione aleatoriamente uma mulher de 47 anos e um homem de 59 anos não-aparentado dessa população. Qual a probabilidade de que ambos sejam não-segurados?
- (b) Qual a probabilidade de que ambos tenham cobertura de seguro saúde?
- (c) Se cinco adultos não-aparentados entre 45 e 64 anos são escolhidos da população, qual a probabilidade de que sejam não-segurados?
12. Consulte a Tabela 5.1, a tábua de vida resumida para os Estados Unidos [1].
- (a) Qual a probabilidade de que um bebê recém-nascido esteja vivo em seu quinto aniversário?
- (b) Qual a probabilidade de que um indivíduo com 60 anos sobreviva nos próximos dez anos?
- (c) Considere um homem e uma mulher que são casados e que tenham 60 anos. Qual probabilidade de que tanto a mulher como o homem estejam vivos em seus 70º aniversário? Assuma que os dois eventos sejam independentes.
- (d) Qual a probabilidade de que tanto a mulher como o marido, mas não ambos, estejam vivos com 70 anos?
13. Um estudo registrou que a sensibilidade da mamografia como teste de triagem para detecção de câncer de mama é 0,85, enquanto sua especificidade é 0,80 [22].
- (a) Qual a probabilidade de um resultado de teste falso negativo?
- (b) Qual a probabilidade de um resultado falso positivo?
- (c) Na população na qual a probabilidade de que uma mulher tenha câncer de mama é 0,0025, qual a probabilidade de que tenha câncer, se sua mamografia for positiva?

14. O Instituto Nacional de Segurança Ocupacional e de Saúde desenvolveu uma definição de caso de síndrome de túnel carpal — uma doença do punho — que incorpora três critérios: sintomas de envolvimento do nervo, história de fatores de risco ocupacional e a presença de materiais de exames físicos [23]. A sensibilidade dessa definição como um teste para a síndrome de túnel carpal é de 0,67; sua especificidade é de 0,58.

- (a) Em uma população cuja prevalência da síndrome de túnel carpal está estimada em 15%, qual o valor preditivo de um resultado positivo de teste?
- (b) Como esse valor preditivo se modifica, se a prevalência for somente 10%? E se for 5%?
- (c) Construa um diagrama — como o da Figura 6.3 — que ilustre os resultados do processo de teste de diagnóstico. Inicie com uma população de 1.000.000 de pessoas cuja prevalência da síndrome de túnel carpal seja 15%.

15. Os dados seguintes são tomados de um estudo que investiga o uso de uma técnica chamada ventriculografia radionuclidica como teste de diagnóstico para se detectar doença da artéria coronária [24].

Teste	Doença		Total
	Presente	Ausente	
Positivo	302	80	382
Negativo	179	372	551
Total	481	452	933

- (a) Qual a sensibilidade da ventriculografia radionuclidica nesse estudo? Qual a sua especificidade?
 - (b) Para uma população cuja prevalência da doença da artéria coronária seja 0,10, calcule a probabilidade de que um indivíduo tenha a doença, sendo que ele apresenta resultado positivo usando a ventriculografia radionuclidica.
 - (c) Qual o valor preditivo de um teste negativo?
16. A tabela abaixo exibe dados tomados de um estudo que compara o *status* de fumantes auto-registrados com o nível sérico de cotinina [25]. Como parte do estudo, o nível de cotinina foi usado como ferramenta de diagnóstico para predizer o *status* de fumante; o *status* auto-registrado foi considerado verdadeiro. Para uma série de pontos de corte, as sensibilidades e as especificidades observadas estão abaixo.

Nível de Cotinina (ng/ml)	Sensibilidade	Especificidade
5	0,971	0,898
7	0,964	0,931
9	0,960	0,946
11	0,954	0,951
13	0,950	0,954
14	0,949	0,956
15	0,945	0,960
17	0,939	0,963
19	0,932	0,965

- (a) Quando o ponto de corte é aumentado, como a probabilidade de um resultado falso positivo se modifica? Como a probabilidade de um resultado falso negativo se modifica?

- (b) Use esses dados para construir uma curva ROC.
 (c) Com base nesse gráfico, que valor de nível sérico de cotinina você escolheria como ponto de corte ideal para predizer o *status* de fumante? Por quê?
17. A Tabela 6.2 mostra as porcentagens de recém-nascidos com HIV positivo para várias regiões do Estado de Nova York [10].
 (a) No Brooklin, qual a probabilidade de um teste com resultado positivo?
 (b) Assuma que a sensibilidade do teste de triagem usado é 0,99 e que sua especificidade é 0,998. Qual a prevalência da infecção de HIV nesse bairro?
 (c) Qual a prevalência da infecção de HIV no Bronx?
18. Por diversos métodos de contracepção, as probabilidades de que uma mulher casada tenha uma gravidez não-planejada durante o primeiro ano de uso estão abaixo [26].

Método de Contracepção	Probabilidade de Gravidez
Nenhum	0,431
Diafragma	0,149
Camisa de Vênus	0,106
DIU	0,071
Pílula	0,037

Para cada método listado, calcule o risco relativo de gravidez para mulheres que usam o método *versus* as que não usam qualquer tipo de proteção. Como o risco se modifica com relação ao método de contracepção?

19. Um estudo sobre doenças respiratórias com base em uma comunidade durante o primeiro ano de vida foi conduzido na Carolina do Norte. Como parte desse estudo, um grupo de crianças foi classificado de acordo com o *status* socioeconômico da família. Os números de crianças em cada grupo que experimentaram sintomas respiratórios persistentes estão mostrados abaixo [27].

Status socioeconômico	Número de crianças	Número com sintomas
Baixo	79	31
Médio	122	29
Alto	192	27

- (a) Use esses dados para calcular a probabilidade de cada grupo socioeconômico sofrer de sintomas respiratórios persistentes. Assuma que os números são suficientemente grandes para satisfazer à definição freqüentista de probabilidade.
 (b) Calcule a chance de os grupos socioeconômicos baixo e médio experimentarem sintomas respiratórios persistentes com relação ao grupo socioeconômico alto.
 (c) Entre o *status* socioeconômico e os sintomas respiratórios, parece haver alguma associação?
20. Um estudo que investiga o uso de glicofita capilar (em inglês FCG – *fasting capillary glycemia*) foi conduzido — o nível de glicose no sangue para indivíduos que não se alimentaram em um número de horas especificado — como um teste de triagem para diabetes [28]. Os pontos de corte do FCG, variando de 3,9 até 8,9 mmol/litro, foram examinados; as sensibilidades e especificidades do teste correspondente a esses diferentes níveis estão contidos no conjunto de dados diabetes (Apêndice B, Tabela B.11). Os níveis de FCG estão salvos sob a variável de nome *fcg*, as sensibilidades sob *sens* e as especificidades sob *spec*.

- (a) Como a sensibilidade do teste de triagem se modifica quando o ponto de corte é elevado de 3,9 para 8,9 mmol/l? Como a especificidade se modifica?
- (b) Use esses dados para construir uma curva ROC.
- (c) Os investigadores que conduziram esse estudo escolheram um nível de FCG de 5,6 mmol/litro como o ponto de corte ideal para predizer diabetes. Você concorda com essa escolha? Por quê?

Bibliografia

- [1] National Center for Health Statistics. KOCHANEK, K. D. e HUDSON, B. L. "Advanced Report of Final Mortality Statistics, 1992". *Monthly Vital Statistics Report*. v. 43, n. 6, 22 mar. 1995.
- [2] National Center for Health Statistics. COLLINS, J. G. "Types of Injuries and Impairments Due to Injuries, United States". *Vital and Health Statistics*. série 10, n. 159, nov. 1986.
- [3] YOBS, A. R., SWANSON, R. A. e LAMOTTE, L. C. "Laboratory Reliability of the Papancolaou Smear". *Obstetrics and Gynecology*. v. 65, fev. 1985. p. 235–244.
- [4] DEVESA, S. S., SILVERMAN, D. T., YOUNG, J. L., POLLACK, E. S., BROWN, C. C., HORM, J. W., PERCY, C. L., MYERS, M. H., MCKAY, F. W. e FRAUMENI, J. F. "Cancer Incidence and Mortality Trends Among Whites in the United States, 1949–1984". *Journal of the National Cancer Institute*. v. 79, out. 1987. p. 701–770.
- [5] HENIG, R. M. "Is the Pap Test Valid?". *The New York Times Magazine*. 28 maio. 1989. p. 37–38.
- [6] FAHEY, M. T., IRWIG, L. e MACASKILL, P. "Meta-analysis of Pap Test Accuracy". *American Journal of Epidemiology*. v. 141, 1º abr. 1995. p. 680–689.
- [7] YERUSHALMY, J., HARKNESS, J. T., COPE, J. H. e KENNEDY, B. R. "The Role of Dual Reading in Mass Radiography". *American Review of Tuberculosis*. v. 61, abr. 1950. p. 443–464.
- [8] Centers for Disease Control. "A Strategic Plan for the Elimination of Tuberculosis in the United States". *Morbidity and Mortality Weekly Report*. v. 38, n. 16, 28 abr. 1989.
- [9] DELONG, E. R., VERNON, W. B. e BOLLINGER, R. R. "Sensitivity and Specificity of a Monitoring Test". *Biometrics*. v. 41, dez. 1985. p. 947–958.
- [10] NOVICK, L. F., GLEBATIS, D. M., STRICOFF, R. L., MACCUBBIN, P. A., LESSNER, L. e BERNS, D. S. "Newborn Seroprevalence Study: Methods and Results". *American Journal of Public Health*. v. 81, maio 1991. p. 15–21.
- [11] CARTER, C. L., JONES, D. Y., SCHATZKIN, A. e BRINTON, L. A. "A Prospective Study of Reproductive, Familial, and Socioeconomic Risk Factors for Breast Cancer Using NHANES I Data". *Public Health Reports*. v. 104, jan./fev. 1989. p. 45–49.
- [12] GARFINKEL, L. e SILVERBERG, E. "Lung Cancer and Smoking Trends in the United States Over the Past 25 Years". *Ca—A Cancer Journal for Clinicians*. v. 41, maio/jun. 1991. p. 137–145.
- [13] HENNEKENS, C. H., SPEIZER, F. E., LIPNICK, R. J., ROSNER, B., BAIN, C., BELANGER, C., STAMPFER, M. J., WILLETT, W. e Peto, R. "A Case-Control Study of Oral Contraceptive Use and Breast Cancer". *Journal of the National Cancer Institute*. v. 72, jan. 1984. p. 39–42.
- [14] COLDITZ, G. A., HANKINSON, S. E., HUNTER, D. J., WILLETT, W. C., MANSON, J. E., STAMPFER, M. J., HENNEKENS, C., ROSNER, B. e Speizer, F. E. "The Use of Estrogens and Progestins and the Risk of Breast Cancer in Postmenopausal Women". *The New England Journal of Medicine*. v. 332, 15 jun. 1995. p. 1589–1593.
- [15] FEUER, E. J., WUN, L. M., BORING, C. C., FLANDERS, W. D., TIMMEL, M. J. e TONG, T. "The Lifetime Risk of Developing Breast Cancer". *Journal of the National Cancer Institute*. v. 85, 2 jun. 1993. p. 892–897.
- [16] National Center for Health Statistics. VENTURA, S. J., MARTIN, J. A., TAFFEL, S. M., MATHEWS, T. J. e CLARKE, S. C. "Advanced Report of Final Natality Statistics, 1992". *Monthly Vital Statistics Report*. v. 43, n. 5, 25 out. 1994.

7

Distribuições Teóricas de Probabilidade

Qualquer característica que pode ser medida ou categorizada é chamada *variável*. Se uma variável pode assumir uma série de valores diferentes tal que qualquer resultado particular seja determinado pela sorte, ela é uma *variável aleatória*. Já vimos diversas variáveis aleatórias nos capítulos anteriores, embora não usássemos esse termo. No Capítulo 2, por exemplo, o nível sérico de colesterol de um homem de 25 a 34 anos, nos Estados Unidos, é uma variável aleatória; no Capítulo 3, o volume expiratório forçado em um segundo para um adolescente que sofre de asma é outra. Variáveis aleatórias são representadas tipicamente por letras maiúsculas tais como X , Y e Z . Uma *variável aleatória discreta* pode assumir somente um número finito ou enumerável de resultados. Um exemplo é o *status conjugal*: um indivíduo pode ser solteiro, casado, divorciado ou viúvo. Outro exemplo seria o número de infecções auditivas que um bebê desenvolve durante seu primeiro ano de vida. Uma *variável aleatória contínua*, tal como peso ou altura, pode tomar qualquer valor dentro de um intervalo especificado ou contínuo.

7.1 Distribuições de Probabilidade

Cada variável aleatória tem uma distribuição de probabilidade correspondente, que aplica a teoria da probabilidade para descrever o comportamento dessa variável. No caso das variáveis discretas, ela especifica todos os resultados possíveis da variável aleatória junto com a probabilidade de que cada um ocorra. No caso das contínuas, ela nos permite determinar as probabilidades associadas com intervalos específicos de valores.

Por exemplo, seja X uma variável aleatória discreta que representa a ordem de nascimento de cada criança nascida de uma mulher residente nos Estados Unidos [1]. Se a criança é o primeiro filho da mulher, $X = 1$; se é o segundo, $X = 2$. Para construir uma distribuição de probabilidade para X , listamos cada valor x que a variável aleatória pode assumir, junto com $P(X = x)$ para cada um. Isso foi feito na Tabela 7.1. Os resultados $X = 8$, $X = 9$ e assim por diante para os inteiros contáveis foram agrupados e chamados de “8+”. Note que usamos um X maiúsculo para denotar a variável aleatória e um x minúsculo para representar o resultado de uma criança em particular.

A Tabela 7.1 se parece com as distribuições de freqüências introduzidas no Capítulo 2. Para uma amostra de observações, a distribuição de freqüências exibe cada resultado observado e o número de vezes que aparece no conjunto de dados. A distribuição de freqüências algumas vezes inclui também a freqüência relativa de cada resultado. Para uma variável aleatória

TABELA 7.1

Distribuição de Probabilidade de uma variável aleatória X que represente a ordem de nascimento de crianças nos Estados Unidos.

x	$P(X = x)$
1	0,416
2	0,330
3	0,158
4	0,058
5	0,021
6	0,009
7	0,004
8+	0,004
Total	1,000

discreta, a distribuição de probabilidade lista cada possível resultado e sua probabilidade correspondente. As probabilidades representam a freqüência relativa de ocorrência de cada resultado x em um grande número de ensaios repetidos sob condições essencialmente idênticas; equivalentemente, elas podem ser imaginadas como as freqüências relativas associadas com uma amostra infinitamente grande. Elas nos contam quais valores são mais prováveis de ocorrer do que outros. Desde que todos os possíveis valores da variável aleatória sejam levados em conta, os resultados são exaustivos; então, a soma de suas probabilidades precisa ser 1.

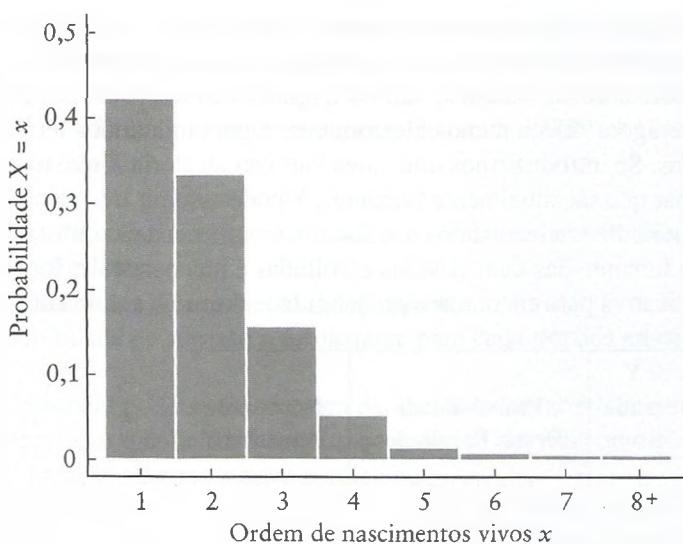
Em muitos casos, podemos exibir uma distribuição de probabilidade por meio de um gráfico ou de uma fórmula matemática. A Figura 7.1, por exemplo, é um histograma da distribuição de probabilidade mostrada na Tabela 7.1. A área de cada barra vertical representa $P(X = x)$, a probabilidade associada com aquele resultado em particular da variável aleatória; a área total do histograma é igual a 1.

A distribuição de probabilidade de X pode ser usada para se fazer afirmações sobre os possíveis resultados da variável aleatória. Suponha que queiramos conhecer a probabilidade de que um recém-nascido aleatoriamente escolhido seja a quarta criança de sua mãe. Ao usarmos a informação na Tabela 7.1, observamos que $P(X = 4) = 0,058$. Qual a probabilidade de que o bebê seja a primeira ou a segunda criança de sua mãe? Ao se aplicar a regra aditiva de probabilidade para eventos mutuamente exclusivos,

$$\begin{aligned} P(X = 1 \text{ ou } X = 2) &= P(X = 1) + P(X = 2) \\ &= 0,416 + 0,330 \\ &= 0,746. \end{aligned}$$

Se uma variável aleatória pode assumir grande número de valores, uma distribuição de probabilidade pode não ser um modo útil de resumir seu comportamento. Tal como a distribuição de freqüências de dados agrupados, no entanto, podemos descrever uma distribuição de probabilidade usando uma medida de tendência central e uma medida de dispersão. O valor médio assumido por uma variável aleatória é conhecido como a *média da população*; a dispersão dos valores relativa a essa média é a *variância da população*, cuja raiz quadrada é o *desvio-padrão da população*.

A distribuição de probabilidade da ordem de nascimento de crianças nos Estados Unidos foi gerada com base na experiência da população do país em 1986. Probabilidades

**FIGURA 7.1**

Distribuição de probabilidade de uma variável aleatória que representa a ordem de nascimento de crianças nos Estados Unidos.

calculadas a partir de uma quantidade finita de dados são chamadas *probabilidades empíricas*. As distribuições de probabilidade de muitas outras variáveis de interesse, no entanto, podem ser determinadas com base em considerações teóricas. As distribuições desse tipo são conhecidas como *distribuições teóricas de probabilidade*.

7.2 A Distribuição Binomial

Considere uma variável aleatória dicotômica Y . Por definição, a variável Y deve assumir um de dois possíveis valores; esses resultados mutuamente exclusivos poderiam representar vida e morte, homem e mulher ou doença e saúde. A título de simplificação, são referidas freqüentemente como “fracasso” e “sucesso”. Uma variável aleatória desse tipo é conhecida como *variável aleatória de Bernoulli*.

Como exemplo, seja Y uma variável aleatória que representa o *status* de fumante; $Y = 1$ se um adulto atualmente é fumante e $Y = 0$ se ele não o é. Os dois resultados de Y são mutuamente exclusivos e exaustivos. Em 1987, 29% dos adultos nos Estados Unidos fumavam cigarros, charutos ou cachimbos [2]; portanto, podemos enumerar as probabilidades associadas com os respectivos resultados de Y como

$$\begin{aligned} P(Y = 1) &= p \\ &= 0,29 \end{aligned}$$

e

$$\begin{aligned} P(Y = 0) &= 1 - p \\ &= 1,00 - 0,29 \\ &= 0,71. \end{aligned}$$

Essas duas equações descrevem completamente a distribuição de probabilidade da variável aleatória dicotômica Y ; ao supormos que os hábitos dos fumantes não tenham mudado desde 1987 (talvez uma hipótese não-razoável), se viajássemos pelos Estados Unidos observan-

do se determinados adultos são ou não fumantes, Y assumiria o valor 1 aproximadamente 29% das vezes e o valor 0 os restantes 71%. Lembre-se de que a proporção de vezes que uma variável aleatória dicotômica assume o valor 1 é igual a sua média da população.

Suponha que agora selecionamos aleatoriamente dois indivíduos da população adulta dos Estados Unidos. Se introduzirmos uma nova variável aleatória X que represente o número de pessoas do par que são atualmente fumantes, X pode assumir três valores possíveis: 0, 1 ou 2. Ambos os indivíduos selecionados não fumam ou um fuma e o outro não ou ambos fumam. O *status* de fumante das duas pessoas escolhidas é independente; logo, podemos aplicar a regra multiplicativa para encontrar a probabilidade de que X assuma um valor particular.

Resultado de Y Primeira Pessoas	Segunda Pessoas	Probabilidade Desses Resultados	Número de Fumantes X
0	0	$(1-p)(1-p)$	0
1	0	$p(1-p)$	1
0	1	$(1-p)p$	1
1	1	pp	2

Com a substituição do valor de p , verificamos que

$$\begin{aligned} P(X = 0) &= (1-p)^2 \\ &= (0,71)^2 \\ &= 0,504, \end{aligned}$$

$$\begin{aligned} P(X = 1) &= p(1-p) + (1-p)p \\ &= 2p(1-p) \\ &= 2(0,29)(0,71) \\ &= 0,412, \end{aligned}$$

e

$$\begin{aligned} P(X = 2) &= p^2 \\ &= (0,29)^2 \\ &= 0,084. \end{aligned}$$

Note-se que existem duas possíveis situações nas quais um adulto fuma e outro não; ou $Y = 1$ para o primeiro indivíduo e $Y = 0$ para o segundo ou $Y = 0$ para o primeiro e $Y = 1$ para o segundo. Os dois resultados são mutuamente exclusivos e aplicamos a regra aditiva de probabilidade para encontrar $P(X = 1)$. Observe também que, por serem considerados todos os resultados possíveis de X , suas probabilidades precisam somar 1; isto é,

$$\begin{aligned} P(X = 0) + P(X = 1) + P(X = 2) &= 0,504 + 0,412 + 0,084 \\ &= 1,000. \end{aligned}$$

A distribuição de probabilidade da variável aleatória discreta X descrita anteriormente é um caso especial da *distribuição binomial*. Em geral, se temos uma seqüência de n ensaios independentes de Bernoulli — ou, equivalentemente, n resultados independentes da variável aleatória de Bernoulli Y — cada uma com probabilidade de “sucesso” p , então o número total de sucessos X é uma variável aleatória binomial. Os números fixados n e p são chamados

de parâmetros da distribuição. Parâmetros são quantidades numéricas que resumem as características de uma distribuição de probabilidade. No exemplo anterior, os parâmetros são $n = 2$, pois dois indivíduos são selecionados e $p = 0,29$, porque a probabilidade de que qualquer adulto escolhido aleatoriamente seja atualmente fumante é 0,29. A distribuição binomial envolve três suposições:

1. Há um número fixo de ensaios n , cada um resulta em um de dois resultados mutuamente exclusivos.
2. Os resultados dos n ensaios são independentes.
3. A probabilidade de sucesso p é constante para cada um dos ensaios.

A distribuição binomial pode ser usada para descrever uma variedade de situações, tais como o número de irmãos que herdarão certo traço genético dos pais ou o número de pacientes que experimenta reação adversa a novo medicamento.

Suponha que continuássemos com o exemplo precedente, selecionando de forma aleatória três adultos a partir da população e não dois. Nesse caso, X seria uma variável aleatória binomial com parâmetros $n = 3$ e $p = 0,29$.

Resultado de Y			Probabilidade Desses Resultados	Número de Fumantes X
Primeira Pessoas	Segunda Pessoas	Terceira Pessoas		
0	0	0	$(1-p)(1-p)(1-p)$	0
1	0	0	$p(1-p)(1-p)$	1
0	1	0	$(1-p)p(1-p)$	1
0	0	1	$(1-p)(1-p)p$	1
1	1	0	$pp(1-p)$	2
1	0	1	$p(1-p)p$	2
0	1	1	$(1-p)pp$	2
1	1	1	ppp	3

Ao substituir o valor de p ,

$$\begin{aligned} P(X = 0) &= (1-p)^3 \\ &= (0,71)^3 \\ &= 0,358, \end{aligned}$$

$$\begin{aligned} P(X = 1) &= p(1-p)^2 + p(1-p)^2 + p(1-p)^2 \\ &= 3(0,29)(0,71)^2 \\ &= 0,439, \end{aligned}$$

$$\begin{aligned} P(X = 2) &= p^2(1-p) + p^2(1-p) + p^2(1-p) \\ &= 3(0,29)^2(0,71) \\ &= 0,179, \end{aligned}$$

$$\begin{aligned} P(X = 3) &= p^3 \\ &= (0,29)^3 \\ &= 0,024. \end{aligned}$$

Essas equações descrevem a distribuição de probabilidade de X . A variável aleatória X pode assumir quatro possíveis valores e

$$\begin{aligned} P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ = 0,358 + 0,439 + 0,179 + 0,024 \\ = 1,000. \end{aligned}$$

Note-se que $P(X = 1)$ e $P(X = 2)$ envolvem a soma de três termos; se temos um total de três indivíduos, existem exatamente três condições nas quais um deles pode ser fumante e três condições nas quais dois podem ser fumantes.

Se prosseguirmos com nosso exemplo e selecionarmos um total de n adultos da população, a probabilidade de que exatamente x deles fumem pode ser escrita como

$$\begin{aligned} P(X = x) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} (0,29)^x (0,71)^{n-x} \end{aligned}$$

onde $n = 1, 2, 3, \dots$ e $x = 0, 1, \dots n$. Essa é a expressão geral para a distribuição de probabilidade de uma variável aleatória binomial X , na qual X é o número de fumantes em uma amostra de tamanho n . Dado um total de n adultos, $n!$ — ou n fatorial — nos permite calcular o número de condições em que os n indivíduos podem ser ordenados; temos n escolhas para a primeira posição, $n - 1$ escolhas para a segunda posição e assim por diante. Assim,

$$n! = n(n-1)(n-2) \cdots (3)(2)(1).$$

Por definição, $0!$ é igual a 1. A expressão

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

é a *combinação* de n objetos escolhidos x por vez; ela representa o número de condições nas quais x objetos podem ser selecionados de um total de n objetos sem considerar a ordem. Por exemplo, se escolhêssemos aleatoriamente três indivíduos da população adulta, haveria

$$\binom{3}{0} = \frac{3!}{0!(3-0)!} = \frac{6}{(1)(6)} = 1$$

condições nas quais poderíamos selecionar zero fumantes; em particular, os três adultos teriam que ser não-fumantes. Haveria

$$\binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{6}{(1)(2)} = 3$$

condições nas quais poderíamos escolher um fumante, desde que o fumante pudesse ser a primeira, a segunda ou a terceira pessoa. Analogamente, haveria

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{6}{(2)(1)} = 3$$

condições nas quais poderíamos selecionar dois fumantes e somente

$$\binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{6}{(6)(1)} = 1$$

condição para os três adultos serem fumantes. Portanto, como verificamos anteriormente,

$$\begin{aligned} P(X=0) &= \binom{3}{0} p^0 (1-p)^{3-0} \\ &= 1(0,29)^0 (0,71)^3 \\ &= 0,358, \end{aligned}$$

$$\begin{aligned} P(X=1) &= \binom{3}{1} p^1 (1-p)^{3-1} \\ &= 3(0,29)(0,71)^2 \\ &= 0,439, \end{aligned}$$

$$\begin{aligned} P(X=2) &= \binom{3}{2} p^2 (1-p)^{3-2} \\ &= 3(0,29)^2 (0,71) \\ &= 0,179, \end{aligned}$$

e

$$\begin{aligned} P(X=3) &= \binom{3}{3} p^3 (1-p)^{3-3} \\ &= 1(0,29)^3 (0,71)^0 \\ &= 0,024. \end{aligned}$$

Em vez de realizar esses cálculos manualmente — supondo que não temos nenhum dos muitos programas de computador para calculá-los — podemos usar a Tabela A.1 no Apêndice A para obtermos as probabilidades binomiais dos valores selecionados de n e p . O número de ensaios n aparece na primeira coluna do lado esquerdo da tabela para $n \leq 20$. O número de sucessos x está na segunda coluna e toma valores inteiros de 0 até n . A probabilidade p aparece na linha de topo. Para valores especificados de n , x e p , a entrada no corpo da tabela representa

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Novamente, suponha que escolhemos aleatoriamente três indivíduos da população adulta e queremos encontrar a probabilidade de que exatamente dois deles sejam fumantes. Primeiro localizamos $n = 3$ no lado esquerdo da tabela e então selecionamos a linha que corresponde a $x = 2$. Arredondando a probabilidade $p = 0,29$ para 0,3, encontramos a coluna correspondente a $p = 0,3$. Isso nos permite aproximar a probabilidade de que exatamente dois dos três adultos são fumantes para 0,189 (esse resultado difere de 0,179, a probabilidade calculada acima, porque foi necessário arredondar o valor de p).

O que resultaria se escolhêssemos três adultos da população e quiséssemos encontrar a probabilidade de que exatamente dois deles não sejam fumantes? Nesse caso, queremos determinar a probabilidade binomial correspondente a $n = 3$, $x = 2$ e $p = 0,71$. Mesmo se arredondássemos p para 0,7, no entanto, a Tabela A.1 não contém qualquer valor de p maior que

0,5. O modo de resolver esse problema é entender que, se dois dos três indivíduos não são fumantes, o outro precisa sê-lo. Portanto, simplesmente usamos a tabela para encontrar $P(X = 1 | n = 3, p = 0,3)$, o que é matematicamente equivalente a $P(X = 2 | n = 3, p = 0,7)$.

Além das probabilidades dos resultados individuais, podemos calcular as medidas-resumo numéricas associadas à distribuição de probabilidade. Por exemplo, o valor médio de uma variável aleatória binomial X — ou o número médio de “sucessos” em amostras repetidas de tamanho n — é obtido pela multiplicação do número de ensaios independentes de Bernoulli pela probabilidade de sucesso de cada ensaio; por isso, o valor médio de X é igual a np . A variância de X é $np(1 - p)$. Essas expressões foram obtidas com o uso de um método análogo àquele para se encontrar a média e a variância de dados agrupados [3]. Com a aplicação das fórmulas, se selecionássemos amostras repetidas de tamanho $n = 10$ da população adulta, o número médio de fumantes por amostra seria

$$\begin{aligned} np &= 10(0,29) \\ &= 2,9, \end{aligned}$$

e o desvio-padrão seria

$$\begin{aligned} \sqrt{np(1 - p)} &= \sqrt{10(0,29)(0,71)} \\ &= \sqrt{2,059} \\ &= 1,4. \end{aligned}$$

A expressão para a variância de uma variável aleatória binomial parece ser bastante razoável. A quantidade $np(1 - p)$ é maior quando p é igual a 0,5 e diminui conforme p se aproxima de 0 ou de 1. Quando p é muito grande ou muito pequeno, quase todos os resultados tomam o mesmo valor — por exemplo, quase todo mundo fuma ou quase ninguém fuma — e a variabilidade entre os resultados é pequena. Paralelamente, se metade da população assume o valor 0 e a outra metade o valor 1, é mais difícil predizer qualquer resultado particular; nesse caso, a variabilidade é relativamente grande.

A Figura 7.2 é um gráfico da distribuição de probabilidade de X — o número de fumantes — para o qual $n = 10$ e $p = 0,29$. Por serem considerados todos os possíveis resultados

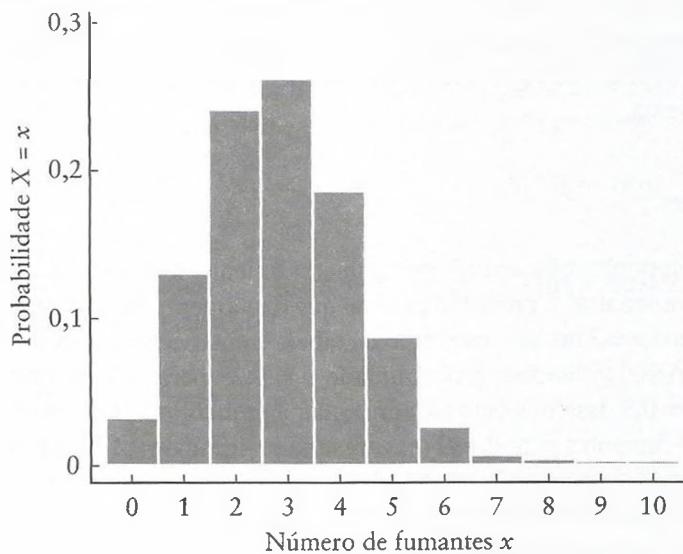
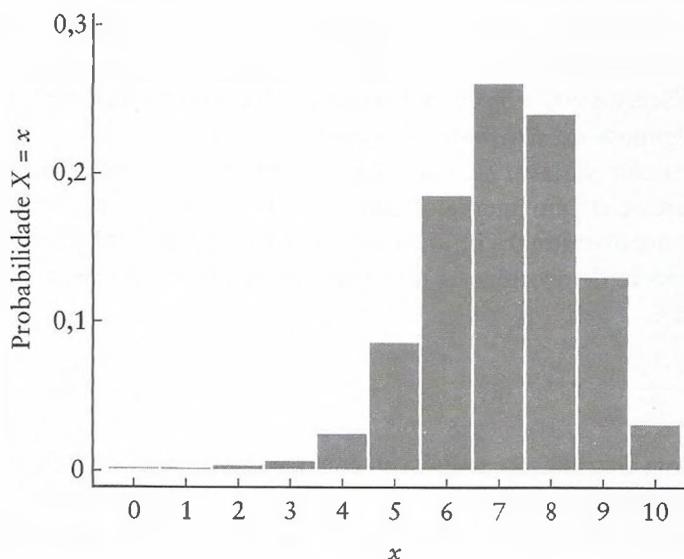


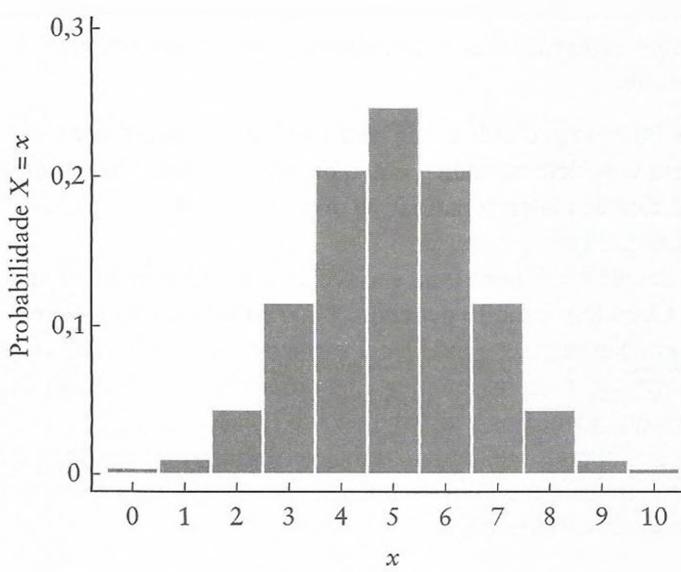
FIGURA 7.2

Distribuição de probabilidade de uma variável aleatória binomial na qual $n = 10$ e $p = 0,29$.

de X , as áreas representadas pela barras verticais somam 1. A Figura 7.3 é a distribuição de probabilidade de outra variável aleatória binomial para a qual $n = 10$ e $p = 0,71$. Note-se que a distribuição é assimétrica à direita quando $p < 0,5$ e assimétrica à esquerda quando $p > 0,5$. Se $p = 0,5$, como é o caso na Figura 7.4, a distribuição de probabilidade é simétrica.

**FIGURA 7.3**

Distribuição de probabilidade de uma variável aleatória binomial na qual $n = 10$ e $p = 0,71$.

**FIGURA 7.4**

Distribuição de probabilidade de uma variável aleatória binomial na qual $n = 10$ e $p = 0,50$.

7.3 A Distribuição de Poisson

Suponha que X seja uma variável aleatória que representa o número de indivíduos envolvidos em um acidente de veículo a motor a cada ano. Nos Estados Unidos, a probabilidade de que um indivíduo em particular esteja envolvido é 0,00024 [4]. Tecnicamente, essa é uma

situação binomial na qual existem dois resultados distintos — acidente ou não acidente. No entanto, n é muito grande; estamos interessados na população inteira dos Estados Unidos. Quando n se torna grande, a combinação de n objetos tomados x de cada vez, $n!/x!(n-x)!$, é muito cansativa de se avaliar. Como resultado, a distribuição binomial pode ser impraticável de se usar como base para os cálculos. No entanto, em situações como essa — quando n é muito grande e p é muito pequeno — a distribuição binomial é bem aproximada por outra distribuição teórica de probabilidade, chamada *distribuição de Poisson*, que é usada para modelar eventos discretos ocorridos com baixa freqüência no tempo ou no espaço; por isso, é algumas vezes chamada de *distribuição de eventos raros*.

Consideremos uma variável aleatória X que representa o número de ocorrências de algum evento de interesse em um intervalo. Como X é uma contagem, pode teoricamente assumir qualquer valor inteiro entre 0 e infinito. Seja λ (a letra grega lambda) uma constante que denota o número médio de ocorrências do evento em um intervalo. Se a probabilidade que X assuma o valor x é

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!},$$

diz-se que X tem uma distribuição de Poisson com parâmetro λ . O símbolo e representa uma constante aproximada por 2,71828; de fato, e é a base dos logaritmos naturais. Tal como a binomial, a distribuição de Poisson envolve um conjunto de suposições básicas:

1. A probabilidade de que um único evento ocorra em um intervalo é proporcional ao comprimento desse intervalo.
2. Em um intervalo único, um número infinito de ocorrências do evento é teoricamente possível. Não estamos restritos a um número fixado de ensaios.
3. Os eventos ocorrem independentemente, no mesmo intervalo ou entre intervalos consecutivos.

A distribuição de Poisson pode ser usada para modelar o número necessário de ambulâncias em uma cidade em uma determinada noite, o número de partículas emitidas a partir de uma quantidade específica de material radioativo ou o número de colônias de bactérias que crescem em uma placa de Petri.

Lembre-se de que a média de uma variável aleatória binomial é igual a np e sua variância é $np(1-p)$. Quando p é muito pequeno, $1-p$ é próximo de 1 e $np(1-p)$ é aproximadamente igual a np . Nesse caso, a média e a variância da distribuição são idênticas e podem ser representadas pelo parâmetro simples λ . A propriedade cuja média é igual à variância é uma característica que identifica a distribuição de Poisson.

Suponha que estejamos interessados em determinar o número de pessoas em uma população de 10.000 que estará envolvido em um acidente de veículo a motor a cada ano. O número médio de pessoas envolvidas seria

$$\begin{aligned}\lambda &= np \\ &= (10.000)(0,00024) \\ &= 2,4;\end{aligned}$$

que também é variância. A probabilidade de que ninguém dessa população esteja envolvido em um acidente em um determinado ano é

$$\begin{aligned}P(X = 0) &= \frac{e^{-2,4}(2,4)^0}{0!} \\ &= 0,091.\end{aligned}$$

A probabilidade de que exatamente uma pessoa esteja envolvida é

$$\begin{aligned} P(X = 1) &= \frac{e^{-2,4}(2,4)^1}{1!} \\ &= 0,218. \end{aligned}$$

Analogamente,

$$\begin{aligned} P(X = 2) &= \frac{e^{-2,4}(2,4)^2}{2!} \\ &= 0,261, \end{aligned}$$

$$\begin{aligned} P(X = 3) &= \frac{e^{-2,4}(2,4)^3}{3!} \\ &= 0,209, \end{aligned}$$

$$\begin{aligned} P(X = 4) &= \frac{e^{-2,4}(2,4)^4}{4!} \\ &= 0,125, \end{aligned}$$

$$\begin{aligned} P(X = 5) &= \frac{e^{-2,4}(2,4)^5}{5!} \\ &= 0,060, \end{aligned}$$

$$\begin{aligned} P(X = 6) &= \frac{e^{-2,4}(2,4)^6}{6!} \\ &= 0,024. \end{aligned}$$

Como os resultados de X são mutuamente exclusivos e exaustivos,

$$\begin{aligned} P(X \geq 7) &= 1 - P(X < 7) \\ &= 1 - (0,091 + 0,218 + 0,261 + 0,209 \\ &\quad + 0,125 + 0,060 + 0,024) \\ &= 0,012. \end{aligned}$$

Em vez de fazermos cálculos manualmente — ou com um programa de computador — podemos usar a Tabela A.2 no Apêndice A para obtermos as probabilidades de Poisson para valores selecionados de λ . O número de sucessos x aparece na primeira coluna do lado esquerdo da tabela; λ está na linha de topo. Para valores específicos de x e λ , a entrada na tabela representa

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Em uma população de 10.000 pessoas, qual a probabilidade de que exatamente três delas estejam envolvidas em um acidente de veículo a motor em determinado ano? Começamos por localizar $x = 3$ na primeira coluna da Tabela A.2. Ao arredondarmos 2,4 para 2,5, encontramos a coluna correspondente a $\lambda = 2,5$. A tabela nos mostra que podemos aproximar a probabilidade de que exatamente três indivíduos estejam envolvidos em um acidente por 0,214 (novamente, esse resultado difere de 0,209, a probabilidade calculada acima, porque foi necessário arredondar o valor do parâmetro λ para o uso da tabela).

A Figura 7.5 é um gráfico da distribuição de probabilidade de X , o número de indivíduos da população envolvida em um acidente de veículo a motor a cada ano. As áreas repre-

sentadas pelas barras verticais somam 1. Como mostrado na Figura 7.6, a distribuição de Poisson é bastante assimétrica para pequenos valores de λ ; conforme λ aumenta, a distribuição se torna mais simétrica.

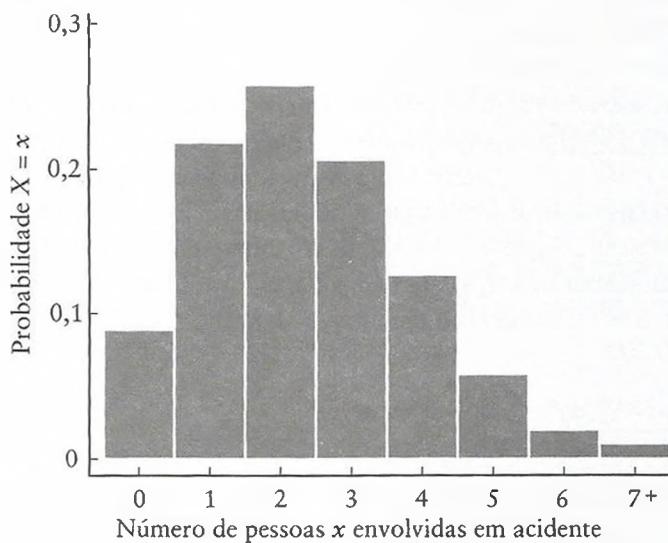


FIGURA 7.5

Distribuição de probabilidade de uma variável aleatória de Poisson para a qual $\lambda = 2,4$.

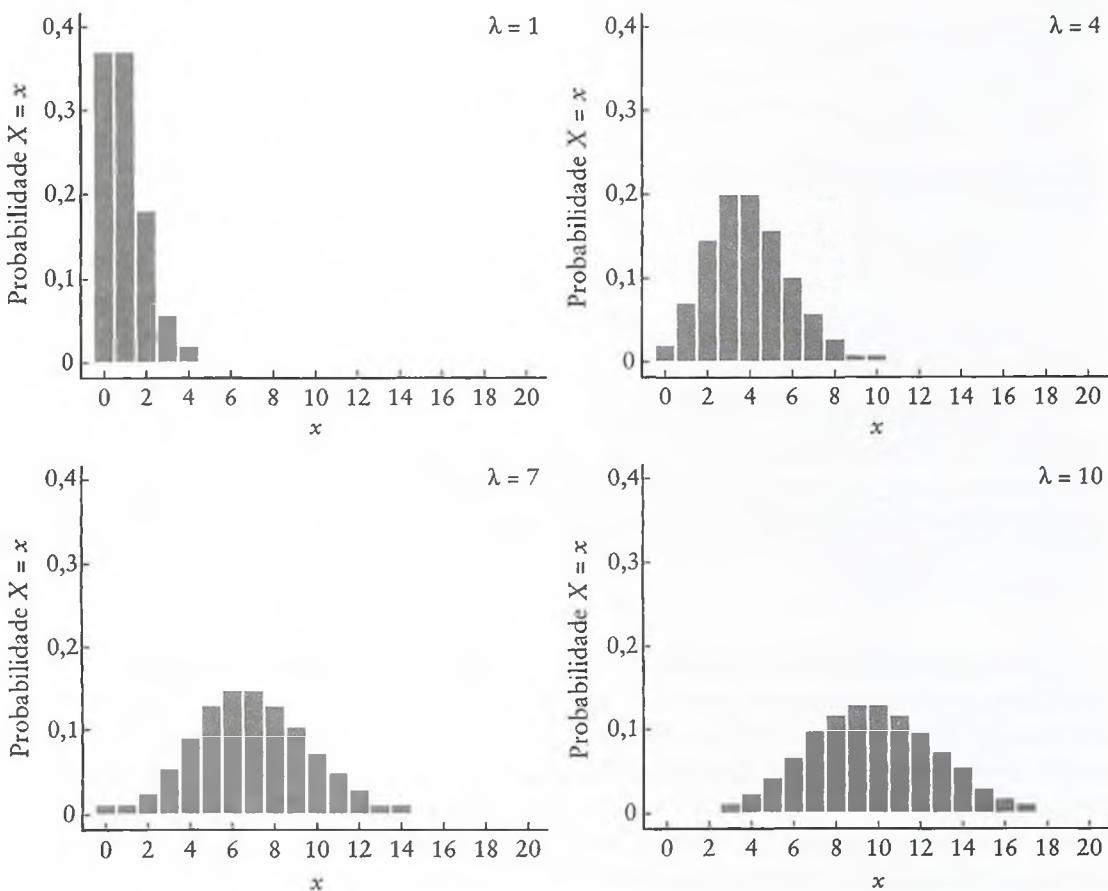


FIGURA 7.6

Distribuições de probabilidade de variáveis aleatórias de Poisson para vários valores de λ .

7.4 A Distribuição Normal

Quando uma variável aleatória X segue uma distribuição binomial ou uma distribuição de Poisson, fica restrita a assumir somente valores inteiros. Sob diferentes circunstâncias, no entanto, os resultados de uma variável aleatória podem não estar limitados a inteiros ou a contagens. Suponha que X represente altura. Raramente um indivíduo tem exatamente 67 ou 68 polegadas de altura; teoricamente, X pode assumir um número infinito de valores intermediários, tais como 67,04 polegadas ou 67,8352. De fato, entre quaisquer dois resultados possíveis de X podemos sempre encontrar um terceiro valor. Embora discutamos filosoficamente que podemos medir somente resultados discretos devido às limitações de nossos instrumentos de medida — talvez possamos medir a altura somente no décimo de polegada mais próximo — tratar essa variável como se fosse contínua permite-nos tirar vantagem de poderosos resultados matemáticos.

Como vimos, a distribuição de probabilidade de uma variável aleatória discreta está representada por uma equação de $P(X = x)$, a probabilidade de que a variável aleatória X assuma o valor específico x . Para uma variável aleatória binomial com parâmetros n e p , por exemplo,

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Essas probabilidades podem ser plotadas *versus* x , como na Figura 7.4. Suponha que o número de resultados possíveis de X vai se tornar muito grande e que a largura dos intervalos correspondentes se torne muito pequena. Na Figura 7.7, por exemplo, $n = 30$ e $p = 0,50$. Geralmente, se o número de valores possíveis de X se aproxima do infinito enquanto a largura dos intervalos se aproxima de zero, o gráfico paulatinamente se parecerá com uma curva suave, que é usada para representar a distribuição de probabilidade de uma variável aleatória contínua. É chamada de *densidade de probabilidade*.

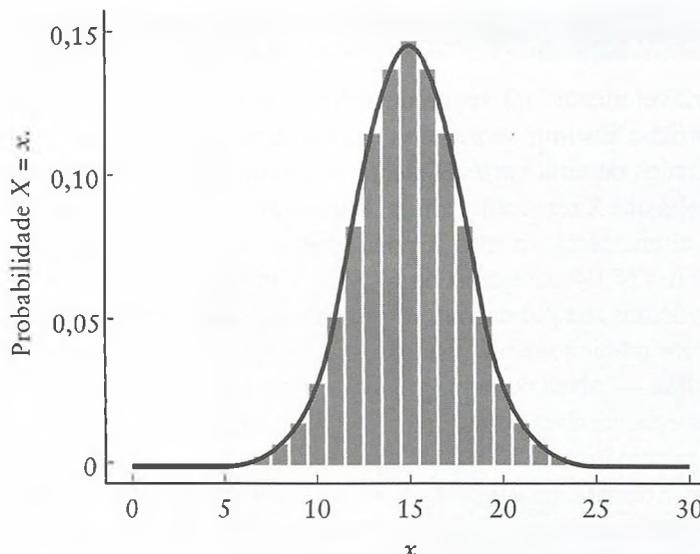
Para qualquer gráfico que ilustra uma distribuição de probabilidade discreta, a área representada pelas barras verticais soma 1. Para a densidade de probabilidade, a área total embaixo da curva também precisa ser 1. Por uma variável aleatória contínua poder tomar um número infinito de valores, a probabilidade associada com qualquer um deles em particular é igual a zero. No entanto, a probabilidade de que X assuma algum valor no intervalo limitado pelos resultados x_1 e x_2 é igual à área embaixo da curva que se encontra entre esses dois valores.

A distribuição contínua mais comum é a *distribuição normal*, também conhecida como *distribuição Gaussiana* ou *curva em forma de sino*. Sua forma é aquela da distribuição binomial para a qual p é constante, mas n se aproxima do infinito ou de uma distribuição de Poisson para a qual λ se aproxima do infinito. Sua densidade de probabilidade é dada pela equação

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},$$

onde $-\infty < x < \infty$. O símbolo π (pi) representa uma constante aproximada por 3,14159. A curva normal é unimodal e simétrica ao redor de sua média μ (mu); nesse caso em especial, a média, a mediana e a moda da distribuição são idênticas. O desvio-padrão, representado por σ (sigma), especifica a quantidade de dispersão ao redor da média. Juntos, os parâmetros μ e σ definem completamente uma curva normal.

O valor de uma distribuição normal se tornará mais aparente quando começarmos a trabalhar com a distribuição amostral da média. De momento, no entanto, é importante no-

**FIGURA 7.7**

Distribuição de probabilidade de uma variável aleatória binomial para a qual $n = 30$ e $p = 0,50$.

tar que muitas variáveis aleatórias de interesse — inclusive a pressão sanguínea, nível sérico de colesterol, altura e peso — têm distribuição aproximadamente normal. A curva normal pode, assim, ser usada para estimar probabilidades associadas a essas variáveis. Por exemplo, em uma população na qual o nível sérico de colesterol é normalmente distribuído com média μ e desvio-padrão σ , poderíamos encontrar a probabilidade de que um indivíduo aleatoriamente escolhido tenha um nível sérico de colesterol maior que 250 mg/100 ml. Talvez esse conhecimento nos auxilie a planejar futuros serviços de atendimento a cardíacos. Como a área total embaixo da curva normal é igual a 1, podemos estimar a probabilidade em questão, ao determinarmos a proporção da área sob a curva à direita do ponto $x = 250$ ou $P(X > 250)$, o que pode ser feito com um programa de computador ou uma tabela de áreas calculadas para a curva normal.

Desde que uma distribuição normal possa ter um número infinito de valores possíveis para sua média e desvio-padrão, é impossível tabular a área associada a cada uma das curvas normais. Ao contrário, somente uma curva é tabulada — o caso em especial para o qual $\mu = 0$ e $\sigma = 1$. Essa curva é conhecida como *distribuição normal padrão*. A Figura 7.8 ilustra a curva normal padrão e a Tabela A.3 do Apêndice A exibe as áreas na extremidade superior da distribuição. Os resultados da variável aleatória Z estão representados por z ; o número inteiro e os décimos das posições decimais de z estão listados na coluna à esquerda da tabela e os centésimos das posições decimais são mostrados na linha de topo. Para um valor particular de z , a entrada no corpo da tabela especifica a área embaixo da curva à direita de z , ou $P(Z > z)$. Alguns valores de amostra de z e suas áreas correspondentes são:

z	Área na extremidade direita
0,00	0,500
1,65	0,049
1,96	0,025
2,58	0,005
3,00	0,001

Assim, por exemplo, $P(Z > 2,58) = 0,005$. Como a distribuição normal padrão é simétrica ao redor de $z = 0$, a área sob a curva à direita de z será igual à área à esquerda de $-z$.

$-z$	Área na Extremidade Esquerda
0,00	0,500
-1,65	0,049
-1,96	0,025
-2,58	0,005
-3,00	0,001

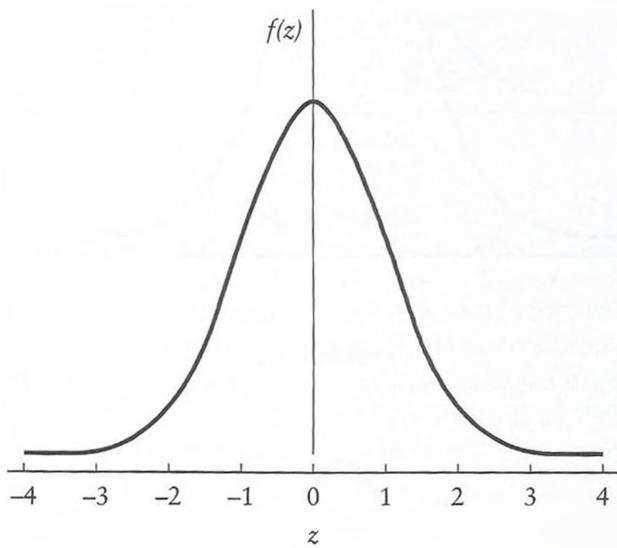


FIGURA 7.8

A curva normal padrão para a qual $\mu = 0$ e $\sigma = 1$.

Suponha que desejamos conhecer a área sob a curva normal padrão que se encontra entre $z = -1,00$ e $z = 1,00$; $\mu = 0$ e $\sigma = 1$, essa é a área contida no intervalo $\mu \pm 1\sigma$, ilustrada na Figura 7.9. Equivalentemente, ela é $P(-1 \leq Z \leq 1)$. Ao observarmos a Tabela A.3, vemos que a área à direita de $z = 1,00$ é $P(Z > 1) = 0,159$. Então, a área à esquerda de $z = -1,00$ precisa também ser 0,159. Os eventos em que $Z > 1$ e $Z < -1$ são mutuamente exclusivos; consequentemente, ao se aplicar a regra aditiva de probabilidade, a soma da área à direita de 1 e à esquerda de -1 é

$$\begin{aligned} P(Z > 1) + P(Z < -1) &= 0,159 + 0,159 \\ &= 0,318. \end{aligned}$$

Desde que a área total sob a curva seja igual a 1, a área entre -1 e 1 precisa ser

$$\begin{aligned} P(-1 \leq Z \leq 1) &= 1 - [P(Z > 1) + P(Z < -1)] \\ &= 1 - 0,318 \\ &= 0,682. \end{aligned}$$

Portanto, para a distribuição normal padrão, aproximadamente 68,2% da área abaixo da curva se encontra dentro de ± 1 desvio-padrão da média.

Poderíamos também calcular a área sob a curva normal padrão contida no intervalo $\mu \pm 2\sigma$ ou $P(-2 \leq Z \leq 2)$. Essa área está ilustrada na Figura 7.10. A Tabela A.3 indica que

a área à direita de $z = 2,00$ é 0,023; a área à esquerda de $z = -2,00$ é 0,023 também. Assim, a área entre $-2,00$ e $2,00$ precisa ser

$$\begin{aligned} P(-2 \leq Z \leq 2) &= 1 - [P(Z > 2) + P(Z < -2)] \\ &= 1,000 - [0,023 + 0,023] \\ &= 0,954. \end{aligned}$$

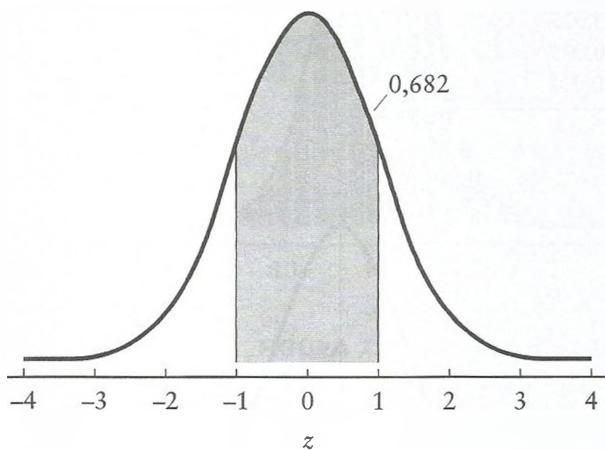


FIGURA 7.9

A curva normal padrão, área entre $z = -1,00$ e $z = 1,00$.

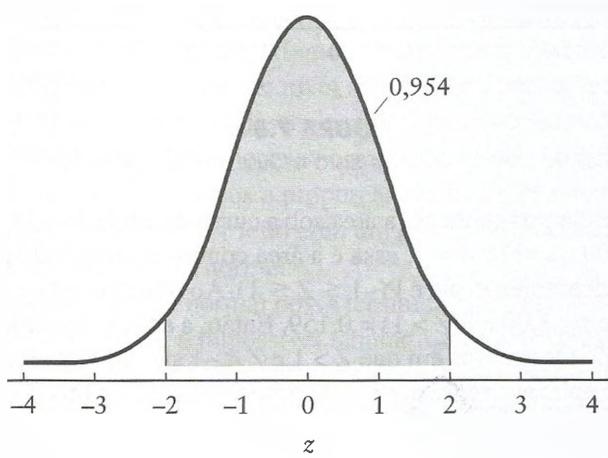
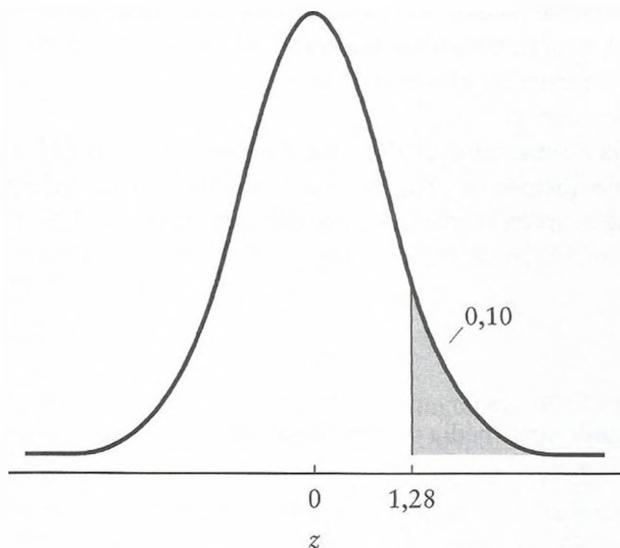


FIGURA 7.10

A curva normal padrão, área entre $z = -2,00$ e $z = 2,00$.

Aproximadamente 95,4% da área sob a curva normal padrão se encontra dentro de ± 2 desvios-padrão da média. Os cálculos anteriores formam a base da regra empírica descrita na Seção 3.4, a qual estabelece que, se uma distribuição de valores é simétrica e unimodal, aproximadamente 67% das observações se encontram dentro de um desvio-padrão da média e cerca de 95% em dois desvios-padrão.

A Tabela A.3 pode também ser usada de outra maneira. Por exemplo, poderíamos desejar encontrar o valor z que limite os 10% superiores da distribuição normal padrão ou o valor de z para o qual $P(Z > z) = 0,10$. Localizando 0,100 no corpo da tabela, observamos que o valor correspondente de z é 1,28. Portanto, 10% da área sob a curva normal padrão se encontra à direita de $z = 1,28$; essa área é mostrada na Figura 7.11. Analogamente, outros 10% da área se encontram à esquerda de $z = -1,28$.

**FIGURA 7.11**

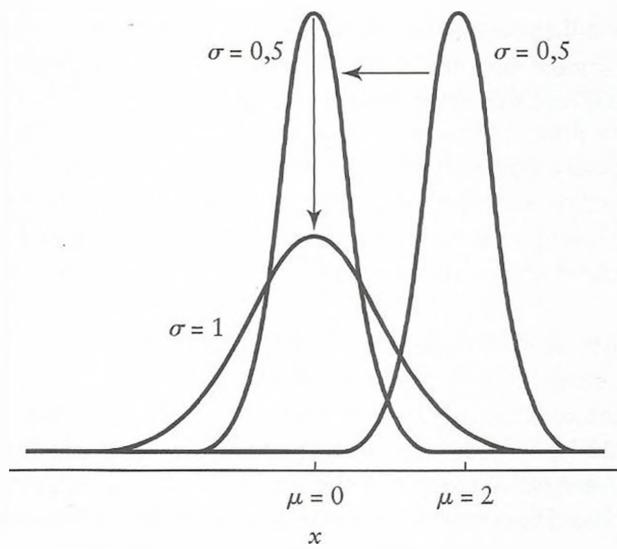
A curva normal padrão, área à direita de $z = 1,28$.

Suponha agora que X seja uma variável aleatória normal com média 2 e desvio-padrão 0,5. Subtraindo-se 2 de X , teríamos uma variável aleatória normal que tem média 0. Como mostrado na Figura 7.12, a distribuição inteira estaria deslocada duas unidades à esquerda. Dividindo-se $(X - 2)$ por 0,5 altera-se a dispersão da distribuição, de modo que temos uma variável aleatória normal com desvio-padrão 1. Então, se X é uma variável aleatória normal com média 2 e desvio-padrão 0,5,

$$Z = \frac{X - 2}{0,5}$$

é uma variável aleatória normal padrão. Geralmente, para qualquer variável aleatória normal arbitrária com média μ e desvio-padrão σ ,

$$Z = \frac{X - \mu}{\sigma}$$

**FIGURA 7.12**

Transformação de uma curva normal com média 2 e desvio-padrão 0,5 em curva normal padrão.

tem uma distribuição normal padrão. Ao transformarmos X em Z , podemos usar uma tabela de áreas calculada para a curva normal padrão a fim de estimar as probabilidades associadas com X . Um resultado da variável aleatória Z , indicada por z , é conhecido como um *desvio normal padrão* ou um *escore z*.

Por exemplo: seja X uma variável aleatória que representa a pressão sanguínea sistólica. Para a população de homens de 18 a 74 anos, nos Estados Unidos, a pressão sanguínea sistólica tem distribuição aproximadamente normal com média de 129 milímetros de mercúrio (mm Hg) e desvio-padrão de 19,8 mm Hg [5]. Essa distribuição é mostrada na Figura 7.13. Note-se que

$$Z = \frac{X - 129}{19,8}$$

é normalmente distribuída com média 0 e desvio-padrão 1.

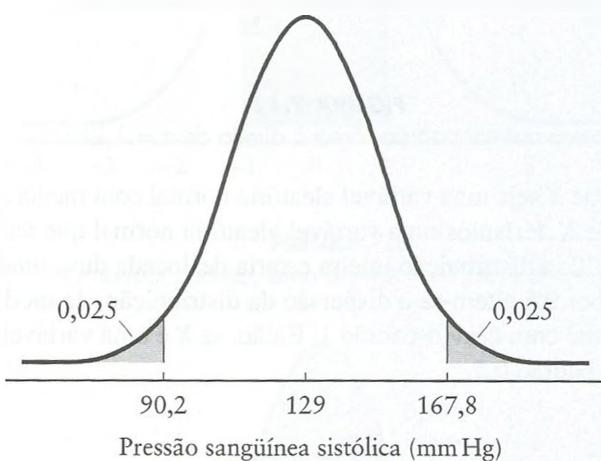


FIGURA 7.13

Distribuição da pressão sanguínea sistólica para homens de 18 a 74 anos, Estados Unidos, 1976–1980.

Suponha que queremos encontrar o valor de x que limite os 2,5% superiores da curva de pressão sanguínea sistólica ou, equivalentemente, o valor de x para o qual $P(X > x) = 0,025$. Usando a Tabela A.3, vemos que a área à direita de $z = 1,96$ é 0,025. Para se obter o valor de x que corresponde a esse valor de z , resolvemos a equação

$$\begin{aligned} z &= 1,96 \\ &= \frac{x - 129}{19,8} \end{aligned}$$

ou

$$\begin{aligned} x &= 129 + (1,96)(19,8) \\ &= 167,8. \end{aligned}$$

Logo, aproximadamente 2,5% dos homens dessa população — uma minúscula minoria — têm pressão sanguínea sistólica maior que 167,8 mm Hg, enquanto 97,5% têm pressão sanguínea menor que 167,8 mm Hg. Em outras palavras, se selecionarmos aleatoriamente um indivíduo dessa população, a probabilidade de que sua pressão sanguínea sistólica seja maior do que 167,8 mm Hg é 0,025.

Porque a curva normal padrão é simétrica ao redor de $z = 0$, sabemos que a área à esquerda de $z = -1,96$ é também 0,025. Resolvendo a equação

$$\begin{aligned} z &= -1,96 \\ &= \frac{x - 129}{19,8} \end{aligned}$$

ou

$$\begin{aligned} x &= 129 + (-1,96)(19,8) \\ &= 90,2, \end{aligned}$$

verificamos que 2,5% dos homens têm pressão sangüínea sistólica menor do que 90,2 mm Hg. Equivalentemente, a probabilidade de que um homem aleatoriamente selecionado tenha pressão sangüínea sistólica menor que 90,2 mm Hg é 0,025. Desde que 2,5% dos homens na população tenham pressões sangüíneas sistólicas maiores que 167,8 mm Hg e 2,5% tenham valores menores do que 90,2 mm Hg, os restantes 95% dos homens precisam ter pressão sangüínea sistólica entre 90,2 e 167,8 mm Hg.

Poderíamos determinar a proporção de homens na população que teria pressões sanguíneas sistólicas maiores do que 150 mm Hg. Nesse caso, damos o resultado da variável aleatória X e precisamos resolver para o desvio normal z :

$$\begin{aligned} z &= \frac{150 - 129}{19,8} \\ &= 1,06. \end{aligned}$$

A área à direita de $z = 1,06$ é 0,145. Então, aproximadamente 14,5% dos homens dessa população têm pressões sangüíneas sistólicas maiores do que 150 mm Hg.

Considere agora a situação mais complicada, na qual há duas variáveis aleatórias normalmente distribuídas. Em um estudo nacional australiano de prevalência de fator de risco, duas das populações investigadas são homens cujas pressões sangüíneas estão em um intervalo normal aceitável e não tomam qualquer medicação corretiva, e homens que têm pressão sangüínea alta, mas que se submetem a terapia com medicamentos anti-hipertensivos [6].

Para a população de homens que não tomou medicação corretiva, a pressão sangüínea diastólica tem distribuição normal aproximada com média $\mu_1 = 80,7$ mm Hg e desvio-padrão $\sigma_1 = 9,2$ mm Hg. Para os homens que usam medicamentos anti-hipertensivos, a pressão sangüínea diastólica tem distribuição normal aproximada distribuída com média $\mu_2 = 94,9$ mm Hg e desvio-padrão $\sigma_2 = 11,5$ mmHg. Essas duas distribuições estão na Figura 7.14. Nosso objetivo é determinar se um homem tem pressão sangüínea normal ou se toma medicação anti-hipertensiva somente com base na leitura de sua pressão sangüínea diastólica. Esse exercício aparentemente sem importância é valioso, pois nos dá fundamentos para os testes de hipóteses.

O primeiro fato a notar é que, devido à grande quantidade de sobreposição entre as duas curvas normais, é difícil fazer distinção entre elas. Não obstante, vamos prosseguir: se nosso objetivo é identificar 90% dos indivíduos que atualmente tomam medicação, que valor de pressão sangüínea diastólica deve ser designada como o ponto de corte mais baixo? Equivalentemente, precisamos encontrar o valor da pressão sangüínea diastólica que limita uma área de 0,10 na extremidade mais baixa da curva normal padrão. Então, resolvendo para x ,

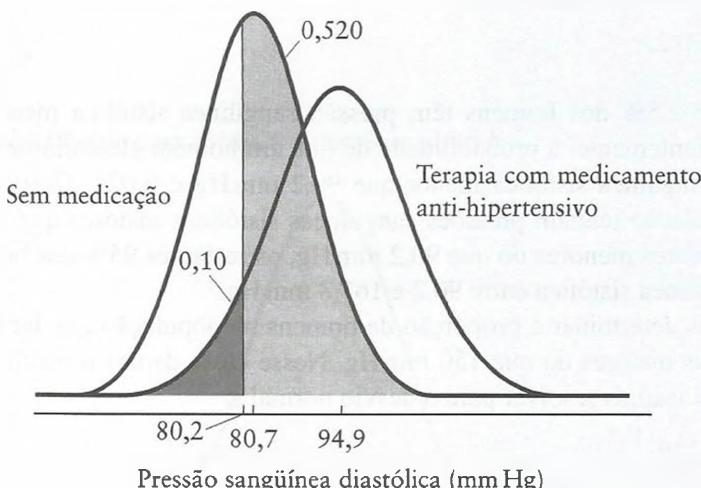
$$z = -1,28$$

$$= \frac{x - 94,9}{11,5}$$

e

$$x = 94,9 + (-1,28)(11,5)$$

$$= 80,2.$$

**FIGURA 7.14**

Distribuição da pressão sanguínea diastólica para duas populações de homens australianos, 1980.

Aproximadamente 90% dos homens que tomam medicamentos anti-hipertensivos têm pressões sanguíneas diastólicas maiores que 80,2 mm Hg. Se usamos esse valor como nosso ponto de corte, os outros 10% dos homens — aqueles com leituras abaixo de 80,2 mm Hg — representam falsos negativos; são indivíduos que atualmente usam medicação e não estão identificados como tais.

Que proporção de homens com pressões sanguíneas normais serão incorretamente rotulados como usuários de medicamentos anti-hipertensivos? Esses são os homens da população livre de medicamentos que têm leituras de pressão sanguínea diastólica maiores que 80,2 mm Hg. Resolvendo para z (note-se que usamos a média e o desvio-padrão da população de homens que não tomam medicação corretiva),

$$z = \frac{80,2 - 80,7}{9,2}$$

$$= -0,05.$$

Uma área de 0,480 encontra-se à esquerda de $-0,05$; portanto, a área à direita de $z = -0,05$ precisa ser

$$1,000 - 0,480 = 0,520.$$

Aproximadamente 52% dos homens com pressões sanguíneas normais estariam incorretamente rotulados como tendo usado medicação. Esses erros são resultados falsos positivos.

Para reduzir a grande proporção de erros falsos positivos, o ponto de corte para identificarmos indivíduos que atualmente usam medicamentos anti-hipertensivos pode ser aumentado. Se o corte fosse 90 mm Hg, por exemplo, então

$$z = \frac{90 - 80,7}{9,2} \\ = 1,01,$$

e somente 15,6% dos homens com pressões sanguíneas normais estariam erroneamente classificados como usuários de medicação.

Entretanto, quando o ponto de corte é aumentado, a proporção dos homens corretamente rotulados como usuários de medicação anti-hipertensiva diminui; note que

$$z = \frac{90 - 94,9}{11,5} \\ = -0,43.$$

A área à esquerda de $z = -0,43$ é 0,334, e

$$1,000 - 0,334 = 0,666;$$

logo, somente 66,6% dos homens que usam medicamentos anti-hipertensivos seriam identificados. Os restantes 33,4% seriam falsos negativos.

Sempre existem concessões quando tentamos manipular as proporções de resultados falsos negativos e de falsos positivos. Esse é o mesmo fenômeno observado quando investigamos a sensibilidade e a especificidade de um teste de diagnóstico. Geralmente, uma proporção menor de erros falsos positivos pode ser obtida com o simples aumento da probabilidade de um resultado falso negativo, assim como a proporção de falsos negativos pode ser reduzida somente com a elevação da probabilidade de falsos positivos. A relação entre esses dois tipos de erros em uma aplicação específica é determinada pela quantidade de sobreposição nas duas populações normais de interesse.

7.5 Aplicações Adicionais

Suponha que queiramos investigar a probabilidade de que um paciente picado com agulha infectada com hepatite B realmente desenvolva a doença. Seja Y uma variável aleatória de Bernoulli que representa o *status* de doença de um paciente exposto a uma agulha infectada; Y assume o valor 1 se o indivíduo desenvolve a hepatite e 0 se não desenvolve. Esses dois resultados são mutuamente exclusivos e exaustivos. Se 30% dos pacientes que estão expostos à hepatite B são infectados [7], então

$$P(Y = 1) = p \\ = 0,30,$$

e

$$P(Y = 0) = 1 - p \\ = 1 - 0,30 \\ = 0,70.$$

Se temos n observações independentes de uma variável aleatória dicotômica tal que cada observação tenha probabilidade constante de “sucesso” p , o número total de “sucessos” X segue uma distribuição binomial. A variável aleatória X pode assumir qualquer valor inteiro entre 0 e n ; a probabilidade de que X assuma um valor x particular pode ser expressa como

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Suponha que selecionamos cinco indivíduos da população de pacientes picados com agulha infectada com hepatite B. O número de pacientes que desenvolve a doença nessa amostra é uma variável aleatória binomial com parâmetro $n = 5$ e $p = 0,30$. Sua distribuição de probabilidade pode ser representada do seguinte modo:

$$\begin{aligned} P(X = 0) &= \binom{5}{0}(0,30)^0(0,70)^{5-0} \\ &= (1)(1)(0,70)^5 \\ &= 0,168, \end{aligned}$$

$$\begin{aligned} P(X = 1) &= \binom{5}{1}(0,30)^1(0,70)^{5-1} \\ &= (5)(0,30)(0,70)^4 \\ &= 0,360, \end{aligned}$$

$$\begin{aligned} P(X = 2) &= \binom{5}{2}(0,30)^2(0,70)^{5-2} \\ &= (10)(0,30)^2(0,70)^3 \\ &= 0,309, \end{aligned}$$

$$\begin{aligned} P(X = 3) &= \binom{5}{3}(0,30)^3(0,70)^{5-3} \\ &= (10)(0,30)^3(0,70)^2 \\ &= 0,132, \end{aligned}$$

$$\begin{aligned} P(X = 4) &= \binom{5}{4}(0,30)^4(0,70)^{5-4} \\ &= (5)(0,30)^4(0,70) \\ &= 0,028, \end{aligned}$$

e

$$\begin{aligned} P(X = 5) &= \binom{5}{5}(0,30)^5(0,70)^{5-5} \\ &= (1)(0,30)^5(1) \\ &= 0,002. \end{aligned}$$

Em vez de calcularmos essas probabilidades manualmente, poderíamos consultar a Tabela A.1 do Apêndice A. Alternativamente, muitos pacotes estatísticos geram probabilidades associadas com uma variável aleatória binomial; a Tabela 7.2 mostra o resultado relevante do Minitab.

TABELA 7.2

Saída do Minitab que exibe a distribuição de probabilidade de uma variável aleatória binomial com parâmetros $n = 5$ e $p = 0,30$.

BINOMIAL WITH N=5 P=0.30	
K	P (X=K)
0	0.1681
1	0.3601
2	0.3087
3	0.1323
4	0.0284
5	0.0024

A probabilidade de que pelo menos três indivíduos entre os cinco desenvolvam a hepatite B é

$$\begin{aligned} P(X \geq 3) &= P(X = 3) + P(X = 4) + P(X = 5) \\ &= 0,132 + 0,028 + 0,003 \\ &= 0,163; \end{aligned}$$

a probabilidade de que no máximo um paciente desenvolva a doença é

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= 0,168 + 0,360 \\ &= 0,528. \end{aligned}$$

Além disso, o número médio de pessoas que desenvolveria a doença em amostras repetidas de tamanho 5 é $np = 5(0,3) = 1,5$ e o desvio-padrão é

$$\sqrt{np(1-p)} = \sqrt{5(0,3)(0,7)} = \sqrt{1,05} = 1,03.$$

Se X representa o número de ocorrências de algum evento em um intervalo específico de tempo ou espaço tal que tanto o número médio de ocorrências como a variância da população são iguais a λ , X tem uma distribuição de Poisson com parâmetro λ . A variável aleatória X pode assumir qualquer valor inteiro entre 0 e ∞ ; a probabilidade de que X assuma um valor particular x é

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Suponha que estejamos preocupados com a possível difusão de difteria e queiramos saber quantos casos podemos esperar em um determinado ano. Seja X o número de casos registrados de difteria nos Estados Unidos em um ano entre 1980 e 1989. A variável aleatória X tem uma distribuição de Poisson com parâmetro $\lambda = 2,5$ [8]; a distribuição de probabilidade de X pode ser expressa como

$$P(X = x) = \frac{e^{-2,5}(2,5)^x}{x!}.$$

Portanto, a probabilidade de que nenhum caso de difteria seja registrado durante um determinado ano é

$$\begin{aligned} P(X = 0) &= \frac{e^{-2,5}(2,5)^0}{0!} \\ &= 0,082. \end{aligned}$$

A probabilidade de que um único caso seja registrado é

$$\begin{aligned} P(X = 1) &= \frac{e^{-2,5}(2,5)^1}{1!} \\ &= 0,205; \end{aligned}$$

analogamente,

$$\begin{aligned} P(X = 2) &= \frac{e^{-2,5}(2,5)^2}{2!} \\ &= 0,257, \end{aligned}$$

$$\begin{aligned} P(X = 3) &= \frac{e^{-2.5}(2.5)^3}{3!} \\ &= 0,214, \end{aligned}$$

$$\begin{aligned} P(X = 4) &= \frac{e^{-2.5}(2.5)^4}{4!} \\ &= 0,134, \end{aligned}$$

e

$$\begin{aligned} P(X = 5) &= \frac{e^{-2.5}(2.5)^5}{5!} \\ &= 0,067. \end{aligned}$$

Poderíamos ter consultado a Tabela A.2 no Apêndice A para determinar essas probabilidades ou ter usado um pacote estatístico.

Desde que os resultados de X são mutuamente exclusivos e exaustivos,

$$\begin{aligned} P(X \geq 4) &= 1 - P(X < 4) \\ &= 1 - (0,082 + 0,205 + 0,257 + 0,214) \\ &= 0,242. \end{aligned}$$

Há uma probabilidade de 24,2% de que observaremos quatro casos de difteria ou mais em determinado ano. Analogamente, a probabilidade de que observaremos seis casos ou mais é

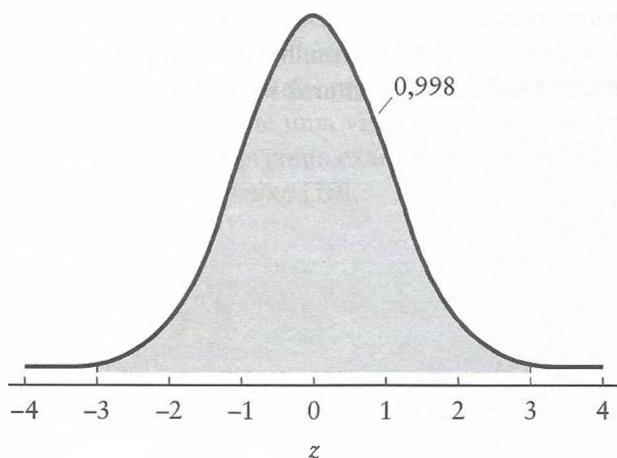
$$\begin{aligned} P(X \geq 6) &= 1 - P(X < 6) \\ &= 1 - (0,082 + 0,205 + 0,257 + 0,214 + 0,134 + 0,067) \\ &= 0,041. \end{aligned}$$

O número médio de casos por ano é $\lambda = 2,5$ e o desvio-padrão é $\sqrt{\lambda} = \sqrt{2,5} = 1,58$.

Se X pode assumir qualquer valor em um intervalo específico em vez de restringir-se a inteiros, então X é uma variável aleatória contínua. A distribuição contínua mais comum é a normal, a qual é definida por dois parâmetros: sua média μ e seu desvio-padrão σ . A média mede o centro da distribuição; o desvio-padrão quantifica a quantidade de espalhamento ou dispersão ao redor da média. A forma da distribuição normal indica que os resultados da variável aleatória X que estiverem próximos da média são mais prováveis de ocorrer do que os valores distantes dela.

A distribuição normal com média $\mu = 0$ e desvio-padrão $\sigma = 1$ é conhecida como a distribuição normal padrão. Devido à sua área ter sido tabulada, ela é usada para se obter probabilidades associadas às variáveis aleatórias normais. Por exemplo, suponha que desejamos conhecer a área sob a curva normal padrão que se encontra entre $z = -3,00$ e $z = 3,00$; equivalentemente, essa é a área no intervalo $\mu \pm 3\sigma$, mostrada na Figura 7.15. Ao consultarmos a Tabela A.3, encontramos a área à direita de $z = 3,00$ sendo 0,001. Uma vez que a curva normal padrão é simétrica, a área à esquerda de $z = -3,00$ também é 0,001. Portanto, a área entre $-3,00$ e $3,00$ é

$$\begin{aligned} P(-3 \leq Z \leq 3) &= 1 - [P(Z < -3) + P(Z > 3)] \\ &= 1 - 0,001 - 0,001 \\ &= 0,998; \end{aligned}$$

**FIGURA 7.15**

A curva normal padrão, área entre $z = -3,00$ e $z = 3,00$.

aproximadamente 99,8% da área sob a curva normal padrão se encontra dentro de ± 3 desvios-padrão da média.

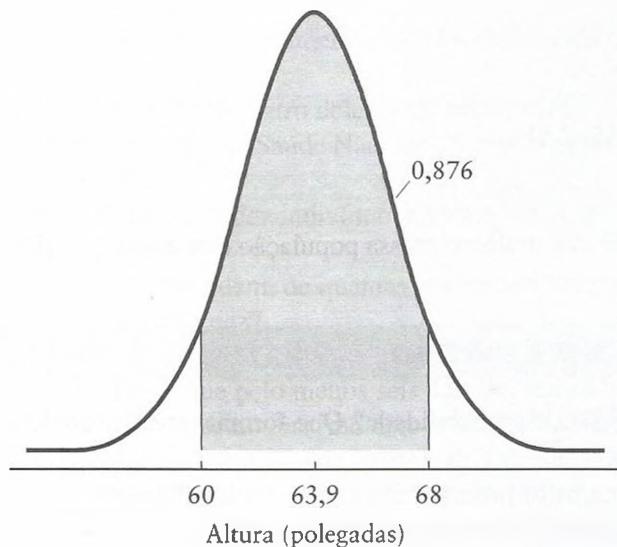
Se X é variável aleatória normal arbitrária com média μ e desvio-padrão σ , então

$$Z = \frac{X - \mu}{\sigma}$$

é variável aleatória normal padrão. Transformando X em Z , podemos usar a tabela de áreas para a curva normal padrão e estimar probabilidades associadas com X .

Por exemplo, seja X uma variável aleatória que representa altura. Para a população de mulheres de 18 a 74 anos nos Estados Unidos, a altura é normalmente distribuída com média $\mu = 63,9$ polegadas e desvio-padrão $\sigma = 2,6$ polegadas [9]. Essa distribuição está ilustrada na Figura 7.16. Observe que

$$Z = \frac{X - 63,9}{2,6}$$

**FIGURA 7.16**

Distribuição de altura para mulheres de 18 a 74 anos, Estados Unidos, 1976–1980.

é uma variável aleatória normal.

Se selecionarmos aleatoriamente uma mulher dessa população, qual a probabilidade de que ela tenha entre 60 e 68 polegadas de altura? Para $x = 60$,

$$\begin{aligned} z &= \frac{60 - 63,9}{2,6} \\ &= -1,50, \end{aligned}$$

e para $x = 68$,

$$\begin{aligned} z &= \frac{68 - 63,9}{2,6} \\ &= 1,58. \end{aligned}$$

Como resultado, a probabilidade de que x — a altura da mulher — se encontre entre 60 e 68 polegadas é igual à probabilidade de que z se encontre entre $-1,50$ e $1,58$ para a curva normal padrão. A área à esquerda de $z = -1,50$ é 0,067 e a área à direita de $z = 1,58$ é 0,057. (No lugar da Tabela A.3, poderíamos usar um pacote estatístico para gerar essas probabilidades.) Desde que a área total sob a curva seja igual a 1, a área entre $-1,50$ e $1,58$ é

$$\begin{aligned} P(60 \leq X \leq 68) &= P(-1,50 \leq Z \leq 1,58) \\ &= 1 - [P(Z < -1,50) + P(Z > 1,58)] \\ &= 1 - [0,067 + 0,057] \\ &= 0,876. \end{aligned}$$

A probabilidade de que a altura da mulher esteja entre 60 e 68 polegadas é 0,876.

Poderíamos também conhecer o valor da altura que limita os 5% superiores dessa distribuição. Da Tabela A.3, observamos que uma área de extremidade de 0,050 corresponde a $z = 1,645$. Resolvendo para x ,

$$\begin{aligned} z &= 1,645 \\ &= \frac{x - 63,9}{2,6} \end{aligned}$$

e

$$\begin{aligned} x &= 63,9 + (1,645)(2,6) \\ &= 68,2. \end{aligned}$$

Aproximadamente 5% das mulheres nessa população têm altura superior a 68,2 polegadas.

7.6 Exercícios de Revisão

- O que é distribuição de probabilidade? Que formas uma distribuição de probabilidade pode tomar?
- O que são parâmetros de uma distribuição de probabilidade?
- Quais são as três propriedades associadas à distribuição binomial?
- Quais são as três propriedades associadas à distribuição de Poisson?
- Quando a distribuição binomial é bem aproximada pela de Poisson?

6. Quais as propriedades da distribuição normal?
7. Explique a importância da distribuição normal padrão.
8. Seja X uma variável aleatória discreta que representa o número de serviços de diagnósticos que uma criança recebe durante uma visita ao consultório de um pediatra; esses serviços incluem procedimentos tais como exames de sangue e de urina. A distribuição de probabilidade para X aparece abaixo [10].

x	$P(X = x)$
0	0,671
1	0,229
2	0,053
3	0,031
4	0,010
5+	0,006
Total	1,000

- (a) Construa um gráfico da distribuição de probabilidade de X .
- (b) Qual a probabilidade de uma criança receber exatamente três serviços de diagnósticos durante uma visita ao consultório de um pediatra?
- (c) Qual a probabilidade de ela receber pelo menos um serviço? Quatro ou mais serviços?
- (d) Qual a probabilidade de a criança receber exatamente três serviços, se ela receber pelo menos um?
9. Suponha que você queira monitorar a poluição do ar de Los Angeles, Califórnia, no período de uma semana. Seja X uma variável aleatória que representa o número de dias dentro de sete nos quais a concentração de monóxido de carbono supera um nível específico. Você acredita que X tenha uma distribuição binomial? Explique.
10. Considere um grupo de sete indivíduos selecionados de uma população de 65 a 74 anos nos Estados Unidos. O número de pessoas que sofre de diabetes nessa amostra é uma variável aleatória binomial com parâmetros $n = 7$ e $p = 0,125$ [11].
 - (a) Se você deseja fazer uma lista de sete pessoas escolhidas, de quantas maneiras elas podem ser ordenadas?
 - (b) Sem se preocupar com a ordem, de quantas maneiras diferentes você pode selecionar quatro indivíduos desse grupo de sete?
 - (c) Qual a probabilidade de que exatamente dois dos indivíduos na amostra sofram de diabetes?
 - (d) Qual a probabilidade de que quatro deles tenham diabetes?
11. De acordo com o Levantamento de Saúde Nacional, 9,8% da população de 18 a 24 anos nos Estados Unidos é canhota [9].
 - (a) Suponha que você selecione dez indivíduos dessa população. De quantas maneiras eles podem ser ordenados?
 - (b) Sem se preocupar com a ordem, de quantas maneiras você pode selecionar quatro indivíduos desse grupo de dez?
 - (c) Qual a probabilidade de que exatamente três das dez pessoas sejam canhotas?
 - (d) Qual a probabilidade de que pelo menos seis das dez pessoas sejam canhotas?
 - (e) Qual a probabilidade de que no máximo dois indivíduos sejam canhotos?
12. De acordo com o Sistema de Vigilância de Fatores de Risco Comportamental, 58% de todos os americanos aderem ao estilo de vida sedentário [12].
 - (a) Se você selecionou amostras repetidas de tamanho 12 da população dos Estados Unidos, qual seria o número médio de indivíduos por amostra que não se exercita regularmente? Qual seria o desvio-padrão?

- (b) Suponha que você selecione uma amostra de 12 indivíduos e encontre dez que não se exercitam regularmente. Ao assumir que o Sistema de Vigilância esteja correto, qual a probabilidade de você obter resultados tão ruins ou piores do que os observados?
13. De acordo com o Departamento de Saúde de Massachusetts, 224 mulheres que deram à luz no Estado de Massachusetts em 1988 apresentaram resultado positivo para anticorpos de HIV. Assuma que 25% dos bebês nascidos dessas mães também se tornarão soro-positivos.
- Se amostras de tamanho 224 forem repetidamente selecionadas da população de crianças nascidas de mães com anticorpos de HIV, qual seria o número médio de crianças infectadas por amostra?
 - Qual seria o desvio-padrão?
 - Use a desigualdade de Chebychev para descrever essa distribuição.
14. O número de casos de tétano registrado nos Estados Unidos durante um único mês, em 1989, tem uma distribuição de Poisson com parâmetro $\lambda = 4,5$ [8].
- Qual a probabilidade de que exatamente um caso de tétano seja registrado durante um determinado mês?
 - Qual a probabilidade de que no máximo dois casos de tétano sejam registrados?
 - Qual a probabilidade de que quatro casos ou mais sejam registrados?
 - Qual o número médio de casos de tétano registrado no período de um mês? Qual é o desvio-padrão?
15. Em um determinado país, o número médio mensal de suicídios é 2,75 [13]. Assuma que o número de suicídios segue uma distribuição de Poisson.
- Qual a probabilidade de que nenhum suicídio seja registrado durante determinado mês?
 - Qual a probabilidade de que no máximo quatro suicídios sejam registrados?
 - Qual a probabilidade de que seis suicídios ou mais sejam registrados?
16. Seja X uma variável aleatória que representa o número de bebês em um grupo de 2.000 que morre antes de atingir o primeiro aniversário. Nos Estados Unidos, a probabilidade de que uma criança morra durante o primeiro ano de vida é 0,0085 [14].
- Qual é o número médio de bebês que morre em um grupo desse tamanho?
 - Qual a probabilidade de que no máximo cinco bebês dentre 2.000 morram em seus primeiros anos de vida?
 - Qual a probabilidade de que entre 15 e 20 bebês morram em seus primeiros anos de vida?
17. Considere a distribuição normal padrão com média $\mu = 0$ e desvio-padrão $\sigma = 1$.
- Qual a probabilidade de que um resultado z seja maior do que 2,60?
 - Qual a probabilidade de que z seja menor do que 1,35?
 - Qual a probabilidade de que z esteja entre -1,70 e 3,10?
 - Que valor de z limita os 15% superiores de uma distribuição normal padrão?
 - Que valor de z limita os 20% inferiores da distribuição?
18. Dentre as mulheres nos Estados Unidos de 18 e 74 anos, a pressão sanguínea diastólica é normalmente distribuída com média $\mu = 77$ mm Hg e desvio-padrão $\sigma = 11,6$ mm Hg [5].
- Qual a probabilidade de que uma mulher selecionada aleatoriamente tenha pressão sanguínea diastólica menor que 60 mm Hg?
 - Qual a probabilidade de que ela tenha pressão sanguínea diastólica maior do que 90 mm Hg?
 - Qual a probabilidade de que tenha pressão sanguínea diastólica entre 60 e 90 mm Hg?

19. A distribuição de pesos para a população masculina nos Estados Unidos é aproximadamente normal com média $\mu = 172,2$ libras e desvio-padrão $\sigma = 29,8$ libras [9].
- Qual a probabilidade de que um homem selecionado aleatoriamente pese menos do que 130 libras?
 - Qual a probabilidade de que pese mais do que 210 libras?
 - Qual a probabilidade de que entre cinco homens selecionados ao acaso da população, pelo menos um tenha peso fora do intervalo entre 130 e 210 libras?
20. No Estudo de Framingham, os níveis séricos de colesterol foram medidos para um grande número de homens sadios. A população foi acompanhada por 16 anos. No final desse período, os homens foram divididos em dois grupos: os que tinham desenvolvido doença cardíaca coronariana e os que não. As distribuições dos níveis séricos de colesterol para cada grupo foram consideradas aproximadamente normais. Entre os indivíduos que eventualmente desenvolveram a doença cardíaca coronariana, o nível sérico médio de colesterol foi $\mu_d = 244$ mg/100 ml e o desvio-padrão foi $\sigma_d = 51$ mg/100 ml; para os que não desenvolveram a doença, o nível médio sérico de colesterol foi $\mu_{nd} = 219$ mg/100 ml e o desvio-padrão foi $\sigma_{nd} = 41$ mg/100 ml [15].
- Suponha que um nível de colesterol inicial de 260 mg/100 ml ou maior seja usado para predizer a doença cardíaca coronariana. Qual é a probabilidade de se prever corretamente essa doença para um homem que virá a desenvolvê-la?
 - Qual é a probabilidade de se predizer a doença do coração para um homem que não a desenvolverá?
 - Qual é a probabilidade de falha em se predizer a doença do coração para um homem que virá a desenvolvê-la?
 - O que aconteceria às probabilidades de erros falso positivo e falso negativo se o ponto de corte para se prever a doença de coração for diminuído para 250 mg/100 ml?
 - Nessa população, os níveis iniciais séricos de colesterol parecem úteis para a previsão da doença cardíaca coronariana? Por quê?

Bibliografia

- [1] National Center for Health Statistics. "Supplements to the Monthly Vital Statistics Reports: Advance Reports, 1986". *Vital and Health Statistics*. série 24, n. 3, mar. 1990.
- [2] Centers for Disease Control. "The Surgeon General's 1989 Report on Reducing the Health Consequences of Smoking: 25 Years of Progress". *Morbidity and Mortality Weekly Report Supplement*. v. 38, 24 mar. 1989.
- [3] ROSS, S. M. *Introduction to Probability Models*. Orlando, Florida: Academic Press, 1985.
- [4] WILSON, R. e CROUCH, E. A. C. "Risk Assessment and Comparisons: An Introduction". *Science*. v. 236, 17 abr. 1987. p. 267–270.
- [5] National Center for Health Statistics. DRIZD, T., DANNENBERG, A. L. e Engel, A. "Blood Pressure Levels in Persons 18–74 Years of Age in 1976–1980, and Trends in Blood Pressure From 1960 to 1980 in the United States". *Vital and Health Statistics*. série 11, n. 234, jul. 1986.
- [6] CASTELLI, W. P. e ANDERSON, K. "Antihypertensive Treatment and Plasma Lipoprotein Levels: The Associations in Data from a Population Study". *American Journal of Medicine Supplement*. v. 80, 14 fev. 1986. p. 23–32.
- [7] TYE, L. "Many States Tackling Issue of AIDS-Infected Health Care Workers". *The Boston Globe*. 27 maio 1991. p. 29–30.
- [8] Centers for Disease Control. "Summary of Notifiable Diseases, United States, 1989". *Morbidity and Mortality Weekly Report*. v. 39, 5 out. 1990.

8

Distribuição Amostral da Média

No capítulo anterior, examinamos diversas distribuições teóricas de probabilidade, tais como a binomial e a normal. Em todos os casos, assumiu-se que os parâmetros relevantes da população eram conhecidos, o que nos permite descrever completamente as distribuições e calcular as probabilidades associadas com os vários resultados. Em muitas das aplicações práticas, no entanto, os valores desses parâmetros não são conhecidos. Então, precisamos tentar descrever ou estimar algumas características da população — tal como sua média ou seu desvio-padrão — com o uso da informação contida na amostra de observações. O processo de extrair conclusões de uma população inteira com base na informação de uma amostra é conhecido como *inferência estatística*.

8.1 Distribuições Amostrais

Suponha que nosso foco esteja em estimar o valor médio de alguma variável aleatória contínua de interesse. Por exemplo, podemos fazer uma afirmação sobre o nível sérico médio de colesterol de todos os homens residentes nos Estados Unidos, com base em uma amostra extraída dessa população. A abordagem óbvia seria usar a média da amostra como estimativa da média μ da população desconhecida. A quantidade \bar{X} é chamada *estimador* do parâmetro μ . Filosoficamente, existem muitas abordagens diferentes do processo de estimação; nesse caso, por ser a população assumida normalmente distribuída, a média da amostra \bar{X} é um *estimador de máxima verossimilhança* [1]. O método de máxima verossimilhança encontra o valor do parâmetro mais provável de ter produzido os dados observados da amostra. Usualmente, pode-se contar com esse método para se obter estimadores razoáveis. Tenha em mente, no entanto, que duas amostras diferentes provavelmente produzam diferentes médias de amostra; consequentemente, um certo grau de incerteza está envolvido. Antes que apliquemos esse procedimento de estimação, portanto, precisamos examinar algumas das propriedades da média da amostra e os modos como ela pode variar.

A população investigada pode ser qualquer grupo que escolhermos. Geralmente, estimamos a média da população μ com maior precisão quando o grupo é relativamente homogêneo. Se há somente uma pequena variação entre os indivíduos, estaremos mais seguros de que as observações em qualquer amostra são representativas do grupo inteiro.

É muito importante que a amostra forneça uma representação precisa da população da qual ela é selecionada. Do contrário, as conclusões sobre a população podem estar distorcidas ou viesadas. Por exemplo, se tencionamos fazer uma afirmação sobre o nível médio sé-

rico de colesterol para todos os homens de 20 a 74 anos nos Estados Unidos, mas amostramos somente homens acima de 60 anos, é provável que nossa estimativa da média da população seja muito alta. É fundamental que a amostra extraída seja *aleatória*; cada indivíduo da população deve ter igual probabilidade de ser selecionado. Esse ponto será discutido posteriormente no Capítulo 22. Além disso, esperamos que, quanto maior for a amostra, mais confiável seja nossa estimativa da média da população.

Suponha que em uma população especificada, a média da variável aleatória contínua nível sérico de colesterol seja μ e o desvio-padrão σ . Selecionamos aleatoriamente uma amostra de n observações da população e calculamos sua média; a qual chamaremos de \bar{x}_1 . Obtemos uma segunda amostra aleatória de n observações e também calculamos sua média, a qual chamaremos de \bar{x}_2 . A menos que todos na população tenham o mesmo nível sérico de colesterol, é muito improvável que \bar{x}_1 seja igual a \bar{x}_2 . Se continuássemos esse processo indefinidamente — selecionar todos as possíveis amostras de tamanho n e calcular suas médias —, terminaríamos com um conjunto de valores que se consistiria inteiramente de médias de amostra. Outro procedimento seria notar que o estimador \bar{X} realmente é uma variável aleatória com resultados $\bar{x}_1, \bar{x}_2, \bar{x}_3$ e assim por diante.

Se cada média nessa série é tratada como uma única observação, sua distribuição de probabilidade coletiva — a distribuição de probabilidade de \bar{X} — é conhecida como *distribuição amostral* de médias das amostras de tamanho n . Por exemplo, se relacionássemos amostras repetidas de tamanho 25 da população masculina residente nos Estados Unidos e calculássemos a média do nível sérico de colesterol para cada uma, terminaríamos com a distribuição amostral dos níveis médios séricos de colesterol de amostras de tamanho 25. Na prática, não é comum selecionar amostras repetidas de tamanho n de determinada população. No entanto, entender as propriedades da distribuição teórica de suas médias permite-nos inferir, fundamentados em uma **única** amostra de tamanho n .

8.2 O Teorema Central do Limite

Desde que a distribuição de níveis séricos de colesterol na população original tenha média μ e desvio-padrão σ , a distribuição de médias da amostra calculadas para amostras de tamanho n tem três propriedades importantes:

1. A média da distribuição amostral é idêntica à média μ da população.
2. O desvio-padrão da distribuição de médias da amostra é igual à σ/\sqrt{n} . Essa quantidade é conhecida como *erro-padrão* da média.
3. Com a condição de que n seja suficientemente grande, a forma da distribuição amostral é aproximadamente normal.

Intuitivamente, poderíamos esperar que as médias de todas as amostras se agrupassem ao redor da média da população da qual foram extraídas. Embora o desvio-padrão da distribuição amostral esteja relacionado com o desvio-padrão σ da população, há menor variabilidade entre as médias da amostra do que entre as observações individuais. Mesmo que uma amostra particular contenha um ou dois valores extremos, é provável que eles estejam deslocados por outras medidas no grupo. Assim, enquanto n for maior do que 1, o erro-padrão da média é sempre menor do que o desvio-padrão da população. Além disso, conforme n aumenta, a quantidade de variação amostral diminui. Finalmente, se n é suficientemente grande, a distribuição das médias amostrais é aproximadamente normal. Esse resultado notável é conhecido como *teorema central do limite* e aplica-se a qualquer população com desvio-padrão finito, independentemente da forma da distribuição original [2]. Entretanto, quanto mais a

população original se afasta da distribuição normal, maior é o valor de n necessário para assegurar a normalidade da distribuição amostral. Se a própria população original é normal, amostras de tamanho 1 são suficientemente grandes. Para uma população que seja bimodal ou notavelmente assimétrica, com freqüência uma amostra de tamanho 30 é suficiente.

O teorema central do limite é muito poderoso. Ele se mantém verdadeiro não somente para níveis séricos de colesterol, como também para quase qualquer outro tipo de medida e se aplica até para variáveis aleatórias discretas. O teorema central do limite nos permite quantificar a incerteza inherente na inferência estatística sem ser necessário fazer grandes suposições que não podem ser verificadas. Independentemente da distribuição de X , devido à distribuição das médias amostrais ser aproximadamente normal com média μ e desvio-padrão σ/\sqrt{n} , sabemos que, se n é suficientemente grande,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

é normalmente distribuída com média 0 e desvio-padrão 1. Simplesmente padronizamos a variável aleatória normal \bar{X} do modo usual. Como resultado, usamos as tabelas de distribuição normal padrão — tal como a Tabela A.3 no Apêndice A — para inferir o valor da média da população.

8.3 Aplicações do Teorema Central do Limite

Considere a distribuição dos níveis séricos de colesterol para todos os homens de 20 a 74 anos que vivem nos Estados Unidos. A média dessa população é $\mu = 211$ mg/100 ml e o desvio-padrão é $\sigma = 46$ mg/100 ml [3]. Se selecionarmos amostras repetidas de tamanho 25 da população, que proporção de amostras terá um valor médio de 230 mg/100 ml ou acima?

Ao se assumir que a amostra de tamanho 25 seja suficientemente grande, o teorema central do limite estabelece que a distribuição de médias de amostras de tamanho 25 é aproximadamente normal com média $\mu = 211$ mg/100 ml e erro-padrão $\sigma/\sqrt{n} = 46/\sqrt{25} = 9,2$ mg/100 ml. Essa distribuição amostral e a distribuição da população original estão na Figura 8.1. Note-se que

$$Z = \frac{\bar{X} - 211}{9,2}$$

é uma variável aleatória normal padrão. Se $\bar{x} = 230$, então

$$\begin{aligned} z &= \frac{230 - 211}{9,2} \\ &= 2,07. \end{aligned}$$

Ao consultarmos a Tabela A.3, verificamos que a área à direita de $z = 2,07$ é 0,019. Sómente cerca de 1,9% das amostras terá média maior do que 230 mg/100 ml. Equivalentemente, se selecionarmos uma amostra simples de tamanho 25 da população de homens de 20 a 74 anos, a probabilidade de que o nível médio sérico de colesterol para essa amostra será 230 mg/100 ml ou maior é de 0,019.

Que valor médio de nível sérico de colesterol limita os 10% mais baixos da distribuição amostral das médias? Ao localizarmos 0,100 no corpo da Tabela A.3, vemos que corresponde ao valor $z = -1,28$. Resolvendo para \bar{x} ,

$$z = -1,28$$

$$= \frac{\bar{x} - 211}{9,2}$$

e

$$\bar{x} = 211 + (-1,28)(9,2)$$

$$= 199,2.$$

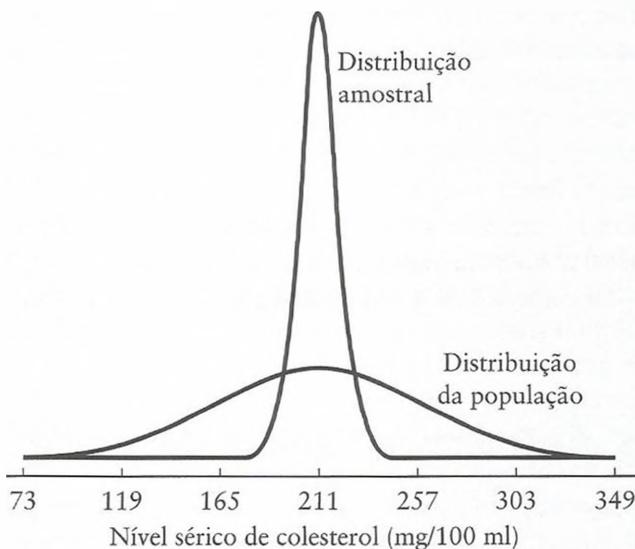


FIGURA 8.1

Distribuições de valores individuais e médias de amostras de tamanho 25 para os níveis séricos de colesterol de homens de 20 a 74 anos, Estados Unidos, 1976–1980.

Portanto, aproximadamente 10% das amostras de tamanho 25 têm médias menores ou iguais a 199,2 mg/100 ml.

Calcularemos agora os limites superior e inferior que incluem 95% das médias das amostras de tamanho 25 extraídas da população. Uma vez que 2,5% da área sob a curva normal padrão se encontram acima de $z = 1,96$ e os outros 2,5% se encontram abaixo de $z = -1,96$,

$$P(-1,96 \leq Z \leq 1,96) = 0,95.$$

Assim, estamos interessados em resultados de Z para os quais

$$-1,96 \leq Z \leq 1,96.$$

Gostaríamos de transformar essa desigualdade em uma afirmação sobre \bar{X} . Substituindo Z por $(\bar{X} - 211)/9,2$,

$$-1,96 \leq \frac{\bar{X} - 211}{9,2} \leq 1,96.$$

Multiplicando os três termos da desigualdade por 9,2 e somando 211, resultará em

$$211 - 1,96(9,2) \leq \bar{X} \leq 211 + 1,96(9,2),$$

ou

$$193,0 \leq \bar{X} \leq 229,0.$$

Isso nos informa que aproximadamente 95% das médias das amostras de tamanho 25 estão entre 193,0 mg/100 ml e 229,0 mg/100 ml. Consequentemente, se selecionarmos uma amostra aleatória de tamanho 25 registrada como a da população de níveis séricos de colesterol para todos os homens de 20 a 74 anos e ela tiver média maior do que 229,0 ou menor do que 193,0 mg/100 ml, seríamos suspeitos ao fazermos essa alegação. Ou a amostra aleatória foi realmente extraída de uma população diferente ou um evento raro se realizou. Para o propósito desta discussão, “um evento raro” é definido como um resultado que ocorre menos que 5% das vezes.

Suponha que tivéssemos selecionado amostras de tamanho 10 da população no lugar de tamanho 25. Nesse caso, o erro-padrão de \bar{X} seria $46/\sqrt{10} = 14,5$ mg/100 ml, e construiríamos a desigualdade

$$-1,96 \leq \frac{\bar{X} - 211}{14,5} \leq 1,96.$$

Os limites superior e inferior que contêm 95% das médias seriam

$$182,5 \leq \bar{X} \leq 239,5.$$

Esse intervalo é mais amplo do que o calculado para as amostras de tamanho 25. Esperamos que a quantidade de variação amostral aumente conforme o tamanho da amostra diminua. Ao extraímos amostras de tamanho 50, teríamos nos limites superior e inferior

$$198,2 \leq \bar{X} \leq 223,8;$$

nada surpreendente, esse intervalo é mais estreito do que o construído para amostras de tamanho 25. Amostras de tamanho 100 produzem os limites

$$202,0 \leq \bar{X} \leq 220,0.$$

Em suma, incluindo o caso para o qual $n = 1$, temos os seguintes resultados:

n	σ/\sqrt{n}	Intervalo que Contém 95% das Médias	Comprimento do Intervalo
1	46,0	$120,8 \leq \bar{X} \leq 301,2$	180,4
10	14,5	$182,5 \leq \bar{X} \leq 239,5$	57,0
25	9,2	$193,0 \leq \bar{X} \leq 229,0$	36,0
50	6,5	$198,2 \leq \bar{X} \leq 223,8$	25,6
100	4,6	$202,0 \leq \bar{X} \leq 220,0$	18,0

Conforme o tamanho das amostras aumenta, a quantidade de variabilidade entre as médias da amostra — quantificada pelo erro-padrão σ/\sqrt{n} — diminui; consequentemente, os limites que englobam 95% dessas médias se aproximam. O comprimento de um intervalo é simplesmente o limite superior menos o limite inferior.

Todos os intervalos que construímos têm sido simétricos ao redor da média da população de 211 mg/100 ml. Claramente, existem outros intervalos que obteriam a proporção apropriada de médias da amostra. Suponha que novamente desejemos construir um intervalo que contenha 95% das médias das amostras de tamanho 25. Desde que 1% da área sob a curva normal padrão se encontre acima de $z = 2,32$ e 4% se encontre abaixo de $z = -1,75$, sabemos que

$$P(-1,75 \leq Z \leq 2,32) = 0,95.$$

Como resultado, estamos interessados nos valores de Z para os quais

$$-1,75 \leq Z \leq 2,32.$$

Ao substituirmos Z por $(\bar{X} - 211)/9,2$, verificamos que o intervalo é

$$194,9 \leq \bar{X} \leq 232,3.$$

Portanto, podemos dizer que aproximadamente 95% das médias das amostras de tamanho 25 se encontram entre 194,9 mg/100 ml e 232,3 mg/100 ml. Normalmente, é preferível, entretanto, construir um intervalo simétrico, essencialmente porque é o menor intervalo que contém a proporção apropriada das médias (uma exceção a essa regra é o intervalo unilateral; vamos retornar a esse caso especial logo abaixo). Nesse exemplo, o intervalo assimétrico tem comprimento $232,3 - 194,9 = 37,4$ mg/100 ml; o comprimento do intervalo simétrico é $229,0 - 193,0 = 36,0$ mg/100 ml.

Agora abordaremos uma questão um pouco mais complicada: de que tamanho as amostras precisam ser para que 95% de suas médias se encontrem a ± 5 mg/100 ml da média μ da população? Para responder, não é necessário conhecer o valor do parâmetro μ : simplesmente verificamos o tamanho da amostra n para o qual

$$P(\mu - 5 \leq \bar{X} \leq \mu + 5) = 0,95,$$

ou

$$P(-5 \leq \bar{X} - \mu \leq 5) = 0,95.$$

Para começar, dividimos os três termos da desigualdade pelo erro-padrão $\sigma/\sqrt{n} = 46/\sqrt{n}$, o que resulta em,

$$P\left(\frac{-5}{46/\sqrt{n}} \leq \frac{\bar{X} - \mu}{46/\sqrt{n}} \leq \frac{5}{46/\sqrt{n}}\right) = 0,95.$$

Uma vez que Z é igual a $(\bar{X} - \mu)/(46/\sqrt{n})$,

$$P\left(\frac{-5}{46/\sqrt{n}} \leq Z \leq \frac{5}{46/\sqrt{n}}\right) = 0,95.$$

Lembre-se de que 95% da área sob a curva normal padrão se encontra entre $z = -1,96$ e $z = 1,96$. Logo, para encontrar o tamanho da amostra n , poderíamos usar o limite superior do intervalo e resolver a equação

$$z = 1,96$$

$$= \frac{5}{46/\sqrt{n}};$$

equivalentemente, poderíamos usar o limite inferior e resolver

$$\begin{aligned} z &= -1,96 \\ &= \frac{-5}{46/\sqrt{n}}. \end{aligned}$$

Se tomarmos

$$1,96 = \frac{5\sqrt{n}}{46}$$

e multiplicarmos ambos os lados da equação por 46/5, verificamos que

$$\sqrt{n} = \frac{1,96(46)}{5}$$

e

$$\begin{aligned} n &= \left[\frac{1,96(46)}{5} \right]^2 \\ &= 325,2. \end{aligned}$$

O convencional, quando se trabalha com amostras, é arredondar seus valores. Assim, amostras de tamanho 326 seriam exigidas para que 95% das suas médias se encontrem a ± 5 mg/100 ml da média μ da população. Outro modo de estabelecer essa convenção é selecionarmos uma amostra de tamanho 326 da população e calcularmos sua média. A probabilidade de que a média da amostra esteja a ± 5 mg/100 ml da verdadeira média μ da população é 0,95.

Até aqui, focalizamos intervalos de duas extremidades; encontramos os limites superior e inferior que contêm a proporção especificada das médias das amostras. Mais especificamente, focalizamos intervalos simétricos. Em algumas situações, no entanto, interessamos por um intervalo com uma extremidade. Por exemplo, poderíamos querer encontrar o limite superior para os 95% dos níveis médios séricos de colesterol de amostras de tamanho 25. Desde que 95% da área sob a curva normal padrão se encontrem abaixo de $z = 1,645$,

$$P(Z \leq 1,645) = 0,95.$$

Conseqüentemente, estamos interessados em valores de Z para os quais

$$Z \leq 1,645.$$

Substituindo Z por $(\bar{X} - 211)/9,2$ produz-se

$$\frac{\bar{X} - 211}{9,2} \leq 1,645,$$

ou

$$\bar{X} \leq 226,1.$$

Aproximadamente 95% das médias das amostras de tamanho 25 se encontram abaixo de 226,1 mg/100 ml.

Se desejamos construir um limite inferior para 95% dos níveis médios séricos de colesterol, focalizaremos os valores de Z que se encontram acima de $-1,645$; nesse caso, resolvemos

$$\frac{\bar{X} - 211}{9,2} \geq -1,645$$

para encontrar

$$\bar{X} \geq 195,9.$$

Aproximadamente 95% das médias de amostras de tamanho 25 se encontram acima de 195,9 mg/100 ml.

Tenha sempre em mente que precisamos ser cuidadosos quando fizermos múltiplas afirmações sobre a distribuição amostral da média. Para as amostras de níveis séricos de colesterol de tamanho 25, verificamos que a probabilidade de a média da amostra se encontrar dentro do intervalo é 0,95.

$$(193,0, 229,0).$$

Dissemos também que a probabilidade é 0,95 de que a média se encontre abaixo de 226,1 mg/100 ml e 0,95 de que está acima de 195,9 mg/100 ml. Embora essas três afirmações sejam corretas individualmente, não são verdadeiras simultaneamente. Os três eventos não são independentes. Para ocorrerem ao mesmo tempo, a média da amostra teria que se encontrar no intervalo

$$(195,9, 226,1);$$

a probabilidade de que isso aconteça não é igual a 0,95.

8.4 Aplicações Adicionais

Considere a distribuição da idade ao morrer da população dos Estados Unidos em 1979–1981. Essa distribuição está na Figura 8.2; tem média $\mu = 73,9$ anos e desvio-padrão $\sigma = 18,1$ anos e está longe de ser normalmente distribuída [4]. O que esperamos que aconteça quando tiramos amostras dessa população de idades?

Em vez de extraímos fisicamente amostras da população, geramos um programa de computador para simular esse processo. O computador é usado para conduzir uma *simulação* e modelar um experimento ou procedimento, de acordo com uma distribuição de proba-

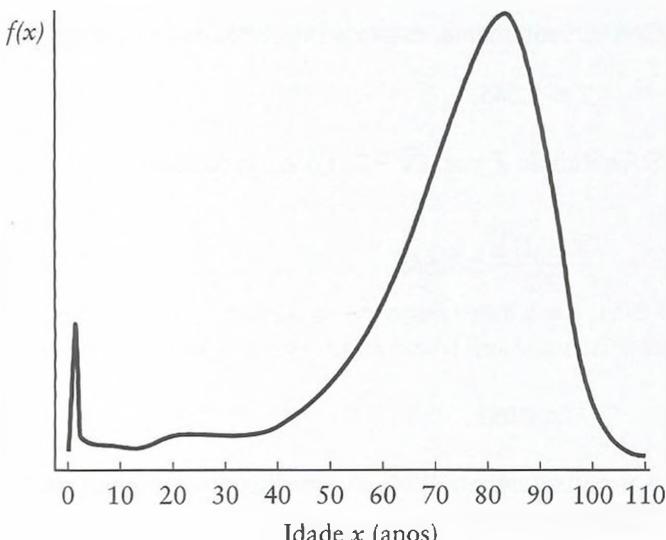


FIGURA 8.2
Distribuição de idade ao morrer, Estados Unidos, 1979–1981.

bilidade específica. Em nosso exemplo, o procedimento consistiria em selecionar uma observação individual da distribuição, mostrada na Figura 8.2. O computador é instruído para repetir o processo determinado número de vezes, mantendo a trilha dos resultados.

Usamos o computador para ilustrar essa técnica e simular a seleção de quatro amostras aleatórias de tamanho 25 da população de idades ao morrer da população dos Estados Unidos. Seus histogramas estão na Figura 8.3; suas médias e desvios-padrão estão a seguir:

Amostra de Tamanho 25	\bar{x}	s
1	71,3	18,1
2	69,2	25,6
3	74,0	14,0
4	76,8	15,0

Note-se que as quatro amostras aleatórias não são idênticas. Cada vez que selecionamos um conjunto de 25 medidas da população, as observações da amostra mudam. Como resultado, os valores de \bar{x} e s — nossas estimativas da média μ e do desvio-padrão σ da população — diferem de amostra para amostra. Essa variação aleatória é conhecida como *variabilidade amostral*. Nas quatro amostras de tamanho 25 selecionadas abaixo, as estimativas de μ variam de 69,2 anos até 76,8 anos. Analogamente, as estimativas de σ variam de 14,0 a 25,6 anos.

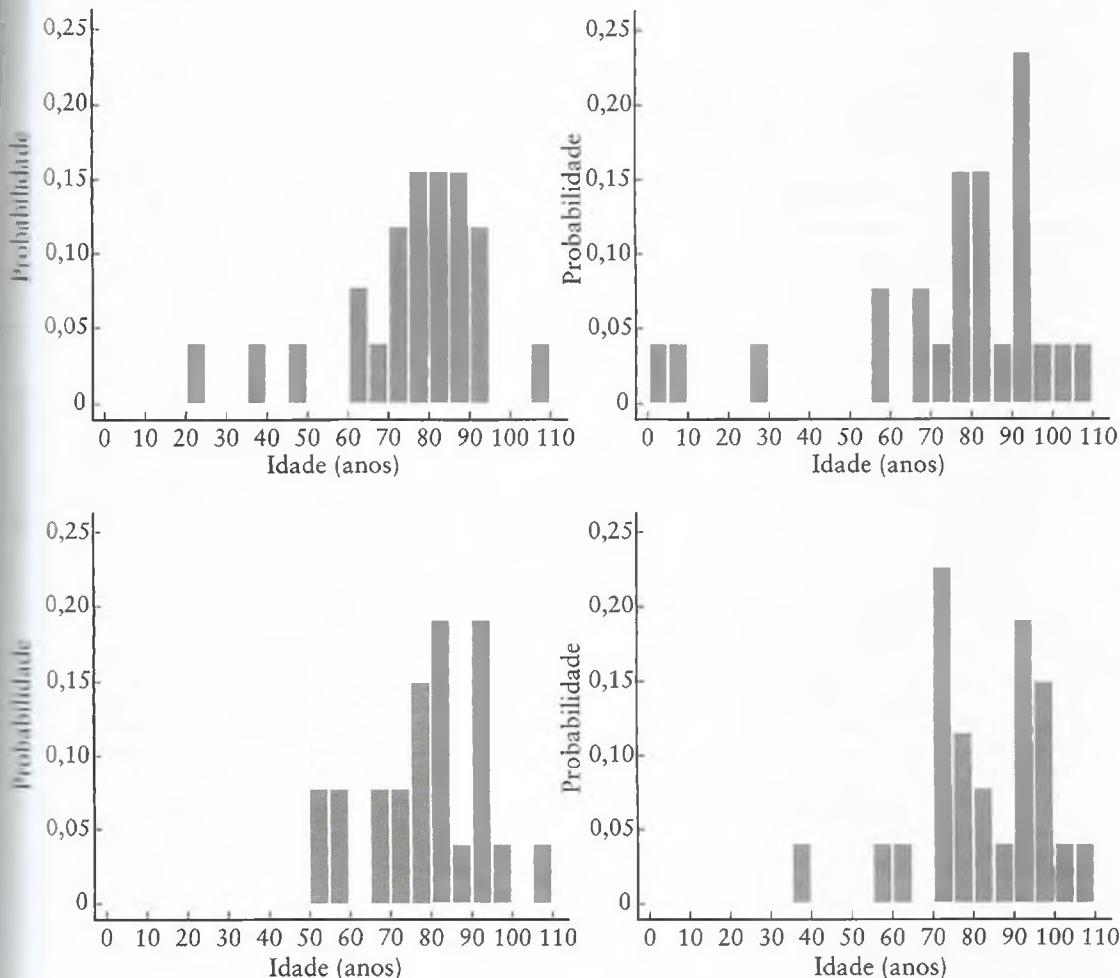


FIGURA 8.3
Histogramas de quatro amostras de tamanho 25.

Suponha agora que, em vez de selecionarmos amostras de tamanho 25, escolhêssemos quatro amostras aleatórias de tamanho 100 da população de idades ao morrer. Novamente usamos o computador para simular esse processo. Os histogramas das amostras são exibidos na Figura 8.4 e suas médias e desvios-padrão estão a seguir.

Amostra de Tamanho 100	\bar{x}	s
1	75,4	16,5
2	75,0	19,9
3	73,5	18,1
4	72,1	20,2

Para essas amostras, as estimativas de μ variam de 72,1 a 75,4 anos e as estimativas de σ de 16,5 a 20,2 anos. Esses intervalos são menores do que os intervalos correspondentes para amostras de tamanho 25. De fato, esperávamos isso; quando o tamanho da amostra aumenta, a quantidade da variabilidade amostral diminui.

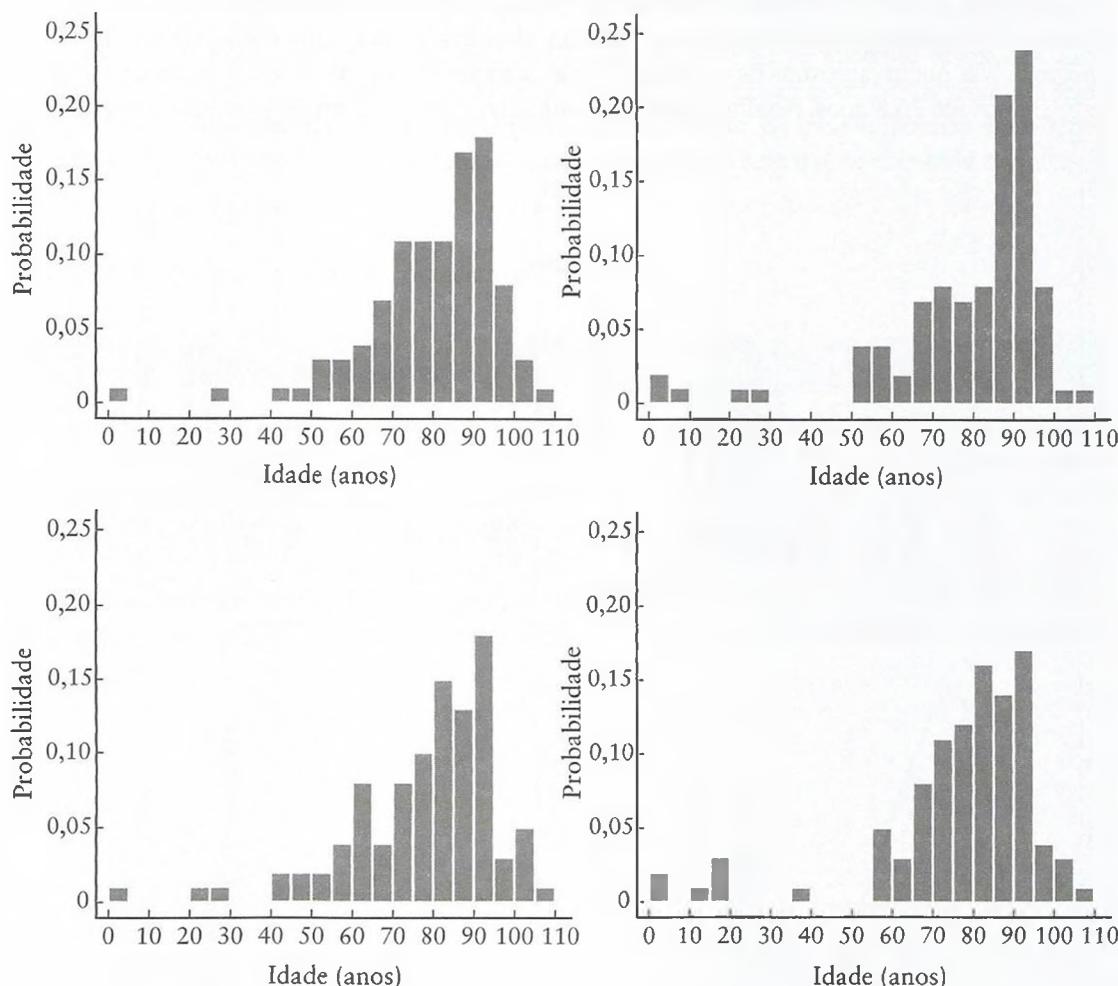


FIGURA 8.4
Histogramas de quatro amostras de tamanho 100.

A seguir, selecionamos quatro amostras aleatórias de tamanho 500, a partir da população de idades ao morrer. Os histogramas aparecem na Figura 8.5 e as médias e os desvios-padrão estão abaixo.

Amostra de Tamanho 500	\bar{x}	s
1	74,3	17,1
2	73,4	18,1
3	73,5	18,6
4	74,2	17,8

Novamente, os intervalos das estimativas tanto para μ como para σ diminuem.

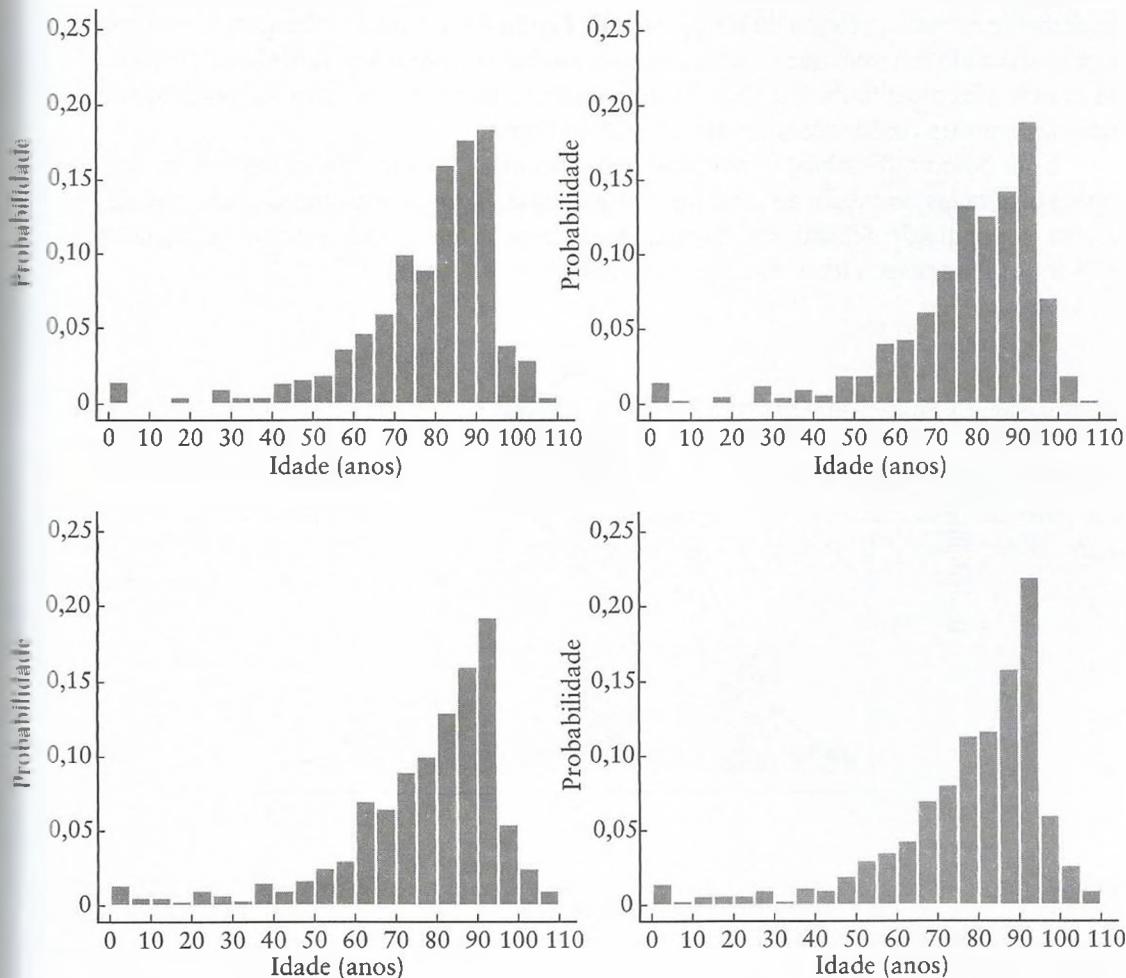


FIGURA 8.5
Histogramas de quatro amostras de tamanho 500.

Ao observarmos as Figuras 8.3 a 8.5, vemos que, conforme o tamanho das amostras aumenta, suas distribuições se aproximam da forma da distribuição da população, conforme Figura 8.2. Embora ainda existam diferenças entre as amostras, a quantidade de variabilidade nas estimativas de \bar{x} e s diminui. Essa propriedade é conhecida como *consistência*; à me-

dida que as amostras selecionadas se tornam maiores, as estimativas dos parâmetros da população se aproximam de seus valores-alvo.

A população de idades ao morrer pode também ser usada para demonstrar uma aplicação do teorema central do limite. Para isso, selecionamos amostras repetidas de tamanho n da população com média $\mu = 73,9$ anos e desvio-padrão $\sigma = 18,1$ anos e examinamos a distribuição das suas médias. Teoricamente, enumeramos todas as possíveis amostras aleatórias; agora, entretanto, selecionamos 100 amostras de tamanho 25. Um histograma das 100 médias amostrais está exibido na Figura 8.6.

De acordo com o teorema central do limite, a distribuição das médias amostrais possui três propriedades. Primeiramente, sua média deve ser igual à média da população $\mu = 73,9$ anos. De fato, a média das 100 médias amostrais é 74,1 anos. Em segundo lugar, esperamos que o erro-padrão das médias das amostras seja $\sigma/\sqrt{n} = 18,1/\sqrt{25} = 3,6$ anos. Na realidade, o erro-padrão é 3,7 anos. Finalmente, a distribuição das médias amostrais deve ser aproximadamente normal. A forma do histograma na Figura 8.6 e a da distribuição normal teórica superposta a ele sugerem que essa terceira propriedade se mantenha verdadeira. Note-se que há grande afastamento da distribuição da população exibida na Figura 8.2 ou de qualquer uma das amostras individuais de tamanho 25 na Figura 8.3.

Com base na distribuição amostral, calculamos as probabilidades associadas com os vários resultados da média da amostra. Por exemplo, entre as amostras de tamanho 25 extraídas da população de idades ao morrer, que proporção tem média que se encontra entre 70 e 78 anos? Para responder a essa questão, precisamos encontrar $P(70 \leq \bar{X} \leq 78)$.

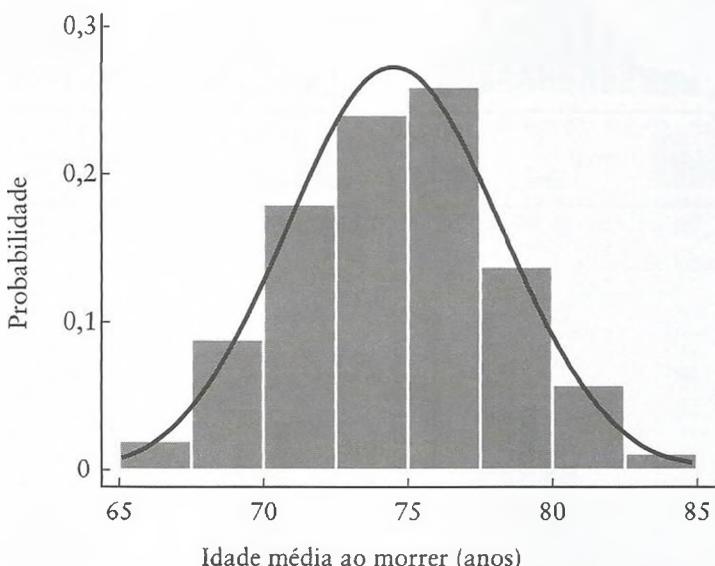


FIGURA 8.6
Histograma de 100 médias amostrais de amostras de tamanho 25.

Conforme foi visto, o teorema central do limite estabelece que a distribuição das médias de amostra de tamanho 25 é aproximadamente normal com média $\mu = 73,9$ anos e erro-padrão $\sigma/\sqrt{n} = 18,1/\sqrt{25} = 3,62$ anos. Logo,

$$\begin{aligned} Z &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \\ &= \frac{\bar{X} - 73,9}{3,62} \end{aligned}$$

é uma variável aleatória normal padrão. Se representamos a desigualdade na expressão

$$P(70 \leq \bar{X} \leq 78)$$

em termos de Z no lugar de \bar{X} , podemos usar a Tabela A.3 para encontrar a proporção de amostras que têm valor médio nesse intervalo.

Começamos subtraindo 73,9 de cada termo na desigualdade e dividindo por 3,62; assim podemos expressar

$$P(70 \leq \bar{X} \leq 78)$$

como

$$P\left(\frac{70 - 73,9}{3,62} \leq \frac{\bar{X} - 73,9}{3,62} \leq \frac{78 - 73,9}{3,62}\right),$$

ou

$$P(-1,08 \leq Z \leq 1,13).$$

Sabemos que a área total sob a curva normal padrão é igual a 1. De acordo com a Tabela A.3, a área à direita de $z = 1,13$ é 0,129 e a área à esquerda de $z = -1,08$ é 0,140. Portanto,

$$\begin{aligned} P(-1,08 \leq Z \leq 1,13) &= 1 - 0,129 - 0,140 \\ &= 0,731. \end{aligned}$$

Aproximadamente 73,1% das amostras de tamanho 25 têm uma média que se encontra entre 70 e 78 anos.

Que proporção das médias de amostras de tamanho 100 se encontra entre 70 e 78 anos? Novamente precisamos encontrar $P(70 \leq \bar{X} \leq 78)$. Dessa vez, no entanto, \bar{X} tem uma distribuição normal com média $\mu = 73,9$ anos e erro-padrão $\sigma/\sqrt{n} = 18,1/\sqrt{100} = 1,81$ anos. Assim, construímos a desigualdade

$$P\left(\frac{70 - 73,9}{1,81} \leq \frac{\bar{X} - 73,9}{1,81} \leq \frac{78 - 73,9}{1,81}\right),$$

ou

$$P(-2,15 \leq Z \leq 2,27).$$

De acordo com a Tabela A.3, a área à direita de $z = 2,27$ é 0,012 e a área à esquerda de $z = -2,15$ é 0,016. Logo,

$$\begin{aligned} P(-2,15 \leq Z \leq 2,27) &= 1 - 0,012 - 0,016 \\ &= 0,972. \end{aligned}$$

Cerca de 97,2% das amostras de tamanho 100 têm uma média que se encontra entre 70 e 78 anos. Se selecionássemos uma amostra aleatória única de tamanho 100 e verificássemos que sua média de amostra é $\bar{x} = 80$ anos, a amostra realmente veio de uma população com média original diferente — maior do que $\mu = 73,9$ anos — ou um evento raro ocorreu.

Para direcionar uma questão diferente, poderíamos querer encontrar os limites superior e inferior que incluem 80% das médias das amostras de tamanho 100. Ao consultarmos a

Tabela A.3, verificamos que 10% da área sob a curva normal padrão se encontra acima de $z = 1,28$ e outros 10% se encontram abaixo de $z = -1,28$. Desde que 80% da área se encontre entre $-1,28$ e $1,28$, estamos interessados em valores de Z para os quais

$$-1,28 \leq Z \leq 1,28,$$

e valores de \bar{X} para os quais

$$-1,28 \leq \frac{\bar{X} - 73,9}{1,81} \leq 1,28.$$

Multiplicando os três termos da desigualdade por 1,81 e somando 73,9, resulta em

$$73,9 + (-1,28)(1,81) \leq \bar{X} \leq 73,9 + (1,28)(1,81),$$

ou, equivalentemente,

$$71,6 \leq \bar{X} \leq 76,2.$$

Logo, 80% das médias das amostras de tamanho 100 se encontram entre 71,6 anos e 76,2 anos.

8.5 Exercícios de Revisão

1. O que é inferência estatística?
2. Por que é importante que uma amostra extraída de uma população seja aleatória?
3. Por que é necessário entender as propriedades de uma distribuição teórica de médias de amostras de tamanho n quando na prática você selecionará somente uma única de tal amostra?
4. O que é erro-padrão de uma média amostral? Como ele se compara ao desvio-padrão da população?
5. Explique o teorema central do limite.
6. O que acontece com a variabilidade da amostra de um conjunto de médias amostrais $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots$ conforme o tamanho das amostras aumenta?
7. O que é consistência?
8. Entre os adultos nos Estados Unidos, a distribuição de níveis de albumina (um tipo de proteína) no fluido cerebroespinal é aproximadamente simétrica com média $\mu = 29,5$ mg/100 ml e desvio-padrão $\sigma = 9,25$ mg/100 ml [5]. Suponha que você selecione amostras repetidas de tamanho 20 dessa população e calcule a média para cada amostra.
 - (a) Se você selecionasse um grande número de amostras aleatórias de tamanho 20, qual seria a média das médias das amostras?
 - (b) Quais seriam seus desvios-padrão? Qual é o outro nome para esse desvio-padrão das médias das amostras?
 - (c) Como o desvio-padrão das médias das amostras se compara com o desvio-padrão dos próprios níveis de albumina?
 - (d) Se você tomasse todas as diferentes médias de amostras e as usasse para construir um histograma, qual seria a forma de sua distribuição?
 - (e) Que proporção das médias de amostras de tamanho 20 é maior do que 33 mg/100 ml?
 - (f) Que proporção das médias é menor do que 28 mg/100 ml?
 - (g) Que proporção das médias está entre 29 e 31 mg/100 ml?

9. Considere uma variável aleatória X que tem uma distribuição normal padrão com média $\mu = 0$ e desvio-padrão $\sigma = 1$.
- O que você pode dizer sobre a distribuição das médias das amostras de tamanho 10 extraídas dessa população? Liste três propriedades.
 - Que proporção das médias das amostras de tamanho 10 é maior do que 0,60?
 - Que proporção das médias é menor do que -0,75?
 - Que valor limita os 20% superiores da distribuição das médias das amostras de tamanho 10?
 - Que valor limita os 10% inferiores da distribuição das médias?
10. Em Denver, Colorado, a distribuição das medidas diárias de ácido nítrico ambiental — um líquido corrosivo — é assimétrica à direita; ela tem média $\mu = 1,81 \mu\text{g}/\text{m}^3$ e desvio-padrão $\sigma = 2,25 \mu\text{g}/\text{m}^3$ [6]. Descreva a distribuição das médias das amostras de tamanho 40 selecionadas dessa população.
11. Na Noruega, a distribuição dos pesos ao nascer cuja idade gestacional é 40 semanas é aproximadamente normal com média $\mu = 3.500$ gramas e desvio-padrão $\sigma = 430$ gramas [7].
- Tomando-se um recém-nascido cuja idade gestacional seja 40 semanas, qual é a probabilidade de que seu peso ao nascer seja menor do que 2.500 gramas?
 - Que valor limita os 5% inferiores da distribuição de pesos ao nascer?
 - Descreva a distribuição das médias das amostras de tamanho 5 extraídas dessa população. Liste três propriedades.
 - Que valor limita os 5% inferiores da distribuição das médias de amostras de tamanho 5?
 - Dada uma amostra de cinco recém-nascidos com idade gestacional de 40 semanas, qual é a probabilidade de que sua média de peso ao nascer seja menor do que 2.500 gramas?
 - Qual a probabilidade de que somente um dos cinco recém-nascidos tenha um peso ao nascer menor que 2.500 gramas?
12. Para a população de mulheres entre 3 e 74 anos que participaram do Levantamento de Entrevistas de Saúde Nacional, a distribuição de níveis de hemoglobina tem média $\mu = 13,3 \text{ g}/100 \text{ ml}$ e desvio-padrão $\sigma = 1,12 \text{ g}/100 \text{ ml}$ [8].
- Se amostras repetidas de tamanho 15 são selecionadas dessa população, que proporção das amostras terá um nível médio de hemoglobina entre 13,0 e 13,6 $\text{g}/100 \text{ ml}$?
 - Se as amostras repetidas são de tamanho 30, que proporção terá uma média entre 13,0 e 13,6 $\text{g}/100 \text{ ml}$?
 - De que tamanho precisam ser as amostras para que 95% de suas médias se encontrem a $\pm 0,2 \text{ g}/100 \text{ ml}$ da média μ da população?
 - De que tamanho precisam ser as amostras para que 95% de suas médias se encontrem a $\pm 0,1 \text{ g}/100 \text{ ml}$ da média da população?
13. Nos Países Baixos, os homens saudáveis entre 65 e 79 anos têm uma distribuição de níveis séricos de ácido úrico aproximadamente normal com média $\mu = 341 \mu\text{mol/l}$ e desvio-padrão $\sigma = 79 \mu\text{mol/l}$ [9].
- Que proporção de homens tem nível sérico de ácido úrico entre 300 e 400 $\mu\text{mol/l}$?
 - Que proporção de amostras de tamanho 5 tem nível médio sérico de ácido úrico entre 300 e 400 $\mu\text{mol/l}$?
 - Que proporção de amostras de tamanho 10 tem nível médio sérico de ácido úrico entre 300 e 400 $\mu\text{mol/l}$?
 - Construa um intervalo que contenha 95% das médias de amostras de tamanho 10. Qual seria menor, um intervalo simétrico ou assimétrico?

14. Para a população de homens adultos nos Estados Unidos, a distribuição de pesos é aproximadamente normal, com $\mu = 172,2$ libras e desvio-padrão $\sigma = 29,8$ libras [10].
- Descreva a distribuição das médias das amostras de tamanho 25 extraídas dessa população.
 - Qual é o limite superior para 90% dos pesos médios das amostras de tamanho 25?
 - Qual é o limite inferior para 80% dos pesos médios?
 - Suponha que você selecione uma amostra aleatória simples de tamanho 25 e verifique que o peso médio para os homens na amostra seja $\bar{x} = 190$ libras. Esse resultado é provável? O que você conclui?
15. No final da Seção 8.3, notou-se que, para as amostras de níveis séricos de colesterol de tamanho 25 — extraídas da população com média $\mu = 211$ mg/100 ml e desvio-padrão $\sigma = 46$ mg/100 ml — a probabilidade de que uma média de amostra \bar{x} se encontre dentro do intervalo (193,0, 229,0) é 0,95. Além disso, a probabilidade de que a média se encontre abaixo de 226,1 mg/100 ml é 0,95 e a de que esteja acima de 195,9 mg/100 ml é 0,95. Para que os três eventos aconteçam simultaneamente, a média da amostra \bar{x} teria que se encontrar no intervalo (195,9, 226,1). Qual a probabilidade dessa ocorrência?

Bibliografia

- [1] LINDGREN, B. W. *Statistical Theory*. New York: Macmillan, 1976.
- [2] SNEDECOR, G. W. e COCHRAN, W. G. *Statistical Methods*. Ames, IA: Iowa State University Press, 1980.
- [3] National Center for Health Statistics. FULWOOD, R., KALSBECK, W., RIFKIND, B., RUSSELL-BRIEFEL, R., MUESING, R., LAROSA, J. e LIPPEL, K. "Total Serum Cholesterol Levels of Adults 20–74 Years of Age: United States, 1976–1980". *Vital and Health Statistics*. série 11, n. 236, maio 1986.
- [4] National Center for Health Statistics. *United States Decennial Life Tables for 1979–1981*. v. I, n. 1, ago. 1985.
- [5] SCULLY, R. E., MCNEELY, B. U. e MARK, E. J. "Case Record of the Massachusetts General Hospital: Weekly Clinicopathological Exercises". *The New England Journal of Medicine*. v. 314, 2 jan., 1986. p. 39–49.
- [6] OSTRO, B. D., LIPSETT, M. J., WIENER, M. B. e Selner, J. C. "Asthmatic Responses to Airborne Acid Aerosols". *American Journal of Public Health*. v. 81, jun. 1991. p. 694–702.
- [7] WILCOX, A. J. e SKJAERVEN, R. "Birth Weight and Perinatal Mortality: The Effects of Gestational Age". *American Journal of Public Health*. v. 82, mar. 1992. p. 378–382.
- [8] National Center for Health Statistics. FULWOOD, R., JOHNSON, C. L., BRYNER, J. D., GUNTER, E. W. e MCGRATH, C. R. "Hematological and Nutritional Biochemistry Reference Data for Persons 6 Months–74 Years of Age: United States, 1976–1980". *Vital and Health Statistics*. série 11, n. 232, dez. 1982.
- [9] LOENEN, H. M. J. A., ESHUIS, H., LOWIK, M. R. H., SCHOUTEN, E. G., HULSHOF, K. F. A. M., ODINK, J. e Kok, F. J. "Serum Uric Acid Correlates in Elderly Men and Women with Special Reference to Body Composition and Dietary Intake (Dutch Nutrition Surveillance System)". *Journal of Clinical Epidemiology*. v. 43, n. 12, 1990. p. 1297–1303.
- [10] National Center for Health Statistics. NAJJAR, M. F. e ROWLAND, M. "Anthropometric Reference Data and Prevalence of Overweight: United States, 1976–1980". *Vital and Health Statistics*. série 11, n. 238, out. 1987.

9

Intervalos de Confiança

Agora que investigamos as propriedades teóricas de uma distribuição amostral da média, estamos prontos para realizar a próxima etapa e aplicar esse conhecimento ao processo de inferência estatística. Lembre-se de que nosso objetivo é descrever ou estimar alguma característica de uma variável aleatória contínua — tal como sua média —, com o uso da informação contida em uma amostra de observações.

Dois métodos de estimação são comumente usados. O primeiro é chamado de *estimação por ponto* e usa os dados da amostra para calcular um único número para estimar o parâmetro de interesse. Por exemplo, poderíamos usar a média da amostra para estimar a média μ da população. O problema é que de duas amostras diferentes é muito provável que os resultados das suas médias não sejam iguais, havendo um grau de incerteza envolvido. Uma estimativa por ponto não fornece nenhuma informação sobre a variabilidade inerente do estimador; não sabemos quão perto \bar{x} está de μ em nenhuma situação. Enquanto é mais provável que \bar{x} esteja mais perto da média verdadeira da população se a amostra em que ela se baseia é grande — lembre-se da propriedade de consistência —, uma estimativa por ponto não fornece informação sobre o tamanho dessa amostra. Conseqüentemente, um segundo método de estimação, conhecido como *estimação por intervalo*, é com freqüência preferido. Essa técnica fornece um intervalo de valores razoável no qual se presume que esteja o parâmetro de interesse — a média μ da população, nesse caso — com certo grau de confiança. Esse intervalo de valores é chamado *intervalo de confiança*.

9.1 Intervalos de Confiança Bilaterais

Para construir um intervalo de confiança para μ , recorreremos ao nosso conhecimento de distribuição amostral da média do capítulo anterior. Dada uma variável aleatória X que tem média μ e desvio-padrão σ , o teorema central do limite estabelece que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tem uma distribuição normal padrão se X for ela própria normalmente distribuída e uma distribuição normal padrão aproximada se ela não for, somente se n for suficientemente grande. Para uma variável aleatória normal padrão, 95% das observações se encontram entre -1,96 e 1,96. Em outras palavras, a probabilidade de que Z assuma um valor entre -1,96 e 1,96 é

$$P(-1,96 \leq Z \leq 1,96) = 0,95.$$

Equivalentemente, poderíamos substituir Z pela quantidade $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ e escrever

$$P\left(-1,96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96\right) = 0,95.$$

Dada essa expressão, podemos manipular a desigualdade dentro do parênteses sem alterar a afirmação da probabilidade. Multiplicamos os três termos da desigualdade pelo erro-padrão σ/\sqrt{n} , logo,

$$P\left(-1,96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

A seguir, subtraímos \bar{X} de cada termo, de modo que

$$P\left(-1,96 \frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq 1,96 \frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 0,95.$$

Finalmente, multiplicamos tudo por -1 . Tenha em mente que multiplicar uma desigualdade por um número negativo inverte a direção da desigualdade. Conseqüentemente,

$$P\left(1,96 \frac{\sigma}{\sqrt{n}} + \bar{X} \geq \mu \geq -1,96 \frac{\sigma}{\sqrt{n}} + \bar{X}\right) = 0,95$$

e, ao rearanjarmos os termos,

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

Note-se que \bar{X} não está mais no centro da desigualdade; ao contrário disso, a afirmação da probabilidade se refere a μ . As quantidades $\bar{X} - 1,96(\sigma/\sqrt{n})$ e $\bar{X} + 1,96(\sigma/\sqrt{n})$ são os limites de confiança de 95% para a média da população; estamos 95% confiantes de que o intervalo

$$\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right)$$

conterá μ . Essa afirmação **não** implica que μ seja uma variável aleatória que assume um valor no intervalo 95% das vezes, nem que 95% dos valores da população se encontrem nesses limites; ao contrário, ela significa que, se selecionássemos 100 amostras aleatórias da população e as usássemos para calcular 100 intervalos de confiança diferentes para μ , aproximadamente 95 dos intervalos conteriam a média verdadeira da população e 5 não.

Saiba que o estimador \bar{X} é variável aleatória, enquanto o parâmetro μ é uma constante. Portanto, o intervalo

$$\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right)$$

é aleatório e tem 95% de probabilidade de conter μ **antes que** uma amostra seja selecionada. Desde que μ tenha valor fixo, uma vez que uma amostra tenha sido extraída e os limites de confiança

$$\left(\bar{x} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}}\right)$$

calculados, μ pode estar dentro do intervalo ou não. Não há mais qualquer probabilidade envolvida.

Embora um intervalo de confiança de 95% seja usado mais freqüentemente na prática, não estamos restritos a essa escolha. Poderíamos ter maior grau de certeza com relação ao valor da média da população; nesse caso, construiríamos um intervalo de confiança de 99% e não um intervalo de 95%. Desde que 99% das observações em uma distribuição normal padrão se encontrem entre $-2,58$ e $2,58$, um intervalo de confiança para μ é

$$\left(\bar{X} - 2,58 \frac{\sigma}{\sqrt{n}}, \bar{X} + 2,58 \frac{\sigma}{\sqrt{n}} \right).$$

Aproximadamente 99 dentre 100 intervalos de confiança, obtidos de 100 amostras aleatórias, independentes de tamanho n extraídas dessa população, conteriam a média verdadeira μ . Como esperávamos, o intervalo de confiança de 99% é maior do que o intervalo de 95%; quanto menor o intervalo de valores que consideramos, menos confiantes estaremos de que o intervalo contenha μ .

Um intervalo de confiança genérico para μ pode ser obtido com a introdução de uma nova notação. Seja $z_{\alpha/2}$ o valor que limita uma área de $\alpha/2$ na extremidade superior da distribuição normal padrão e $-z_{\alpha/2}$ o valor que limita uma área de $\alpha/2$ na extremidade inferior da distribuição. Se $\alpha = 0,05$, por exemplo, então $z_{0,05/2} = 1,96$ e $-z_{0,05/2} = -1,96$. Então, a forma geral para um intervalo de confiança de $100\% \times (1 - \alpha)$ para μ — um intervalo de confiança de 95% se $\alpha = 0,05$ — é

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

Esse intervalo tem uma probabilidade de $100\% \times (1 - \alpha)$ de conter μ antes que uma amostra aleatória seja selecionada.

Se desejamos tornar um intervalo menor sem reduzir o nível de confiança, necessitamos de mais informação sobre a média da população; assim, precisamos selecionar uma amostra maior. Conforme o tamanho n da amostra aumenta, o erro-padrão σ/\sqrt{n} diminui, o que resultar em um intervalo de confiança mais estreito. Considere os limites de confiança de 95% de $\bar{X} \pm 1,96(\sigma/\sqrt{n})$. Se escolhermos uma amostra de tamanho 10, os limites de confiança serão $\bar{X} \pm 1,96(\sigma/\sqrt{10})$. Se a amostra selecionada for de tamanho 100, então os limites serão $\bar{X} \pm 1,96(\sigma/\sqrt{100})$. Para uma amostra ainda maior, de tamanho 1.000, os limites de confiança de 95% seriam $\bar{X} \pm 1,96(\sigma/\sqrt{1.000})$. Ao resumirmos esses cálculos, temos:

n	Limites de Confiança de 95% para μ	Comprimento do Intervalo
10	$\bar{X} \pm 0,620\sigma$	$1,240\sigma$
100	$\bar{X} \pm 0,196\sigma$	$0,392\sigma$
1.000	$\bar{X} \pm 0,062\sigma$	$0,124\sigma$

Conforme selecionamos amostras cada vez maiores, a variabilidade de \bar{X} — nosso estimador da média μ da população — se torna menor. Entretanto, a variabilidade inerente da população original, medida por σ , está sempre presente.

Considere a distribuição de níveis séricos de colesterol para todos os homens hipertensos e fumantes nos Estados Unidos. Essa distribuição é aproximadamente normal com uma média μ desconhecida e desvio-padrão $\sigma = 46$ mg/100 ml. (Mesmo que a média seja dife-

rente, assumimos, por ora, que σ é igual ao que era para a população geral dos homens adultos que vivem nos Estados Unidos.) Estamos interessados em estimar o nível médio sérico de colesterol dessa população. Antes que selecionemos uma amostra aleatória, a probabilidade de que o intervalo

$$\left(\bar{X} - 1,96 \frac{46}{\sqrt{n}}, \bar{X} + 1,96 \frac{46}{\sqrt{n}} \right)$$

contenha a média μ verdadeira da população é 0,95.

Suponha que extraímos uma amostra de tamanho 12 da população de fumantes hipertensos e que esses homens tenham um nível médio sérico de colesterol de $\bar{x} = 217$ mg/100 ml [1]. Com base nessa amostra, um intervalo de confiança de 95% para μ é

$$\left(217 - 1,96 \frac{46}{\sqrt{12}}, 217 + 1,96 \frac{46}{\sqrt{12}} \right)$$

ou

$$(191, 243).$$

Enquanto 217 mg/100 ml for nosso melhor palpite para o nível médio sérico de colesterol da população de homens fumantes e hipertensos, o intervalo de 191 a 243 fornece um intervalo de valores razoáveis para μ . Note que ele contém o valor 211 mg/100 ml, o nível médio de colesterol para todos os homens de 20 a 74 anos nos Estados Unidos, independentemente do *status* de hipertensão ou de fumante [2]. Estamos 95% confiantes de que os limites de 191 e 243 contêm a média verdadeira μ . Não dizemos que há uma probabilidade de 95% de que μ se encontre entre esses valores; μ é fixo e pode estar entre 191 e 243 ou não.

Como mencionamos anteriormente, esse intervalo de confiança também tem uma interpretação freqüentista. Suponha que o nível médio sérico verdadeiro de colesterol da população de homens fumantes e hipertensos seja igual a 211 mg/100 ml, o nível médio para os homens adultos nos Estados Unidos. Se extraíssemos 100 amostras aleatórias de tamanho 12 dessa população e usássemos cada uma delas para construir um intervalo de confiança de 95%, esperaríamos que, na média, 95 dos intervalos conteriam a média $\mu = 211$ verdadeira da população e 5 não. Esse procedimento foi simulado e os resultados são ilustrados na Figura 9.1. A única quantidade que varia de amostra para amostra é \bar{X} . Embora os centros dos intervalos sejam diferentes, todos têm o mesmo comprimento. Cada um dos intervalos de confiança que não contém o verdadeiro valor de μ está marcado por um ponto; note-se que exatamente cinco intervalos ficam nessa categoria.

Em vez de gerar um intervalo de confiança de 95% para o nível médio sérico de colesterol, poderíamos calcular um intervalo de confiança de 99% para μ . Ao usarmos a mesma amostra de 12 fumantes e hipertensos, encontramos os limites como sendo

$$\left(217 - 2,58 \frac{46}{\sqrt{12}}, 217 + 2,58 \frac{46}{\sqrt{12}} \right),$$

ou

$$(183, 251).$$

Estamos 99% confiantes de que esse intervalo contém o nível médio sérico verdadeiro de colesterol da população. Como observado anteriormente, esse intervalo é maior do que o correspondente intervalo de confiança de 95%.

No exemplo anterior, o comprimento do intervalo de confiança de 99% é 251 – 183 = 68 mg/100 ml. Qual o tamanho da amostra que precisamos para reduzir o comprimento desse

**FIGURA 9.1**

Conjunto de intervalos de confiança de 95% construídos de amostras de tamanho 12 extraídas de uma população normal com média 211 (marcada pela linha vertical) e desvio-padrão 46.

intervalo a somente 20 mg/100 ml? Uma vez que o intervalo está centrado ao redor da média da amostra $\bar{x} = 217$ mg/100 ml, estamos interessados no tamanho de amostra necessário para produzir o intervalo

$$(217 - 10, 217 + 10)$$

ou

$$(207, 227).$$

Lembre-se de que o intervalo de confiança de 99% é da forma

$$\left(217 - 2,58 \frac{46}{\sqrt{n}}, 217 + 2,58 \frac{46}{\sqrt{n}} \right).$$

Portanto, para encontrar o tamanho n da amostra requerido, precisamos resolver a equação

$$10 = \frac{2,58(46)}{\sqrt{n}}.$$

Ao multiplicarmos ambos os lados da igualdade por \sqrt{n} e dividirmos por 10, encontramos

$$\sqrt{n} = \frac{2,58(46)}{10}$$

e

$$n = 140,8.$$

Necessitáramos de uma amostra de 141 homens para reduzir o comprimento do intervalo de confiança de 99% para 20 mg/100 ml. Embora a média de amostra de 217 mg/100 ml se encontre no centro do intervalo, não toma parte na determinação do comprimento, pois ele é uma função de σ , n e do nível de confiança.

9.2 Intervalos de Confiança Unilaterais

Em algumas situações, preocupamo-nos com um limite superior para a média μ da população ou com um limite inferior para μ , mas não com ambos. Considere a distribuição de níveis de hemoglobina — proteína condutora de oxigênio encontrada nas células vermelhas do sangue — para a população de crianças com até seis anos, expostas a elevados níveis de chumbo. Essa distribuição tem média μ desconhecida e desvio-padrão $\sigma = 0,85$ g/100 ml [3]. Sabemos que crianças envenenadas por chumbo tendem a ter níveis muito mais baixos de hemoglobina do que as que não o foram. Logo, poderíamos encontrar um limite superior para μ .

Para construir um intervalo de confiança unilateral, consideramos somente a área em uma extremidade da distribuição normal padrão. Consultando a Tabela A.3, verificamos que 95% das observações para uma variável aleatória normal padrão se encontram acima de $z = -1,645$. Logo,

$$P(Z \geq -1,645) = 0,95.$$

Ao substituirmos Z por $(\bar{X} - \mu)/(\sigma/\sqrt{n})$,

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq -1,645\right) = 0,95.$$

Ao multiplicarmos ambos os lados da desigualdade dentro da afirmação de probabilidade por σ/\sqrt{n} e subtrairmos \bar{X} , verificamos que

$$P\left(-\mu \geq -\bar{X} - 1,645 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

e

$$P\left(\mu \leq \bar{X} + 1,645 \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

Portanto, $\bar{X} + 1,645(\sigma/\sqrt{n})$ é um limite de confiança superior de 95% para μ . Analogamente, poderíamos mostrar que $\bar{X} - 1,645(\sigma/\sqrt{n})$ é o correspondente limite de confiança inferior de 95%.

Suponha que selecionemos uma amostra de 74 crianças expostas a elevados níveis de chumbo, as quais têm um nível médio de hemoglobina de $\bar{x} = 10,6$ g/100 ml [4]. Com base nessa amostra, um intervalo de confiança unilateral de 95% para μ — o limite superior somente — é

$$\begin{aligned}\mu &\leq 10,6 + 1,645 \left(\frac{0,85}{\sqrt{74}} \right) \\ &\leq 10,8.\end{aligned}$$

Estamos 95% confiantes de que a média verdadeira de nível de hemoglobina para essa população de crianças é no máximo 10,8 g/100 ml. Na realidade, uma vez que o valor de μ é fixo, a média verdadeira é menor do que 10,8 ou não. No entanto, se selecionássemos 100 amostras aleatórias de tamanho 74 e usássemos cada uma delas para construir um intervalo de confiança unilateral de 95%, aproximadamente 95 de tais intervalos conteriam a média verdadeira μ .

9.3 Distribuição t de Student

Até aqui, assumimos que σ , o desvio-padrão da população, é conhecido, quando calculamos os intervalos de confiança para uma média μ da população desconhecida. Na realidade, é improvável que esse seja o caso. Se μ for desconhecido, σ provavelmente também o será. Nessa situação, os intervalos de confiança são calculados do mesmo modo que já vimos. Em vez de usarmos a distribuição normal padrão, a análise depende da distribuição de probabilidade conhecida como distribuição t de Student — pseudônimo do estatístico que descobriu essa distribuição.

Para construir um intervalo de confiança bilateral para uma média μ da população, notamos que

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

tem uma distribuição normal padrão aproximada, se n é suficientemente grande. Quando o desvio-padrão da população não é conhecido, parece lógico substituir σ por s , o desvio-padrão de uma amostra extraída da população. E é, de fato, o que tem sido feito. No entanto, a razão

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

não tem distribuição normal padrão. Além da variabilidade amostral inerente de \bar{X} — a qual está sendo usada como um estimador da média μ da população — há também a variação de s . É provável que o valor de s varie de amostra para amostra. Portanto, precisamos levar em conta o fato de que s pode não ser uma estimativa confiável de σ , especialmente se a amostra com a qual trabalhamos for pequena.

Se X é normalmente distribuída e a amostra de tamanho n é aleatoriamente escolhida dessa população original, a distribuição de probabilidade da variável aleatória

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

é conhecida como *distribuição t de Student* com $n - 1$ graus de liberdade. Fazemos essa representação com a notação t_{n-1} . Tal como a distribuição normal padrão, a distribuição t é unimodal e simétrica ao redor de sua média 0. A área total sob a curva é igual a 1. No entanto, ela tem extremidades mais densas do que a distribuição normal; valores extremos são mais prováveis de ocorrer com a distribuição t do que com a normal padrão. Essa diferença está ilustrada na Figura 9.2. A forma da distribuição t reflete a variabilidade extra introduzida pelo estimador s . Além disso, a distribuição t tem uma propriedade chamada *graus de liberdade*, abreviada *gl* (em inglês, *df*, *degrees of freedom*). Os graus de liberdade medem o volume de informação disponível nos dados que podem ser usados para estimar σ^2 ; por isso, medem a confiabilidade de s^2 como um estimador de σ^2 . (Os graus de liberdade são

$n - 1$ e não n , porque perdemos 1 gl, ao estimarmos a média da amostra.) Lembre-se de que $gl = n - 1$ é a quantidade pela qual dividimos a soma dos desvios da média ao quadrado, $\sum_{i=1}^n (x_i - \bar{x})^2$, a fim de obtermos a variância da amostra.

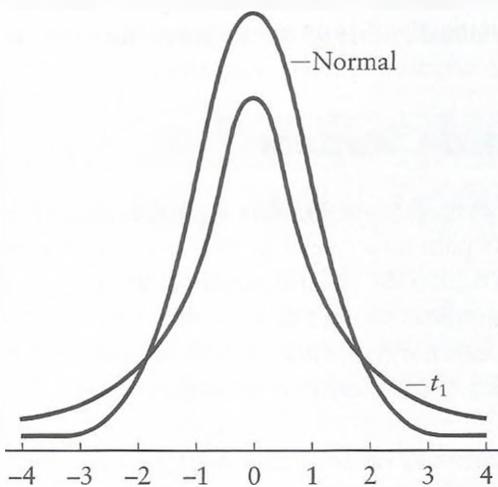


FIGURA 9.2

A distribuição normal padrão e a distribuição t de Student com 1 grau de liberdade.

Para cada possível valor dos graus de liberdade, há uma diferente distribuição t . As distribuições com menores graus de liberdade são mais dispersas; conforme gl aumenta, a distribuição t se aproxima da distribuição normal padrão. Isso ocorre porque, conforme o tamanho da amostra aumenta, s torna-se uma estimativa mais confiável de σ ; se n é muito grande, conhecer o valor de s é quase equivalente a conhecer σ .

Como há uma diferente distribuição t para cada valor dos graus de liberdade, seria bastante inconveniente trabalhar com uma tabela completa de áreas correspondente a cada uma. Contamos com um programa de computador ou com uma tabela condensada que lista as áreas sob a curva somente para percentis selecionados da distribuição que, por exemplo, poderia conter os 5,0, 2,5, 1,0, 0,5 e 0,05% superiores da distribuição. Quando não houver um computador disponível, as tabelas condensadas são suficientes para a maioria das aplicações que envolve a construção dos intervalos de confiança.

A Tabela A.4 do Apêndice A é condensada por áreas calculadas para a família de distribuições t . Para um valor particular de gl, a entrada na tabela representa o valor de t_{n-1} que limita a área especificada na extremidade superior da distribuição. Dada uma distribuição t com 10 graus de liberdade, por exemplo, $t_{10} = 2,228$ limita os 2,5% superiores da área sob a curva. Uma vez que a distribuição é simétrica, $t_{10} = -2,228$ limita os 2,5% inferiores. Os valores de t_{n-1} que limitam os 2,5% superiores das distribuições com vários graus de liberdade estão listados abaixo.

$gl(n - 1)$	t_{n-1}
1	12,706
2	4,303
5	2,571
10	2,228
20	2,086
30	2,042
40	2,021
60	2,000
120	1,980
∞	1,960

Para a curva normal padrão, $z = 1,96$ limita os 2,5% superiores da distribuição. Observe que, conforme n aumenta, t_{n-1} se aproxima desse valor. De fato, quando temos mais do que 30 graus de liberdade, somos capazes de substituir a distribuição t pela distribuição normal padrão e errarmos nossos cálculos por menos de 5%.

Considere uma amostra aleatória de dez crianças selecionadas da população de bebês que recebe antiácidos que contêm alumínio e são freqüentemente usados para tratar desarranjos pépticos e digestivos. A distribuição de níveis de alumínio no plasma é conhecida como aproximadamente normal; no entanto, sua média μ e seu desvio-padrão σ não são conhecidos. O nível médio de alumínio para a amostra de dez bebês é $\bar{x} = 37,2 \text{ } \mu\text{g/l}$ e seu desvio padrão é $s = 7,13 \text{ } \mu\text{g/l}$ [5].

Uma vez que o desvio padrão da população não é conhecido, precisamos usar a distribuição t para encontrar os limites de confiança de 95% para μ . Para uma distribuição com $10 - 1 = 9$ graus de liberdade, 95% das observações se encontram entre -2,262 e 2,262. Então, substituindo σ por s , um intervalo de confiança de 95% para a média μ da população é

$$\left(\bar{x} - 2,262 \frac{s}{\sqrt{n}}, \bar{x} + 2,262 \frac{s}{\sqrt{n}} \right).$$

Ao substituirmos os valores de \bar{x} e de s , o intervalo se torna

$$\left(37,2 - 2,262 \frac{7,13}{\sqrt{10}}, 37,2 + 2,262 \frac{7,13}{\sqrt{10}} \right),$$

ou

$$(32,1, 42,3).$$

Estamos 95% confiantes de que esses limites contenham o nível médio verdadeiro de alumínio no plasma para a população de bebês que recebe antiácidos. Se fornecermos a informação adicional de que o nível médio de alumínio no plasma para essa população é $4,3 \text{ } \mu\text{g/l}$ — um valor não-plausível de μ para bebês que os recebem, de acordo com o intervalo de confiança de 95% — estaremos sugerindo que dar antiácidos aumenta em muito os níveis de alumínio no plasma das crianças.

Se o desvio-padrão σ da população fosse conhecido e igual ao valor da amostra de $7,13 \text{ } \mu\text{g/l}$, o intervalo de confiança de 95% para μ seria

$$\left(37,2 - 1,96 \frac{7,13}{\sqrt{10}}, 37,2 + 1,96 \frac{7,13}{\sqrt{10}} \right),$$

ou

$$(32,8, 41,6).$$

Nesse caso, o intervalo de confiança é levemente menor. Na maioria das vezes, os intervalos de confiança com base na distribuição t são maiores do que os correspondentes fundamentados na distribuição normal padrão. Entretanto, essa generalização nem sempre se aplica; devido à natureza da variabilidade amostral, é possível que o valor da estimativa s seja consideravelmente menor do que σ para uma determinada amostra.

No exemplo anterior, examinamos a distribuição dos níveis séricos de colesterol para todos os homens hipertensos e fumantes nos Estados Unidos. Lembre-se de que o desvio-padrão dessa população foi assumido em 46 mg/100 ml . Do lado esquerdo, a Figura 9.3 mos-

tra os intervalos de confiança de 95% para μ , calculados a partir de 100 amostras aleatórias e exibidos na Figura 9.1. O lado direito da figura mostra 100 intervalos adicionais calculados com as mesmas amostras; em cada caso, no entanto, o desvio-padrão não foi assumido como conhecido. Uma vez mais, 95 dos intervalos contêm a média μ verdadeira e os outros 5 não. Note-se que agora, porém, os intervalos variam no comprimento.

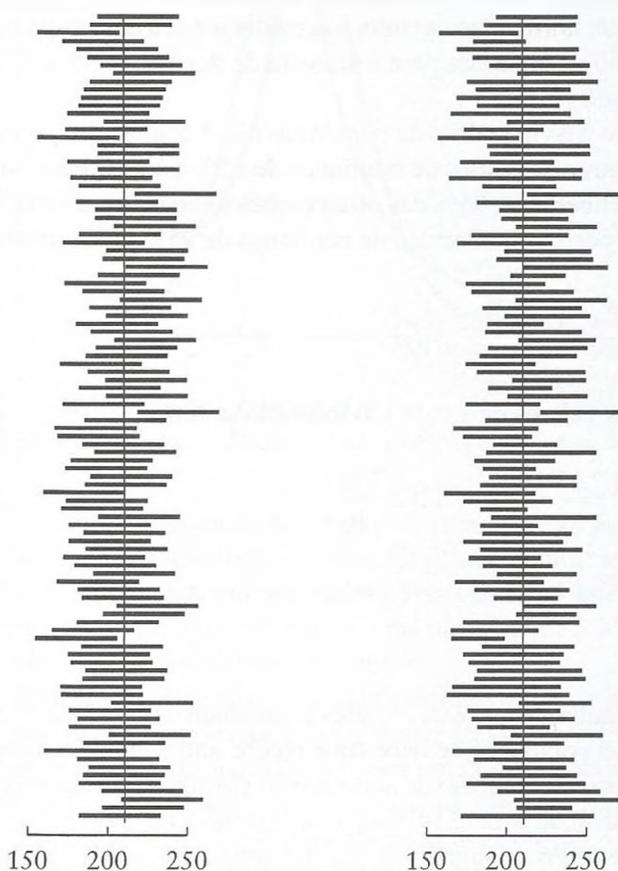


FIGURA 9.3

Dois conjuntos de intervalos de confiança de 95%, construídos a partir de amostras de tamanho 12 extraídas de populações normais com média 211 (marcada pelas linhas verticais), um com desvio-padrão 46 e o outro com desvio-padrão desconhecido.

9.4 Aplicações Adicionais

Considere a distribuição de alturas para a população de indivíduos com idades entre 12 e 40 anos que sofrem de síndrome alcoólica fetal — final severo do espectro de deficiências causadas pelo uso materno de álcool durante a gravidez. A distribuição de alturas é aproximadamente normal com média μ desconhecida. Desejamos encontrar uma estimativa por ponto e um intervalo de confiança para μ , o qual fornece uma série de valores razoáveis para o parâmetro de interesse.

Quando construímos um intervalo de confiança para a média de uma variável aleatória contínua, a técnica usada pode diferir, se o desvio-padrão da população original for ou não conhecido. Para os dados de altura, o desvio-padrão é assumido como $\sigma = 6$ centímetros

[6]. Então, usamos a distribuição normal padrão para nos auxiliar a construir um intervalo de confiança de 95%. Antes que uma amostra seja extraída dessa população, o intervalo

$$\left(\bar{X} - 1,96 \frac{6}{\sqrt{n}}, \bar{X} + 1,96 \frac{6}{\sqrt{n}} \right)$$

tem uma probabilidade de 95% de conter a verdadeira média μ da população.

Uma amostra aleatória de 31 pacientes é selecionada da população original; a altura média desses indivíduos é $\bar{x} = 147,4$ cm. Essa é a estimativa por ponto — nosso melhor palpite — para a média μ da população. Um intervalo de confiança de 95% baseado na amostra é

$$\left(147,4 - 1,96 \frac{6}{\sqrt{31}}, 147,4 + 1,96 \frac{6}{\sqrt{31}} \right),$$

ou

$$(145,3, 149,5).$$

Estamos 95% confiantes que esses limites contêm a altura média verdadeira da população dos indivíduos entre 12 e 40 anos que sofrem de síndrome alcoólica fetal. Na realidade, o valor fixo de μ está entre 145,3 cm e 149,5 cm ou não.

Em vez de gerar o intervalo de confiança manualmente, poderíamos usar um computador para fazer os cálculos. A Tabela 9.1 mostra a saída relevante do Minitab. Além do tamanho da amostra, a tabela exibe a média da amostra, o desvio-padrão assumido, o erro-padrão da média e o intervalo de confiança de 95%. É também possível calcular intervalos com diferentes níveis de confiança. A Tabela 9.2 mostra um intervalo de confiança de 90% para μ . O intervalo de confiança de 90% é mais curto do que o intervalo de 95%; temos menos confiança de que esse intervalo contenha a média verdadeira μ .

Como segundo exemplo, o metilfenidato é uma droga amplamente usada no tratamento do distúrbio de déficit de atenção. Como parte de um *estudo crossover*, dez crianças entre 7 e 12 anos que sofreram desse distúrbio foram designadas para receber a droga e dez receberam placebo [7]. Depois de um certo período, o tratamento foi suspenso para as 20 crianças. Subseqüentemente, às crianças que receberam o metilfenidato foi dado o placebo e as que tinham recebido o placebo agora receberam o medicamento (isso é o que se entende por *estudo crossover*). Medidas da atenção e do *status comportamental* de cada criança, tanto no medicamento como no placebo, foram obtidas por meio de um instrumento chamado “Parent Rating Scale”. As distribuições desses escores são aproximadamente normais

TABELA 9.1

Saída do Minitab exibindo um intervalo de confiança de 95%, desvio-padrão conhecido.

THE ASSUMED SIGMA = 6.000

	N	MEAN	STDEV	SE MEAN	95.0 PERCENT C.I.
HEIGHT	31	147.4	6.000	1.078	(145.288, 149.512)

TABELA 9.2

Saída do Minitab exibindo um intervalo de confiança de 90%, desvio-padrão conhecido.

THE ASSUMED SIGMA = 6.000

	N	MEAN	STDEV	SE MEAN	90.0 PERCENT C.I.
HEIGHT	31	147.4	6.000	1.078	(145.627, 149.173)

com médias e desvios-padrões desconhecidos. Em geral, escores baixos indicam aumento na atenção. Desejamos estimar o escore médio da avaliação de atenção para as crianças que tomam metilfenidato e para as que tomam placebo.

Uma vez que o desvio-padrão não é conhecido para nenhuma das populações, usamos a distribuição t para nos auxiliar a construir os intervalos de confiança de 95%. Para a distribuição t com $20 - 1 = 19$ graus de liberdade, 95% das observações se encontram entre $-2,093$ e $2,093$. Portanto, antes que uma amostra de tamanho 20 seja extraída da população, o intervalo

$$\left(\bar{X} - 2,093 \frac{s}{\sqrt{20}}, \bar{X} + 2,093 \frac{s}{\sqrt{20}} \right)$$

tem uma probabilidade de 95% de conter a média μ verdadeira.

A amostra aleatória de 20 crianças arroladas no estudo tem escore médio da avaliação da atenção $\bar{x}_m = 10,8$ e desvio padrão $s_m = 2,9$ quando tomado metilfenidato e $\bar{x}_p = 14,0$ e $s_p = 4,8$ quando tomado placebo. Logo, em um intervalo de confiança de 95% para μ_m , o escore médio da avaliação da atenção para crianças que tomam o medicamento é

$$\left(10,8 - 2,093 \frac{2,9}{\sqrt{20}}, 10,8 + 2,093 \frac{2,9}{\sqrt{20}} \right),$$

ou

$$(9,44, 12,16).$$

e um intervalo de confiança de 95% para μ_p , o escore médio da avaliação da atenção para crianças que tomam o placebo é

$$\left(14,0 - 2,093 \frac{4,8}{\sqrt{20}}, 14,0 + 2,093 \frac{4,8}{\sqrt{20}} \right),$$

ou

$$(11,75, 16,25).$$

A saída relevante do Stata para esses dois intervalos está na Tabela 9.3. Quando observamos os intervalos, parece que o escore médio da avaliação da atenção é mais baixo quando as crianças com distúrbio de déficit de atenção tomam metilfenidato, o que implica melhora da atenção. No entanto, há certa sobreposição entre os dois intervalos.

TABELA 9.3

Saída do Stata exibindo intervalos de confiança de 95%, desvio-padrão desconhecido.

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
rating	20	10.8	.6484597	9.442758 12.15724

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
rating	20	14	1.073313	11.75353 16.24647

9.5 Exercícios de Revisão

1. Explique a diferença entre estimativa por ponto e estimativa por intervalo.
2. Descreva o intervalo de confiança de 95% para uma média μ da população. Como o intervalo é interpretado?
3. Quais os fatores que afetam o comprimento de um intervalo de confiança para uma média? Explique brevemente.
4. Descreva as similaridades e diferenças entre a distribuição t e a distribuição normal padrão. Se você tentasse construir um intervalo de confiança, quando usaria uma em lugar da outra?
5. As distribuições das pressões sanguíneas sistólicas e diastólicas para mulheres diabéticas entre 30 e 34 anos têm médias desconhecidas. No entanto, seus desvios-padrão são $\sigma = 11,8$ mm Hg e $\sigma = 9,1$ mm Hg, respectivamente [8].
 - (a) Uma amostra aleatória de dez mulheres é selecionada dessa população. A pressão sanguínea sistólica média para a amostra é $\bar{x} = 130$ mm Hg. Calcule um intervalo de confiança bilateral de 95% para μ , a verdadeira pressão sanguínea sistólica média.
 - (b) Interprete esse intervalo de confiança.
 - (c) A pressão sanguínea diastólica média para a amostra de tamanho 10 é $\bar{x} = 84$ mm Hg. Encontre um intervalo de confiança bilateral de 90% para μ , a verdadeira pressão sanguínea diastólica média da população.
 - (d) Calcule um intervalo de confiança bilateral de 99% para μ .
 - (e) Como o intervalo de confiança de 99% se compara ao intervalo de 90%?
6. Considere a distribuição t com 5 graus de liberdade.
 - (a) Que proporção da área sob a curva se encontra à direita de $t = 2,015$?
 - (b) Que proporção da área se encontra à esquerda de $t = -3,365$?
 - (c) Que proporção da área se encontra entre $t = -4,032$ e $t = 4,032$?
 - (d) Que valor de t limita os 2,5% superiores da distribuição?
7. Considere a distribuição t com 21 graus de liberdade.
 - (a) Que proporção da área sob a curva se encontra à esquerda de $t = -2,518$?
 - (b) Que proporção da área se encontra à direita de $t = 1,323$?
 - (c) Que proporção da área se encontra entre $t = -1,721$ e $t = 2,831$?
 - (d) Que valor de t limita os 2,5% inferiores da distribuição?
8. Antes de começar um estudo que investiga a habilidade do medicamento heparina para evitar broncoconstricção, os valores de base da função pulmonar foram medidos para uma amostra de 12 indivíduos com histórico de asma induzida por exercícios [9]. O valor médio da capacidade vital forçada (FVC, do inglês, *Forced Vital Capacity*) para a amostra é $\bar{x}_1 = 4,49$ litros e o desvio-padrão é $s_1 = 0,83$ litros; a média do volume expiratório forçado em um segundo (FEV₁, do inglês *Forced Expiratory Volume*) é $\bar{x}_2 = 3,71$ litros e o desvio-padrão é $s_2 = 0,62$ litros.
 - (a) Calcule um intervalo de confiança bilateral de 95% para μ_1 , o FVC médio verdadeiro da população.
 - (b) No lugar de um intervalo de confiança de 95%, construa um de 90% para o FVC médio verdadeiro. Como o comprimento do intervalo se modifica?
 - (c) Calcule um intervalo de confiança de 95% para μ_2 , o FEV₁ médio verdadeiro da população.
 - (d) Para construir esses intervalos de confiança, que suposição é feita sobre as distribuições originais de FVC e FEV₁?
9. Para a população de bebês submetidos a cirurgia fetal para anomalias congênitas, a distribuição das idades gestacionais ao nascer é aproximadamente normal com média μ e

desvio-padrão σ desconhecidos. Uma amostra aleatória de 14 desses bebês tem uma idade gestacional média de $\bar{x} = 29,6$ semanas e desvio-padrão de $s = 3,6$ semanas [10].

- Construa um intervalo de confiança de 95% para a verdadeira média μ da população.
- Qual é o comprimento desse intervalo?
- De que tamanho uma amostra deveria ser para que o intervalo de confiança de 95% tenha comprimento de 3 semanas? Assuma que o desvio-padrão σ da população seja conhecido e que $\sigma = 3,6$ semanas.
- De que tamanho uma amostra deveria ser para que o intervalo de confiança de 95% tenha comprimento de 2 semanas?

10. As porcentagens do peso corporal ideal foram determinadas para 18 diabéticos dependentes de insulina aleatoriamente selecionados e são mostradas abaixo [11]. Uma porcentagem de 120 significa que um indivíduo pesa 20% mais do que seu peso corporal ideal; uma porcentagem de 95 significa que o indivíduo pesa 5% a menos do que o ideal.

107	119	99	114	120	104	88	114	124
116	101	121	152	100	125	114	95	117 (%)

- Calcule um intervalo de confiança bilateral de 95% para a verdadeira porcentagem média do peso corporal ideal para a população de diabéticos dependentes de insulina.
- Esse intervalo de confiança contém o valor de 100%? O que a resposta a essa questão diz a você?

11. Quando oito pessoas em Massachusetts sofreram um episódio não explicado de intoxicação de vitamina D que exigiu hospitalização, foi sugerido que essas ocorrências não-usuais poderiam ser resultado de suplementação excessiva de leite [12]. Os níveis de cálcio e de albumina no sangue para cada indivíduo no momento da internação no hospital são exibidos abaixo.

Cálcio (mmol/l)	Albumina (g/l)
2,92	43
3,84	42
2,37	42
2,99	40
2,67	42
3,17	38
3,74	34
3,44	42

- Construa um intervalo de confiança unilateral de 95% — um limite inferior — para o nível médio verdadeiro de cálcio de indivíduos que sofreram a intoxicação de vitamina D.
- Calcule um limite de confiança inferior de 95% para o nível médio verdadeiro de albumina desse grupo.
- Para indivíduos saudáveis, o intervalo normal de valores de cálcio é de 2,12 a 2,74 mmol/l e o intervalo de níveis de albumina é de 32 a 55 g/l. Você acredita que os pacientes que sofrem de intoxicação de vitamina D têm níveis normais de cálcio e de albumina no sangue?

12. Níveis séricos de zinco para 462 homens entre 15 e 17 anos estão salvos sob a variável de nome `zinc` no conjunto de dados `serzinc` [13] (Apêndice B, Tabela B.1). As unidades de medida do nível sérico de zinco são microgramas por decilitro.
- Encontre um intervalo de confiança bilateral de 95% para μ , o nível médio sérico de zinco para essa população de homens.

- (b) Interprete esse intervalo de confiança.
(c) Calcule um intervalo de confiança de 90% para μ .
(d) Como o intervalo de confiança de 90% se compara ao intervalo de 95%?
13. O conjunto de dados `lowbwt` contém informação registrada para uma amostra de 100 bebês com baixo peso ao nascer, nascidos em dois hospitais-escola em Boston, Massachusetts [14] (Apêndice B, Tabela B.7). As medidas da pressão sanguínea sistólica estão salvas sob a variável de nome `sbp`, enquanto indicadores do sexo — onde 1 representa um menino e 0 uma menina — estão salvos sob o nome `sex`.
- (a) Calcule um intervalo de confiança de 95% para a pressão sanguínea sistólica média verdadeira de bebês do sexo masculino com baixo peso ao nascer.
(b) Calcule um intervalo de confiança de 95% para a pressão sanguínea sistólica média verdadeira de bebês do sexo feminino com baixo peso ao nascer.
(c) Você acha que é possível meninos e meninas terem a mesma pressão sanguínea sistólica média? Explique brevemente.

Bibliografia

- [1] KAPLAN, N. M. "Strategies to Reduce Risk Factors in Hypertensive Patients Who Smoke". *American Heart Journal*, v. 115, jan. 1988. p. 288–294.
- [2] National Center for Health Statistics. FULWOOD, R., KALSBECK, W., RIFKIND, B., RUSSELL-BIEFEL, R., MUESING, R., LAROSA, J. e LIPPEL, K. "Total Serum Cholesterol Levels of Adults 20–74 Years of Age: United States, 1976–1980". *Vital and Health Statistics*. série 11, n. 236, maio 1986.
- [3] National Center for Health Statistics. FULWOOD, R., JOHNSON, C. L., BRYNER, J. D., GUNTER, E. W. e MCGRATH, C. R. "Hematological and Nutritional Biochemistry Reference Data for Persons 6 Months–74 Years of Age: United States, 1976–1980". *Vital and Health Statistics*, série 11, n. 232, dez. 1982.
- [4] CLARK, M., ROYAL, J. e SEELER, R. "Interaction of Iron Deficiency and Lead and the Hematologic Findings in Children with Severe Lead Poisoning". *Pediatrics*, v. 81, fev. 1988. p. 247–253.
- [5] TSOU, V. M., YOUNG, R. M., HART, M. H. e VANDERHOOF, J. A. "Elevated Plasma Aluminum Levels in Normal Infants Receiving Antacids Containing Aluminum". *Pediatrics*, v. 87, fev. 1991. p. 148–151.
- [6] STREISSGUTH, A. P., AASE, J. M., CLARREN, S. K., RANDELS, S. P., LADUE, R. A. e SMITH, D. F. "Fetal Alcohol Syndrome in Adolescents and Adults". *Journal of the American Medical Association*, v. 265, 17 abr. 1991. p. 1961–1967.
- [7] TIROSH, E., ELHASID, R., KAMAH, S. C. B. e COHEN, A. "Predictive Value of Placebo Methylphenidate". *Pediatric Neurology*, v. 9, n. 2, 1993. p. 131–133.
- [8] KLEIN, B. E. K., KLEIN, R. e MOSS, S. E. "Blood Pressure in a Population of Diabetic Persons Diagnosed After 30 Years of Age". *American Journal of Public Health*, v. 74, abr. 1984. p. 336–339.
- [9] AHMED, T., GARRIGO, J. e DANTA, I. "Preventing Bronchoconstriction in Exercise-Induced Asthma with Inhaled Heparin". *The New England Journal of Medicine*, v. 329, 8 jul. 1993. p. 90–95.
- [10] LONGAKER, M. T., GOLBUS, M. S., FILLY, R. A., ROSEN, M. A., CHANG, S. W. e HARRISON, M. R. "Maternal Outcome After Open Fetal Surgery". *Journal of the American Medical Association*, v. 265, 13 fev. 1991. p. 737–741.
- [11] SAUDEK, C. D., SELAM, J. L., PITTS, H. A., WAXMAN, K., RUBIO, M., JEANDIDIER, N., TURNER, D., FISCHELL, R. E. e CHARLES, M. A. "A Preliminary Trial of the Programmable Implantable Medication System for Insulin Delivery". *The New England Journal of Medicine*, v. 321, 31 ago. 1989. p. 574–579.

22

Teoria da Amostragem

Quando estudamos inferência, aprendemos que um dos objetivos principais da estatística é descrever algumas características de uma população usando a informação contida em uma amostra de observações. Nos capítulos anteriores em que estimamos uma média, assumimos que a população original — tal como a de níveis séricos de colesterol de todos os homens adultos nos Estados Unidos — era infinita com média μ e desvio-padrão σ . Dessa população, uma amostra aleatória de tamanho n foi selecionada. O teorema central do limite nos diz que a distribuição da média dos valores da amostra era aproximadamente normal com média μ e desvio-padrão σ/\sqrt{n} . Era fundamental que a amostra fosse representativa da população, de modo que as conclusões extraídas fossem válidas. Este capítulo fornece detalhes adicionais de algumas questões importantes com relação à teoria da amostragem.

22.1 Esquemas de Amostragem

Suponha que, em vez de ser infinita, nossa população original seja finita e consista de um total de N indivíduos ou objetos. Se N é grande, pode ainda não ser viável avaliar todos os elementos da população. Então, queremos novamente fazer inferência sobre uma característica específica da população, usando a informação contida em uma amostra de indivíduos.

Os elementos individuais na população de interesse são chamados de *unidades de estudo*, ou *unidades amostrais*; uma unidade de estudo pode ser uma pessoa, uma família, uma cidade, um objeto ou alguma coisa mais que seja a unidade de análise na população. Por exemplo, suponha que queiramos determinar a quantidade média de álcool consumida cada semana por pessoas de 15 a 17 anos que vivem no Estado de Massachusetts. Nesse caso, as unidades de estudo serão adolescentes entre 15 e 17 anos que residem em Massachusetts em uma data determinada.

A população ideal que queremos descrever é conhecida como *população-alvo*. No exemplo anterior, a população-alvo consiste de todos os adolescentes de 15 a 17 anos que vivem em Massachusetts. Em muitas situações, a população-alvo inteira não está acessível. Se usarmos os registros escolares para selecionar nossa amostra de adolescentes, por exemplo, os indivíduos que não freqüentam o curso colegial não teriam chances de ser incluídos. Depois que consideramos as restrições práticas, o grupo do qual podemos realmente amostrar é conhecido como a *população de estudo*. Uma lista dos elementos da população de estudo é chamada de *sistema de referência*. Note-se que uma amostra aleatória, embora representativa da população de estudo da qual é selecionada, pode não ser representativa da

população-alvo. Se os dois grupos diferem de algum modo importante — talvez a população de estudo seja mais jovem na média do que a população-alvo — diz-se que a amostra selecionada é tendenciosa. O viés de seleção é uma tendência sistemática de se excluir certos membros da população-alvo.

22.1.1 Amostragem Aleatória Simples

O tipo mais elementar de amostra que pode ser extraída da população de estudo é uma *amostra aleatória simples*, na qual as unidades são independentemente selecionadas, uma de cada vez, até que o tamanho desejado da amostra seja atingido. Como determinada unidade só pode ser escolhida uma vez, essa estratégia é um exemplo de *amostragem sem reposição*. Toda unidade de estudo na população finita tem chance de ser incluída na amostra; a probabilidade de que uma unidade particular seja escolhida é n/N , em que n é o tamanho da amostra e N é o tamanho da população original. A quantidade n/N é a *fração amostral* da população.

Quando a população original de tamanho N tem média μ e desvio-padrão σ , uma versão finita do teorema central do limite estabelece que a distribuição da média da amostra \bar{x} tem média μ e desvio-padrão $\sqrt{1 - (n/N)}(\sigma/\sqrt{n})$. Note-se que o desvio-padrão da média da amostra da população finita difere daquele de uma população infinita por um fator de $\sqrt{1 - (n/N)}$. O quadrado dessa quantidade, ou $1 - (n/N)$, é chamado de *fator de correção da população finita*. Se n tem algum valor fixo e N é muito grande, n/N é próximo de 0. Nesse caso, o fator de correção da população finita é aproximadamente 1 e retornamos à situação familiar na qual o desvio-padrão da média da amostra é σ/\sqrt{n} . Se a população inteira está incluída na amostra, então n/N é igual a 1 e o desvio-padrão é 0. Quando a população inteira é avaliada, não há variabilidade amostral na média.

Um modo de se escolher uma amostra aleatória simples é listar e numerar cada unidade de estudo, misturá-las inteiramente e então selecionar unidades desse sistema de referência até que o tamanho exigido de amostra seja atingido. Outro modo é usar um computador ou uma tabela de números aleatórios para identificar as unidades a serem incluídas na amostra. Em ambos os casos, cada unidade deve ter uma probabilidade igual de ser escolhida. Desse modo, a possibilidade de viés é grandemente reduzida. Por exemplo, para se determinar a quantidade média de álcool consumida em cada semana por adolescentes de 15 a 17 anos em Massachusetts, não seria viável entrevistar cada um deles. Em vez disso, uma lista numerada desses indivíduos pode ser compilada e um dos métodos precedentes usado para se selecionar uma amostra de tamanho n . A informação exigida seria então obtida somente a partir dos membros desse grupo, em vez da população inteira.

22.1.2 Amostragem Sistemática

Se uma lista completa de N elementos de uma população está disponível, a amostragem sistemática pode ser usada. A amostragem sistemática é similar à aleatória simples, mas é mais fácil de se aplicar, na prática. Se uma amostra de tamanho n é desejada, a fração amostral da população é n/N , o que equivale a uma fração amostral de $1/(N/n)$ ou 1 em N/n . Então, a unidade amostral inicial é aleatoriamente selecionada das primeiras k unidades da lista, na qual k é igual a N/n . Se N/n não é um inteiro, seu valor é arredondado para o número inteiro mais próximo. A seguir, conforme percorremos a lista, cada k -ésima unidade consecutiva é escolhida. Por exemplo, para selecionar uma amostra de n adolescentes de 15 a 17 anos que vivem em Massachusetts, primeiramente selecionamos aleatoriamente um número entre 1 e $k = N/n$. Suponha que escolhemos 4. Obtemos então a

informação sobre a quarta pessoa da lista, assim como as pessoas $4 + k$, $4 + 2k$, $4 + 3k$, e assim por diante.

Idealmente, um sistema de referência deve ser uma lista completa de todos os membros da população-alvo. Entretanto, esse raramente é o caso na realidade. Em algumas situações, é impossível idealizar um sistema de referência para a população que desejamos estudar. Suponha que estejamos interessados nos indivíduos que usarão um serviço particular de cuidados com a saúde no próximo ano. Mesmo quando um sistema de referência não esteja disponível, a amostragem aleatória sistemática freqüentemente pode ser aplicada. Ainda queremos amostrar uma fração de 1 em N/n . Se o tamanho da população N não é conhecido, precisa ser estimado. Nesse caso, a unidade de estudo inicial é aleatoriamente selecionada das primeiras k unidades que se tornam disponíveis. Depois disso, toda k -ésima unidade consecutiva é escolhida. Assim, o sistema de referência é compilado conforme o estudo progride.

Diferentemente da amostragem aleatória simples, a amostragem sistemática exige somente a seleção de um único número aleatório. Ela também distribui a amostra homogeneamente na lista inteira da população. Tendências podem surgir se houver algum tipo de seqüência periódica ou cíclica na lista; no entanto, tais padrões são raros. Se pudermos assumir que a lista é aleatoriamente ordenada, podemos tratar a amostra resultante como uma amostra aleatória simples.

22.1.3 Amostragem Estratificada

A amostragem aleatória simples não considera qualquer informação que seja conhecida sobre os elementos de uma população e que possa afetar a característica de interesse. Nesse esquema de amostragem, é também possível que um subgrupo particular da população não esteja representado em uma determinada amostra puramente ao acaso. Quando se seleciona adolescentes em Massachusetts, por exemplo, podemos não escolher nenhum adolescente do sexo masculino de 17 anos. Isso poderia ocorrer meramente como o resultado da variação amostral. Se sentirmos que é importante incluir adolescentes do sexo masculino de 17 anos, podemos evitar esse problema ao selecionarmos uma *amostra aleatória estratificada*. Nosso estrato pode consistir de várias combinações de sexo e de idade: 15 anos homens, 15 anos mulheres, 16 anos homens, 16 anos mulheres, 17 anos homens e 17 anos mulheres. Ao usarmos esse método, dividimos a população em H subgrupos distintos, ou estratos, tais que o h -ésimo estrato tenha tamanho N_h . Uma amostra aleatória simples separada de tamanho n_h é então selecionada de cada estrato, resultando em uma fração amostral de n_h/N_h para o h -ésimo subgrupo. Esse método assegura que cada estrato esteja representado na amostra global. No entanto, ele não requer que todas as unidades de estudo tenham uma probabilidade igual de serem selecionadas. Certos subgrupos pequenos de uma população podem ser sobre-amostrados para fornecer números suficientes para uma análise mais profunda.

Quando a amostragem estratificada é usada, a média da população é estimada como a média ponderada das médias das amostras específicas dos estratos. Além disso, a variância é calculada como a média ponderada das variâncias dentro dos estratos. As variâncias entre as unidades de estudo em subgrupos diferentes não contribuem. Então, podemos tornar a variância global — ou o desvio-padrão — menor, escolhendo subgrupos tais que as unidades de estudo dentro de um estrato particular sejam tão homogêneas quanto possível, diminuindo as variâncias dentro dos estratos, enquanto as unidades nos distintos estratos são tão diferentes quanto possível, aumentando as variâncias entre os estratos. Se os tamanhos das amostras específicas dos estratos são escolhidos apropriadamente, a média estimada de uma amostra estratificada tem variância menor e é consequentemente mais precisa do que a média de uma amostra aleatória simples.

22.1.4 Amostragem por Conglomerados

Se as unidades de estudo formam grupos naturais ou se uma lista adequada da população inteira for difícil de compilar, a *amostragem por conglomerados* pode ser considerada, o que envolve selecionar uma amostra aleatória de grupos ou conglomerados e considerar todas as unidades de estudo dentro dos grupos escolhidos. Na *amostragem em dois estágios*, uma amostra aleatória de conglomerados é selecionada e então, dentro de cada conglomerado, seleciona-se uma amostra aleatória de unidades de estudo. Se nossos conglomerados fossem escolas em Massachusetts, por exemplo, poderíamos começar selecionando uma amostra de escolas seguida por uma amostra de adolescentes de 15 a 17 anos dentro de cada escola. A amostragem por conglomerados normalmente é mais econômica do que os outros tipos de amostragens: economiza tempo e dinheiro. Para amostras de mesmo tamanho, no entanto, ela produz variâncias maiores do que a amostragem aleatória simples.

22.1.5 Amostragem Não-Probabilística

Todas as estratégias de amostragem descritas resultam em *amostras probabilísticas*. Porque a probabilidade de se incluir cada indivíduo de uma população na amostra é conhecida, inferências válidas e confiáveis podem ser feitas, o que não pode ser dito das *amostras não-probabilísticas*, nas quais a probabilidade de um indivíduo ser incluído não é conhecida. Exemplos de amostras não-probabilísticas são amostras de conveniências e amostras constituídas de voluntários, tais como os indivíduos autopsiados e doadores de sangue. Esses tipos de amostras têm propensão a serem tendenciosas e não podem ser assumidas como representativas de alguma população-alvo.

Em geral, a escolha de uma estratégia de amostragem depende de vários fatores, incluindo os objetivos do estudo e a disponibilidade dos recursos [1]. Os custos e benefícios dos vários métodos devem ser pesados cuidadosamente. Na prática, um pesquisador freqüentemente combina duas ou mais estratégias diferentes.

22.2 Fontes de Tendência

Não importa qual seja o esquema de amostragem. Quando escolhemos uma amostra, a tendência da seleção não é a única fonte potencial de erro. Uma segunda fonte de tendência é a *não-resposta*. Em situações nas quais as unidades de estudo são pessoas, existem indivíduos que tipicamente não podem ser atingidos ou que não fornecerão a informação solicitada. A tendência está presente se esses não-respondentes diferem sistematicamente dos indivíduos que respondem à informação solicitada.

Considere os resultados do estudo seguinte, no qual uma amostra de 5.574 psiquiatras dos Estados Unidos foram pesquisados. Um total de 1.442, ou somente 26%, retornaram o questionário [2]. Dos 1.057 homens que responderam às perguntas, 7,1% admitiram ter contato sexual com uma ou mais pacientes. Das 257 mulheres, 3,1% confessaram ter essa prática. Como pode a não-resposta afetar essas estimativas?

Uma vez que o juramento de Hipócrates proíbe expressamente o contato sexual entre o psiquiatra e seus pacientes, parece improvável que houvesse alegação de contato sexual se realmente não ocorreu. Também parece provável que existam psiquiatras que tiveram contato sexual com um paciente e subseqüentemente recusaram-se a retornar a pesquisa. Conseqüentemente, é provável que as porcentagens calculadas dos dados da pesquisa subestimam as proporções verdadeiras da população; por quanto, não se sabe.

Uma fonte potencial de tendência adicional em uma pesquisa de amostragem resulta do fato de que um respondente prefira mentir em vez de revelar algo que lhe seja sensível ou incriminatório, fato que pode ser verdadeiro para os psiquiatras mencionados, por exemplo. Outra situação na qual um indivíduo pode não falar a verdade é em um estudo que investiga padrões de abusos substanciais durante a gravidez. Em alguns estados, uma mulher que confessa usar cocaína durante a gravidez corre o risco de ter sua criança tomada; esse risco pode ser incentivo suficiente para que ela minta.

Em outras circunstâncias, uma pessoa pode mentir, mesmo que as consequências não sejam tão horríveis. Por décadas, levantamentos de opinião pública têm consistentemente registrado que 40% dos americanos freqüentam igrejas ou sinagogas pelo menos uma vez por semana. Essa porcentagem é mais alta do que na maioria de outras nações ocidentais. Um estudo conduzido em 1993 verificou os números de freqüência nos serviços religiosos em um município selecionado de Ohio, assim como de várias outras igrejas ao redor do município, e encontrou que a freqüência verdadeira estava mais perto de 20% do que de 40% [3]. Estudos de acompanhamento sugeriram que é freqüente a maioria dos membros comprometidos com um grupo religioso exagerarem seus envolvimentos. Mesmo que não tenham freqüentado um serviço durante a semana em questão, acham que podem responder afirmativamente, pois usualmente freqüentam o culto.

Um modo de minimizar o problema da mentira ou omissão em pesquisas de amostragem é aplicar a técnica da *resposta aleatorizada*. Introduzindo-se um grau extra de incerteza nos dados, podemos mascarar as respostas de indivíduos específicos, enquanto ainda podemos fazer inferência sobre a população como um todo. Se funcionar, a resposta aleatorizada reduz a motivação para a mentira.

Por exemplo, suponha que a quantidade que queremos estimar seja a prevalência da população de alguma característica, representada por π . Poderia ser a proporção de psiquiatras que tiveram contato sexual com um ou mais pacientes. Uma amostra aleatória de indivíduos da população foi questionada se possuía essa característica ou não. Em vez de lhes dizer para responderem a pergunta de uma maneira direta, uma certa proporção anônima de pessoas que respondem ao questionário — representada por a , onde $0 < a < 1$ — é instruída para responder “sim”, sob todas as circunstâncias. Os indivíduos restantes são solicitados a falarem a verdade. Então, em uma amostra de tamanho n , aproximadamente $n\pi$ pessoas sempre darão uma resposta afirmativa; as outras $n(1 - \pi)$ responderão com a verdade. Desses $n(1 - a)$, $n(1 - a)\pi$ dirão “sim” e $n(1 - a)(1 - \pi)$ dirão “não”. Se n^* é o número total de pessoas que responderam “sim”, então, na média,

$$n^* = na + n(1 - a)\pi.$$

Se subtrairmos na de cada lado dessa equação e dividirmos por $n(1 - a)$, a prevalência π da população pode ser estimada como

$$\hat{\pi} = \frac{(n^*/n) - a}{1 - a}.$$

Como exemplo, citamos um estudo conduzido na cidade de Nova York que comparou as respostas obtidas por telefone, questionando-se diretamente, com as obtidas pelo uso da resposta aleatorizada. O estudo investigou o uso ilícito de quatro drogas diferentes: cocaína, heroína, PCP e LSD. A cada indivíduo questionado solicitou-se que dispusessem de três moedas; ele tinha de lançá-las antes que lhe fosse feita uma pergunta à qual devia responder de acordo com o resultado do lançamento das moedas. As regras eram um pouco mais complicadas do que as descritas anteriormente. Se todas as moedas dessem cara, a pessoa era instruída a responder afirmativamente; se todas fossem coroa, teria de responder “não”. Se os resulta-

dos fossem uma mistura de cara e coroa, a pessoa tinha de responder com a verdade. Assim, a proporção de indivíduos que sempre respondiam “sim” era

$$\begin{aligned}a_1 &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \\&= \frac{1}{8}.\end{aligned}$$

Analogamente, a proporção que sempre fornecia uma resposta negativa era

$$\begin{aligned}a_2 &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \\&= \frac{1}{8}.\end{aligned}$$

O restante

$$\begin{aligned}1 - \frac{1}{8} - \frac{1}{8} &= \frac{6}{8} \\&= \frac{3}{4}\end{aligned}$$

foram instruídos a falar a verdade. Em uma amostra de tamanho n , aproximadamente $(3/4)n\pi$ dessas pessoas responderam “sim” e $(1/4)n(1 - \pi)$ responderam não. Se n^* é o número total de indivíduos que responderam “sim”, então

$$n^* = \frac{1}{8}n + \frac{3}{4}n\pi.$$

Conseqüentemente

$$\hat{\pi} = \frac{8n^* - n}{6n}$$

seria a estimativa da proporção que usa determinada droga. Para três das quatro drogas em questão, as proporções de indivíduos que admitiram o uso foi maior quando as respostas foram obtidas por meio da resposta aleatorizada do que quando o questionamento direto foi usado; para a cocaína, a percentagem aumentou de 11% para 21% e para a heroína foi de 3% para 10% [4]. Isso sugere que alguns indivíduos não estavam sendo completamente verdadeiros quando foram questionados diretamente.

A principal vantagem da resposta aleatorizada é reduzir a proporção de indivíduos que fornecem respostas não-verdadeiras. Embora seja impossível identificar as respostas individuais, a informação agregada ainda pode ser obtida. No entanto, desde que essa técnica introduza uma fonte extra de incerteza na análise, o estimador tem uma variância maior do que na situação em que um dispositivo de mascaramento não é usado e se assume que todos respondem à pergunta honestamente. O que perdemos na clareza, no entanto, ganhamos na acurácia.

22.3 Aplicações Adicionais

Nos capítulos anteriores, estudamos uma amostra de bebês com baixo peso ao nascer nascidos em dois hospitais-escola de Boston, Massachusetts. Para ilustrar algumas questões práticas da amostragem, tratamos agora essas crianças como se constituíssem uma população finita de tamanho $N = 100$. Podemos descrever uma característica particular dessa popula-

ção — talvez sua idade gestacional média. As 100 medidas da idade gestacional dos bebês são exibidas na Tabela 22.1 [5]; a média verdadeira da população é $\mu = 28,9$ semanas. Suponha que não conheçamos essa informação e que também não tenhamos as fontes para obtermos dados necessários de cada criança. Em vez disso, precisamos estimar a média, para obtermos a informação da fração representativa dos recém-nascidos. Como procedemos?

TABELA 22.1

Medidas da idade gestacional de uma população de 100 bebês com baixo peso ao nascer.

Identificação	Idade	Identificação	Idade	Identificação	Idade	Identificação	Idade
1	29	26	28	51	23	76	31
2	31	27	29	52	27	77	30
3	33	28	28	53	28	78	27
4	31	29	29	54	27	79	25
5	30	30	30	55	27	80	25
6	25	31	31	56	26	81	26
7	27	32	30	57	25	82	29
8	29	33	31	58	23	83	29
9	28	34	29	59	26	84	34
10	29	35	27	60	24	85	30
11	26	36	27	61	29	86	29
12	30	37	27	62	29	87	33
13	29	38	32	63	27	88	30
14	29	39	31	64	30	89	29
15	29	40	28	65	30	90	24
16	29	41	30	66	32	91	33
17	29	42	29	67	33	92	25
18	33	43	28	68	27	93	32
19	33	44	31	69	31	94	31
20	29	45	27	70	26	95	31
21	28	46	25	71	27	96	31
22	30	47	30	72	27	97	29
23	27	48	28	73	35	98	32
24	33	49	28	74	28	99	33
25	32	50	25	75	30	100	28

Para selecionar uma amostra aleatória simples de tamanho n , escolhemos unidades de estudo independentes de uma lista de elementos da população — conhecida como sistema de referência — até atingir o tamanho da amostra desejado. Suponha que queiramos extrair uma amostra de tamanho $n = 10$. Um modo de fazê-lo seria escrever os números de 1 a 100 em tiras de papel. Depois de misturá-los completamente, selecionamos 10 números diferentes. Se trabalhássemos com uma população bastante grande, esse método seria impraticável; em vez disso, poderíamos usar um computador para gerar os números aleatórios. Em ambos os casos, cada unidade de estudo tem uma probabilidade igual de ser selecionada. A probabilidade de que uma unidade particular seja escolhida é

$$\frac{n}{N} = \frac{10}{100} \\ = 0,10.$$

A razão n/N é a fração amostral da população.

Suponha que seguimos esse procedimento para extrair uma amostra aleatória simples e que selecionamos o seguinte conjunto de números

Ao retornarmos à população de bebês com baixo peso ao nascer, determinamos a idade gestacional de cada recém-nascido escolhido; os valores apropriados estão marcados na Tabela 22.1 e listados abaixo:

32 26 28 25 24 23 29 30 27 28

Note-se que as observações selecionadas em uma amostra aleatória simples não necessitam de distribuição homogênea no sistema de referência inteiro. Ao usarmos essa amostra aleatória, estimamos a média da população como

$$\bar{x} = \frac{32 + 26 + 28 + 25 + 24 + 23 + 29 + 30 + 27 + 28}{10}$$

$$= 27,2 \text{ semanas.}$$

Esse valor é um pouco menor do que a média verdadeira da população de 28,9 semanas.

Como alternativa ao procedimento descrito, podemos aplicar a técnica da amostragem sistemática. Quando uma lista completa dos N elementos de uma população está disponível, a amostragem sistemática é mais fácil de realizar, pois exige a seleção de somente um número aleatório. Como foi exemplificado, a fração amostral desejada da população de bebês com baixo peso ao nascer é 0,10, ou 1 em 10. Portanto, começaremos selecionando a unidade de estudo inicial das 10 primeiras unidades da lista.

Suponha que escrevemos os números de 1 até 10 em tiras de papel e que aleatoriamente escolhemos o número 5. Além da identificação da idade gestacional do quinto bebê da lista, determinamos a idade gestacional de cada décima criança consecutiva — a 15^a, a 25^a e assim por diante. As idades apropriadas estão exibidas abaixo.

Identificação	Idade
5	30
15	29
25	32
35	27
45	27
55	27
65	30
75	30
85	30
95	31

Essas observações estão homogeneamente distribuídas no sistema de referência. Se a lista da população estiver aleatoriamente ordenada — e não temos razões para acreditar que não esteja — uma amostra sistemática pode ser tratada como uma amostra aleatória simples. Nesse caso, portanto, estimamos a média da população como

$$\bar{x} = \frac{30 + 29 + 32 + 27 + 27 + 27 + 30 + 30 + 30 + 31}{10}$$

$$= 29,3 \text{ semanas.}$$

Agora, nossa estimativa é levemente maior do que a média verdadeira da população.

Se acharmos importante incluir números representativos de bebês dos sexos masculino e feminino em nossa amostra — se acharmos que o sexo esteja associado com a idade gestacional — podemos selecionar uma amostra aleatória estratificada. Para fazê-lo, primeiro precisamos dividir a população de bebês com baixo peso ao nascer em dois subgrupos dis-

tintos de 44 meninos e 56 meninas. Os 100 valores da idade gestacional da população, ordenados por sexo, estão exibidos na Tabela 22.2. Embora trabalhemos com duas subpopulações separadas, ainda gostaríamos de ter uma fração amostral global de 1/10. Então, devemos selecionar uma amostra aleatória simples de tamanho

$$44 \times \frac{1}{10} = 4,4 \\ \approx 4$$

do grupo de bebês do sexo masculino e uma amostra de tamanho

$$56 \times \frac{1}{10} = 5,6 \\ \approx 6$$

do grupo de bebês do sexo feminino.

Ao usarmos a amostragem aleatória simples, escolhemos as observações 2, 85, 61 e 54 para os bebês do sexo masculino. Para os bebês do sexo feminino, selecionamos 51, 14, 33,

TABELA 22.2

Medidas da idade gestacional de uma população de 100 bebês com baixo peso ao nascer, estratificada pelo sexo.

Sexo Masculino				Sexo Feminino			
Identificação	Idade	Identificação	Idade	Identificação	Idade	Identificação	Idade
1	29	72	27	3	33	49	28
2	31	75	30	4	31	50	25
6	25	76	31	5	30	51	23
7	27	77	30	8	29	55	27
15	29	85	30	9	28	57	25
16	29	86	29	10	29	58	23
21	28	87	33	11	26	59	26
23	27	88	30	12	30	60	24
24	33	89	29	13	29	62	29
26	28	90	24	14	29	65	30
28	28	91	33	17	29	66	32
31	31	92	25	18	33	67	33
34	29	95	31	19	33	68	27
37	27	96	31	20	29	69	31
39	31	97	29	22	30	70	26
41	30	98	32	25	32	73	35
42	29			27	29	74	28
43	28			29	29	78	27
47	30			30	30	79	25
48	28			32	30	80	25
52	27			33	31	81	26
53	28			35	27	82	29
54	27			36	27	83	29
56	26			38	32	84	34
61	29			40	28	93	32
63	27			44	31	96	31
64	30			45	27	99	33
71	27			46	25	100	28

25, 62 e 74. Essas observações estão marcadas na Tabela 22.2. Assim, vemos que as médias das amostras específicas dos estratos são

$$\bar{x}_{\text{masculino}} = \frac{31 + 30 + 29 + 27}{4}$$

$$= 29,3 \text{ semanas.}$$

e

$$\bar{x}_{\text{feminino}} = \frac{23 + 29 + 31 + 32 + 29 + 28}{6}$$

$$= 28,7 \text{ semanas.}$$

A média verdadeira da população é estimada como uma média ponderada dessas quantidades; portanto,

$$\bar{x} = \frac{4(29,3) + 6(28,6)}{10}$$

$$= 28,9 \text{ semanas.}$$

Por acaso, esse valor é idêntico à média verdadeira da população μ .

22.4 Exercícios de Revisão

- Quando você conduz uma pesquisa, como a população de estudo está relacionada com a população-alvo? O que é o sistema de referência?
- Como a versão finita do teorema central do limite difere da versão mais comumente usada, em que a população original é assumida ser infinita?
- Quando você pode usar a amostragem sistemática em vez da amostragem aleatória simples? Quando você prefere a amostragem estratificada?
- Como pode a não-resposta resultar em uma amostra tendenciosa? O que você pode fazer para tentar minimizar a não-resposta?
- Um estudo foi conduzido para examinar os efeitos do uso maternal de maconha e cocaína no crescimento fetal. A exposição à droga foi avaliada de duas maneiras: as mães foram questionadas diretamente durante uma entrevista e foi realizado um exame de urina [6].
 - Suponha que seja necessário confiar inteiramente na informação fornecida pelas mães. Como pode a não-resposta afetar os resultados do estudo?
 - Uma estratégia alternativa pode ser entrevistar somente mães que concordam em ser questionadas. Você acha que esse método fornece uma amostra representativa da população original de mães grávidas? Por quê?
- A cada ano, o Departamento de Agricultura dos Estados Unidos usa a renda coletada de impostos para estimar o número de cigarros consumidos no país. No período de 11 anos, de 1974 a 1985, no entanto, pesquisas repetidas de usos do fumo puderam contabilizar somente 72% do consumo total [7].
 - Como você pode explicar essa discrepância entre as estimativas de consumo de cigarro?
 - Que fonte é mais digna de crédito, a de renda de taxas de impostos ou as pesquisas de usos do fumo?
- O conjunto de dados `lowbwt` contém informações que descrevem 100 bebês com baixo peso ao nascer nascidos em Boston, Massachusetts [5] (Apêndice B, Tabela B.7). As-

suma que esses bebês constituem uma população finita. Suas medidas de pressão sanguínea sistólica estão salvas sob a variável de nome *sbp*; a pressão sanguínea sistólica média é $\mu = 47,1$ mm Hg. Suponha que não conheçamos a média verdadeira da população e que queiramos estimá-la usando uma amostra de 20 recém-nascidos.

- (a) Qual é a fração amostral da população?
 - (b) Selecione uma amostra aleatória simples e use-a para estimar a pressão sanguínea sistólica média dessa população de bebês com baixo peso ao nascer.
 - (c) Extraia uma amostra sistemática da mesma população e novamente estime a pressão sanguínea sistólica média.
 - (d) Suponha que você acredite que um diagnóstico de toxemia de uma mãe grávida possa afetar a pressão sanguínea sistólica de sua criança. Divida a população de bebês com baixo peso ao nascer em dois grupos: aqueles cujas mães foram diagnosticadas com toxemia e aqueles cujas mães não foram diagnosticadas. Selecione uma amostra aleatória estratificada de tamanho 20. Use essas pressões sanguíneas para estimar a média verdadeira da população.
 - (e) Quais são as frações amostrais em cada um dos dois estratos?
 - (f) Pode a amostragem por conglomerados ser aplicada nesse problema? Em caso positivo, como?
8. Suponha que você esteja interessado em conduzir sua própria pesquisa para estimar a proporção de psiquiatras que tiveram contato sexual com um ou mais pacientes. Como você desenvolveria esse estudo? Justifique seu método de coleta de dados. Inclua uma discussão de como você tentaria minimizar a tendência.
 9. Retorne ao primeiro exercício do Capítulo 1, que solicitou a você o planejamento de um estudo dirigido para investigar uma questão que você acredita poder influenciar a saúde do mundo. Releia sua proposta original e critique sua concepção de planejamento de estudo. O que você faria diferentemente agora?

Bibliografia

- [1] MENDENHALL, W., OTT, L. e SCHEAFFER, R. L. *Elementary Survey Sampling*. Belmont, Califórnia: Wadsworth, 1971.
- [2] GATRELL, N., HERMAN, J., OLARTE, S., FELDSTEIN, M. e LOCALIO, R. "Psychiatrist-Patient Sexual Contact: Results of a National Survey, Prevalence". *American Journal of Psychiatry*. v. 143, set. 1986. p. 1126–1131.
- [3] OWEN, Karen. "Honesty No Longer Assumed Where Religion Is Concerned". *The Owingsboro Messenger-Inquirer*. 16 jan. 1999.
- [4] WEISSMAN, A. N., STEER, R. A. e LIPTON, D. S. "Estimating Illicit Drug Use Through Telephone Interviews and the Randomized Response Technique". *Drug and Alcohol Dependence*. v. 18, 1986. p. 225–233.
- [5] LEVITON, A., FENTON, T., KUBAN, K. C. K. e PAGANO, M. "Labor and Delivery Characteristics and the Risk of Germinal Matrix Hemorrhage in Low Birth Weight Infants". *Journal of Child Neurology*. v. 6, out. 1991. p. 35–40.
- [6] ZUCKERMAN, B., FRANK, D. A., HINGSON, R., AMARO, H., LEVENSON, S. M., KAYNE, H., PARKER, S., VINCI, R., ABOAGYE, K., FRIED, L. E., CABRAL, H., TIMPERI, R. e BAUCHNER, H. "Effects of Maternal Marijuana and Cocaine Use on Fetal Growth". *The New England Journal of Medicine*. v. 320, jun. 1989. p. 762–768.
- [7] HATZIANDREU, E. J., PIERCE, J. P., FIORE, M. C., GRISE, V., NOVOTNY, T. E e DAVIS, R. M. "The Reliability of Self-Reported Cigarette Consumption in the United States". *American Journal of Public Health*. v. 79, ago. 1989. p. 1020–1023.