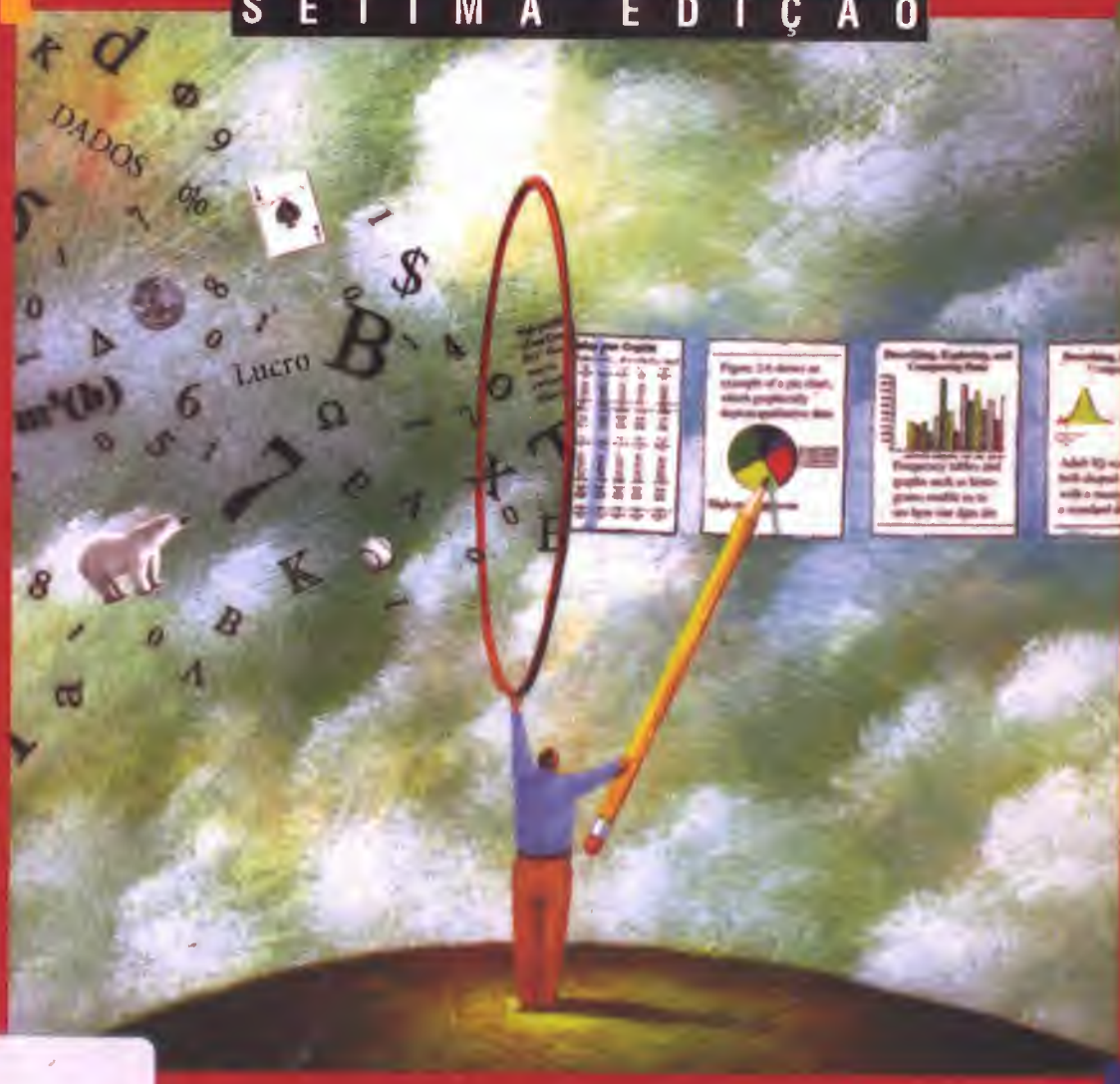


INTRODUÇÃO À ESTATÍSTICA

S É T I M A E D I Ç Ã O



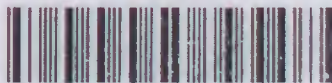
MARIO F. TRIOLA

Estado de Goiás
ACADEMIA DE POLÍCIA MILITAR
BIBLIOTECA

INTRODUÇÃO À ESTATÍSTICA

SÉTIMA EDIÇÃO

Biblioteca



00003621

00003621
0510 99
T034

ex. 3

ABPDEA
ABPDEA
ABPDEA
ABPDEA
ABPDEA
ABPDEA

Associação Brasileira para
a Proteção dos Direitos
Editoriais e Autorais

RESPEITE O AUTOR
NÃO FAÇA CÓPIA
www.abpdea.org.br

INTRODUÇÃO À ESTATÍSTICA

SÉTIMA EDIÇÃO

MARIO F. TRIOLA

Tradução

Alfredo Alves de Farias

Professor Adjunto / UFMG

Revisão técnica

Eliana Farias e Soares, Ph.D.

Professora Adjunta / UFMG

com a colaboração de

Vera Regina L. F. Flores, M. Sc.

Professora Adjunta / UFMG

No interesse de difusão da cultura e do conhecimento, os autores e os editores envidaram o máximo esforço para localizar os detentores dos direitos autorais de qualquer material utilizado, dispondo-se a possíveis acertos posteriores caso, inadvertidamente, a identificação de algum deles tenha sido omitida.

Elementary Statistics

Copyright © 1998 Addison Wesley Longman, Inc.

Published by arrangement with Addison Wesley Longman, Inc.

Capa: Barbara Atkinson

Ilustração de capa: Dave Cutler

Direitos exclusivos para a língua portuguesa

Copyright © 1999 by

LTC — Livros Técnicos e Científicos Editora S.A.

Travessa do Ouvidor, 11

Rio de Janeiro, RJ — CEP 20040-040

Tel.: 21-2221-9621

Fax: 21-2221-3202

Reservados todos os direitos. É proibida a duplicação ou reprodução deste volume, no todo ou em parte, sob quaisquer formas ou por quaisquer meios (eletrônico, mecânico, gravação, fotocópia, distribuição na Web ou outros), sem permissão expressa da Editora.

2

Descrição, Exploração e Comparação de Dados

2-1 Aspectos Gerais

O capítulo apresenta tabelas, gráficos e medidas importantes que podem ser utilizados para descrever ou explorar um conjunto de dados, ou comparar dois ou mais conjuntos. Em capítulos posteriores serão utilizados muitos conceitos importantes ora introduzidos.

2-2 Resumo de Dados com Tabelas de Frequência

Descreve-se a construção de tabelas de frequência, tabelas de frequência relativa e tabelas de frequência acumulada. Essas tabelas são úteis para condensar grandes conjuntos de dados, facilitando o seu manuseio.

2-3 Representação Pictórica de Dados

Apresentamos métodos de construção de histogramas, histogramas de frequências relativas, gráficos por pontos, gráficos tipo ramo-e-folha, gráfico em setores, diagramas de Pareto e diagramas de dispersão. Tais gráficos auxiliam grandemente a visualização de características dos dados que, de outra forma, permaneceriam encobertas.

2-4 Medidas de Tendência Central

As medidas de tendência central são tentativas de determinação de valores que representam conjuntos de

dados. Definimos as seguintes medidas de tendência central: média, mediana, moda, ponto médio e média ponderada. Abordamos também o conceito de assimetria.

2-5 Medidas de Variação

As medidas de variação são números que refletem o grau de dispersão entre os valores de um conjunto de dados. Definem-se as seguintes medidas de variação: amplitude, desvio-padrão, desvio médio e variância. Tais medidas têm extrema importância em análise estatística.

2-6 Medidas de Posição

Define-se o escore padronizado (ou escore z), mostrando como identificar valores atípicos. Definem-se também percentis, quartis e decis, utilizados para comparar valores dentro do mesmo conjunto de dados.

2-7 Análise Exploratória de Dados (EDA – Exploratory Data Analysis)

Apresentamos técnicas para explorar dados com o resumo de cinco números e com diagramas de caixas (*boxplots*). Estes últimos são especialmente adaptados para comparar diferentes conjuntos de dados.

Problema do Capítulo

As latas de alumínio de 12 oz podem ter menor espessura para reduzir o custo?

O Conjunto de Dados 15 do Apêndice B inclui estas duas amostras:

1. Latas de alumínio de 12 oz com espessura de 0,0109 in. (0,0278 cm) (reproduzido como Tabela 2.1)
2. Latas de alumínio com espessura de 0,0111 in. (0,0282 cm)

Exploraremos os valores da Tabela 2.1, que relaciona as cargas axiais (em libras) da amostra de latas de alumínio de 0,0109 in. de espessura. Este conjunto de dados foi fornecido por um estudante que utilizou a edição anterior deste livro. Trata-se de uma funcionária da companhia que fabrica essas latas; ela utiliza métodos aprendidos em seu curso introdutório de estatística. O autor agradece essa contribuição.

A carga axial de uma lata é o peso máximo suportado por seus lados, e é medida utilizando-se uma placa para aplicar uma pressão crescente ao topo da lata, até que ela ceda. É importante termos uma carga axial suficientemente grande a fim de a lata não ceder quando se coloca a tampa sob pressão. Nesse processo de fabricação, os topos das latas são colocados no lugar com uma pressão que varia de 158 a 165 libras.

As latas menos espessas têm a vantagem óbvia de utilizar menos material, o que reduz o custo, mas não são provavelmente tão resistentes quanto as mais espessas. A empresa que fabrica essas latas costuma utilizar uma espessura de 0,0111 in. mas está testando latas de menor espessura. Com os métodos deste capítulo, exploraremos o conjunto de dados (reproduzido na Tabela 2-1) para essas latas menos espessas (0,0109 in. de espessura). E determinaremos, afinal, se essas latas menos espessas podem realmente ser usadas.



TABELA 2-1 Cargas Axiais de Latas de 0,0109 in. (0,0278 cm)

270	273	258	204	254	228	282
(122)*	(124)	(117)	(93)	(115)	(103)	(128)
278	201	264	265	223	274	280
(126)	(91)	(119)	(120)	(101)	(124)	(104)
250	275	281	271	263	277	275
(113)	(125)	(127)	(123)	(119)	(126)	(125)
278	260	262	273	274	286	236
(126)	(118)	(119)	(124)	(124)	(130)	(207)
290	286	278	283	262	277	295
(132)	(130)	(126)	(128)	(119)	(126)	(134)
274	272	265	275	263	251	289
(124)	(123)	(120)	(125)	(119)	(114)	(131)
242	284	241	276	200	278	283
(110)	(129)	(109)	(125)	(91)	(126)	(128)
269	282	267	282	272	277	261
(122)	(128)	(121)	(128)	(123)	(126)	(118)
257	278	295	270	268	286	262
(117)	(126)	(134)	(122)	(122)	(130)	(119)
272	268	283	256	206	277	252
(123)	(122)	(128)	(116)	(93)	(126)	(114)
265	263	281	268	280	289	283
(120)	(119)	(127)	(121)	(127)	(131)	(128)
263	273	209	259	287	269	277
(119)	(124)	(95)	(117)	(130)	(122)	(126)
234	282	276	272	257	267	204
(104)	(128)	(125)	(123)	(117)	(121)	(93)
270	285	273	269	284	276	286
(122)	(129)	(123)	(122)	(129)	(125)	(130)
273	289	263	270	279	206	270
(124)	(131)	(119)	(122)	(127)	(93)	(122)
270	268	218	251	252	284	278
(122)	(122)	(99)	(114)	(114)	(129)	(126)
277	208	271	208	280	269	270
(126)	(94)	(123)	(94)	(127)	(122)	(122)
294	292	289	290	215	284	283
(133)	(132)	(131)	(132)	(96)	(129)	(128)
279	275	223	220	281	268	272
(127)	(125)	(101)	(100)	(127)	(121)	(123)
268	279	217	259	291	291	281
(122)	(127)	(98)	(117)	(132)	(132)	(127)
230	276	225	282	276	289	288
(104)	(125)	(102)	(128)	(125)	(131)	(131)
268	242	283	277	285	293	248
(122)	(110)	(128)	(126)	(129)	(133)	(112)
278	285	292	282	287	277	266
(126)	(129)	(132)	(128)	(130)	(126)	(121)
268	273	270	256	297	280	256
(122)	(124)	(122)	(116)	(135)	(127)	(116)
262	268	262	288	290	274	292
(119)	(122)	(119)	(133)	(132)	(124)	(132)

*Os números entre parênteses são as cargas axiais em kg.

2-1 Aspectos Gerais

Às vezes coletamos dados visando a um fim específico. Por exemplo, um estudo sobre a segurança dos elevadores de um edifício exigiria dados relativos ao peso médio das pessoas que os utilizam. Em outros casos, coletamos ou obtemos dados não com uma finalidade específica, mas porque desejamos explorá-los para ver o que pode ser revelado. A um geólogo podem interessar os intervalos de tempo entre as erupções do gêiser Old Faithful — são elas igualmente espaçadas ao longo do tempo, ou alguns intervalos de tempo são mais frequentes do que outros? Em ambas as circunstâncias, necessitamos de uma diversidade de recursos que contribuam para entendermos o conjunto de dados. Este capítulo apresenta tais recursos.

Ao analisarmos um conjunto de dados, devemos determinar em primeiro lugar se se trata de uma *amostra* ou de uma *população completa*. Essa determinação afetará não somente os métodos utilizados, mas também as conclusões a que chegarmos. Utilizamos métodos de **estatística descritiva** para resumir ou descrever as características importantes de um conjunto conhecido de dados populacionais, e recorremos a métodos de **inferência estatística** quando utilizamos dados amostrais para fazer inferências (ou generalizações) sobre uma população. Quando um professor calcula a média final de um exame para determinada turma, o resultado é um exemplo de uma estatística descritiva, se consideramos a população como toda a turma. Mas se afirmamos que o resultado é uma estimativa da média do exame final de todas as turmas, estamos fazendo uma inferência que ultrapassa o âmbito dos dados conhecidos.

A estatística descritiva e a inferência estatística são dois grandes ramos da estatística. Neste capítulo abordamos os conceitos básicos da estatística descritiva.

Características Importantes dos Dados

Com os recursos da estatística descritiva, podemos entender melhor um conjunto de dados através de suas características. As três características seguintes são extremamente importantes e proporcionam uma visão bastante satisfatória:

1. A natureza ou forma da distribuição dos dados, como forma de sino, uniforme ou assimétrica.
2. Um valor representativo, como uma média.
3. Uma medida de dispersão ou variação.

Podemos conhecer alguma coisa da natureza ou forma da distribuição organizando os dados e construindo gráficos, como nas Seções 2-2 e 2-3. Na Seção 2-4, veremos como obter valores representativos. Avaliaremos a extensão da dispersão, ou variação entre dados, com auxílio dos recursos da Seção 2-5. Na Seção 2-6 definiremos medidas de posição que nos permitem melhor analisar ou comparar diversos valores. E na Seção 2-7 estudaremos métodos de exploração de conjuntos de dados.

2-2 Resumo de Dados com Tabelas de Frequência

Ao estudarmos grandes conjuntos de dados, é conveniente organizá-los e resumí-los, construindo uma tabela de frequências.

DEFINIÇÃO

Uma **tabela de frequências** relaciona categorias (ou classes) de valores, juntamente com contagens (ou frequências) do número de valores que se enquadram em cada categoria.

A Tabela 2-2 é uma tabela de frequências com 10 classes (ou categorias). A frequência de determinada classe é o número de observações originais que se enquadram naquela classe. Por exemplo, a primeira classe na Tabela 2-2 tem uma frequência de 9, indicando que há 9 valores entre 200 e 209 inclusive.

TABELA 2-2 Cargas Axiais de Latas de Alumínio

Carga Axial	Frequência
200-209	9
210-219	3
220-229	5
230-239	4
240-249	4
250-259	14
260-269	32
270-279	52
280-289	38
290-299	14

Começaremos apresentando alguns termos-padrão no estudo de tabelas de frequência e, em seguida, descreveremos um processo para construí-las. (Há vários pacotes estatísticos que constroem essas tabelas automaticamente.)

DEFINIÇÕES

Limites Inferiores de Classes são os menores números que podem efetivamente pertencer às diferentes classes. (Na Tabela 2-2 os limites inferiores de classe são 200, 210, ..., 290.)

Limites Superiores de Classes são os maiores números que podem efetivamente pertencer às diferentes classes. (A Tabela 2-2 tem os limites superiores de classe 209, 219, ..., 299.)

Fronteiras de Classes são os números usados para separar classes, mas sem as lacunas criadas pelos limites de classe. São obtidos como segue: Determinamos o tamanho da lacuna entre o limite superior de uma classe e o limite inferior da classe seguinte, adicionamos metade desse valor a cada limite superior de classe, obtendo as fronteiras superiores de classes; subtraímos metade daquele valor de cada limite inferior de classe, obtendo as fronteiras inferiores de classe. (Na Tabela 2-2 as fronteiras de classe são 199,5, 209,5, 219,5, ..., 299,5.)

Marcas de Classe são os pontos médios das classes. (Na Tabela 2-2 os pontos médios são 204,5, 214,5, ..., 294,5.) Cada marca de classe é obtida somando-se o limite inferior ao limite superior correspondente, e dividindo-se o resultado por 2.

Amplitude de Classe é a diferença entre dois limites de classe inferiores consecutivos ou entre duas fronteiras inferiores de classe consecutivas. (Na Tabela 2-2 a amplitude de classe é 10.)

As definições de marca de classe e fronteira de classe podem ser enganosas. Devemos ter o cuidado de evitar o erro de tomar como amplitude de classe a diferença entre o limite inferior de classe e o correspondente limite superior. Veja a Tabela 2-2 e note que a amplitude de classe é 10, e não 9. (Os estudantes costumam ter dificuldade com as fronteiras de classe. Veja a discussão na seção seguinte.) Observe os limites de classe na Tabela 2-2 e note que há uma lacuna entre 209 e 210, outra entre 219 e 220 e assim por diante. As fronteiras de classe basicamente dividem diferenças e preenchem as lacunas, facilitando a construção de certos gráficos. Examine cuidadosamente, durante algum tempo, a definição de fronteira de classe, até ter entendido perfeitamente.

O processo de construção de uma tabela de frequência envolve os seguintes passos:

Passo 1: *Decidir o número de classes de sua tabela de frequência.* A título de orientação, o número de classes deve ficar entre 5 e 20. O número efetivo de classes pode depender da conveniência de utilizar números arredondados ou de outros fatores subjetivos. Com notas de testes, por exemplo, pode ser conveniente utilizar 10 classes: 50-54, 55-59, 60-64, ..., 95-99.

Passo 2: *Determinar a amplitude de classe, dividindo a amplitude pelo número de classes.* (A amplitude é a diferença entre o maior e o menor valor.) Arredonde o resultado para mais, até um número conveniente. Esse arredondamento para mais não somente é conveniente como também garante que todos os valores sejam incluídos na tabela de frequências. (Se o número de classes divide exatamente a amplitude, é preciso acrescentar mais uma classe para que todos os dados sejam incluídos.)

$$\text{Amplitude de classe} = \frac{\text{amplitude}}{\text{número de classes}} \text{ arredondado para mais}$$

Passo 3: *Escolher como limite inferior da primeira classe o menor valor observado ou um valor ligeiramente inferior a ele.* Esse valor serve como ponto de partida.

Passo 4: *Some a amplitude de classe ao ponto de partida, obtendo o segundo limite inferior de classe.* Adicione a amplitude de classe ao segundo limite inferior para obter o terceiro; e assim por diante.

Passo 5: *Relacione os limites inferiores de classe em uma coluna e introduza os limites superiores, que podem ser facilmente determinados a esta altura.*

Passo 6: *Represente cada observação por um pequeno traço na classe apropriada e, com auxílio desses traços, determine a frequência total de cada classe.*

Como a determinação do número de classes ainda não é uma imposição legal, podemos tomar um número diferente de classes que resulte em uma tabela de frequências diferente e igualmente correta. Novamente frisamos que a prioridade deve ser a obtenção de uma tabela com valores convenientes e compreensíveis.



Autores Identificados

Em 1787-88, Alexander Hamilton, John Jay e James Madison publicaram anonimamente os famosos panfletos *Federalist*, como uma tentativa de convencer os nova-iorquinos a ratificarem a nascente Constituição. A identidade da maioria dos autores dos panfletos tornou-se conhecida, mas a autoria de doze deles foi contestada. Através da análise estatística das frequências de diversas palavras, podemos agora concluir que James Madison foi o autor provável desses 12 panfletos. Em muitos deles, a evidência da autoria de Madison é esmagadora, a ponto de podermos considerá-la praticamente certa.

EXEMPLO Construa uma tabela de frequências para as 175 cargas axiais de latas de alumínio da Tabela 2-1.

SOLUÇÃO Indicaremos os passos que conduzem à tabela de frequências mostrada na Tabela 2-2.

Passo 1: Começamos escolhendo 10 como o número de classes. (Muitos estatísticos recomendam de modo geral o uso de 10 classes, mas utilizam um número menor de classes para conjuntos menores de dados, e um número maior para conjuntos maiores.)

Passo 2: Com um mínimo de 200 e um máximo de 297, a amplitude total é $297 - 200 = 97$.

$$\begin{aligned} \text{intervalo de classe} &= \text{arredondamento de } \frac{97}{10} \text{ para cima} \\ &= \text{arredondamento de } 9,7 \text{ para cima} \\ &= 10 \text{ (arredondamento para cima pela conveniência de termos um número inteiro)} \end{aligned}$$

Passo 3: O menor valor é 200. Como é um valor conveniente, tomamo-lo como ponto de partida e limite inferior da primeira classe.

Passo 4: Adicionando a amplitude de classe 10 ao limite inferior 200, obtemos o próximo limite inferior 210. Prosseguindo, obtemos os outros limites 220, 230 etc.

Passo 5: Esses limites inferiores sugerem os seguintes limites superiores de classe:

200	209
210	219
etc.	

Passo 6: A coluna direita da Tabela 2-2 apresenta as contagens, ou frequências.

A Tabela 2-2 nos dá informações úteis tornando a lista de cargas axiais mais inteligível, mas perdemos a precisão dos dados originais. Por exemplo, a primeira classe 200-209 indica 9 observações, mas não há maneira de sabermos, pela tabela, quais são precisamente esses valores. Não podemos reconstruir os 175 valores iniciais das cargas axiais com base na tabela de frequências; sacrificamos a exatidão dos dados originais para termos dados mais compreensíveis.

Na construção de tabelas de frequência, devemos observar as seguintes diretrizes:

1. *As classes devem ser mutuamente excludentes.* Ou seja, cada valor original deve pertencer exatamente a uma, e uma só classe.
2. *Todas as classes devem ser incluídas, mesmo as de frequência zero.*
3. *Procurar utilizar a mesma amplitude para todas as classes, embora eventualmente seja impossível evitar intervalos com extremidade aberta, como "65 anos ou mais".*
4. *Escolher números convenientes para limites de classe.* Arredondar para cima a fim de ter menos casas decimais, ou utilizar números adequados à situação.
5. *Utilizar entre 5 e 20 classes.*
6. *A soma das frequências das diversas classes deve ser igual ao número de observações originais.*

Tabela de Frequências Relativas

Uma modalidade importante da tabela básica de frequência utiliza **frequências relativas**, que se obtêm dividindo a frequência de cada classe pela frequência total. A **tabela de frequências relativas** tem os mesmos limites de classe que a tabela de frequências; apenas, apresenta frequências relativas em lugar das frequências absolutas.

$$\text{frequência relativa} = \frac{\text{frequência da classe}}{\text{frequência total}}$$

A Tabela 2-3 apresenta as frequências relativas das 175 cargas axiais resumidas na Tabela 2-2. A primeira classe tem uma frequência relativa de $9/175 = 0,051$. (As frequências relativas também podem ser apresentadas como porcentagens; isto é, 0,051 pode expressar-se como 5,1%.) A segunda classe tem uma frequência relativa de $3/175 = 0,017$ etc. Quando calculadas corretamente, a soma das frequências relativas deve ser 1 (ou 100%), admitindo-se pequenas discrepâncias como consequência de arredondamentos.

TABELA 2-3 Frequência Relativa das Cargas Axiais de Latas de Alumínio

Carga Axial	Frequência Relativa
200-209	0,051
210-219	0,017
220-229	0,029
230-239	0,023
240-249	0,023
250-259	0,080
260-269	0,183
270-279	0,297
280-289	0,217
290-299	0,080

TABELA 2-4 Frequência Acumulada das Cargas Axiais

Carga Axial	Frequência Acumulada
Menos de 210	9
Menos de 220	12
Menos de 230	17
Menos de 240	21
Menos de 250	25
Menos de 260	39
Menos de 270	71
Menos de 280	123
Menos de 290	161
Menos de 300	175

As tabelas de frequência relativa facilitam a compreensão da distribuição e a comparação de diferentes conjuntos de dados. Assim, é mais fácil dizer que 5,1% das latas têm carga axial entre 200 e 209 lb do que dizer que 9 das 175 latas têm carga axial entre aqueles valores. Veja também o Exercício 21, para exemplo de uma situação em que a comparação é facilitada pelo uso de tabelas de frequência relativa.

Tabela de Frequências Acumuladas

Obtemos outra variante da tabela de frequências quando desejamos as frequências acumuladas. A **frequência acumulada** de uma classe é a soma das frequências daquela classe e de todas as classes que a antecedem. A Tabela 2-4, que representa as mesmas 175 latas de alumínio da Tabela 2-2, é um exemplo de **tabela de frequência acumulada**, onde se registram as frequências acumuladas em lugar das frequências das classes individuais. A comparação da coluna de frequências da Tabela 2-2 com a coluna de frequências acumuladas da Tabela 2-4 mostra que os valores das frequências acumuladas se obtêm partindo da frequência da primeira classe e somando sucessivamente as frequências de cada classe subsequente. Por exemplo, há 9 valores inferiores a 210, $9 + 3 = 12$ valores inferiores a 220 e assim por diante. Construída corretamente, a última frequência acumulada deve ser igual ao total de observações no conjunto.

Com as tabelas de frequência, podemos identificar a natureza geral da distribuição dos dados, bem como construir gráficos que facilitem a visualização dessa distribuição. Na próxima seção estudaremos esses gráficos.

2-2 Exercícios A: Habilidades e Conceitos Básicos

Nos Exercícios 1-4, identifique, para cada tabela de frequências, a amplitude da classe, os pontos médios das classes e as fronteiras de classe.

1. Ausências	Frequência	2. Ausências	Frequência
0-5	39	0-9	22
6-11	41	10-19	40
12-17	38	20-29	71
18-23	40	30-39	44
24-29	42	40-49	23

2-3 Representação Pictórica de Dados

Na Seção 2-2, utilizamos tabelas de frequências para transformar coleções de dados brutos em sumários organizados e compreensíveis. O objetivo principal desta seção é apresentar métodos de representação de dados em uma forma pictórica que nos permita visualizar facilmente a natureza da distribuição.

Histogramas e a Forma dos Dados

Um recurso gráfico, comum e importante, para apresentação de dados é o histograma, do qual temos um exemplo na Figura 2-1. Um **histograma** consiste em uma escala horizontal para os valores dos dados a serem representados, uma escala vertical para as frequências e barras para representar os valores das frequências das diversas classes. Em geral, a construção de um histograma para representar um conjunto de valores é precedida de uma tabela completa de frequências daqueles valores. Cada barra é delimitada pela fronteira inferior de classe à esquerda e pela fronteira superior de classe à direita. Obtém-se, entretanto, melhor legibilidade tomando-se os pontos médios das classes em lugar das fronteiras das classes. O histograma da Figura 2-1 corresponde diretamente à tabela de frequências (Tabela 2-2 da seção anterior).

Antes de construir um histograma com base em uma tabela de frequências, devemos atentar para as escalas usadas nos eixos vertical e horizontal. A frequência máxima (ou maior número mais próximo conveniente) deve sugerir o maior valor para a escala vertical; 0 deve ser a base. Na Figura 2-1, a escala vertical vai de 0 a 60. A escala horizontal deve ser construída de modo a abranger todas as classes da tabela de frequências. Idealmente, devemos procurar seguir a regra empírica, segundo a qual a altura vertical do histograma deve ser cerca de três quartos da largura total. Ambos os eixos devem ser demarcados sem qualquer ambigüidade.

Um **histograma de frequências relativas** tem a mesma forma e a mesma escala horizontal que um histograma, mas a escala vertical apresenta frequências relativas em lugar de frequências absolutas, como na Figura 2-2. A Figura 2-1 pode ser modificada para um histograma de frequências relativas simplesmente

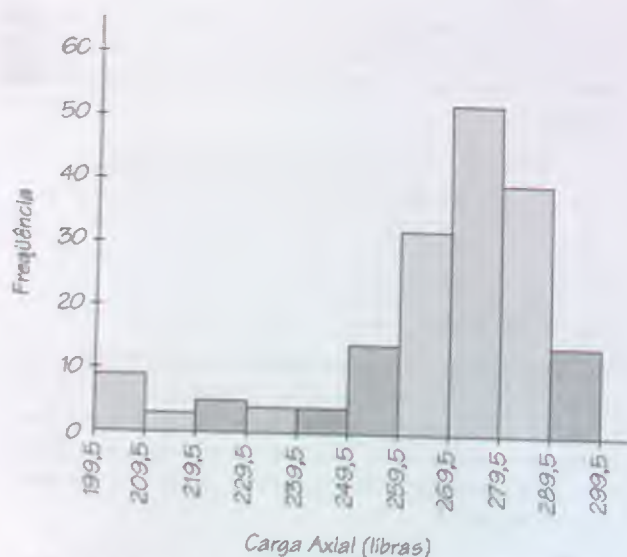


Fig. 2-1 Histograma das cargas axiais de latas de alumínio.

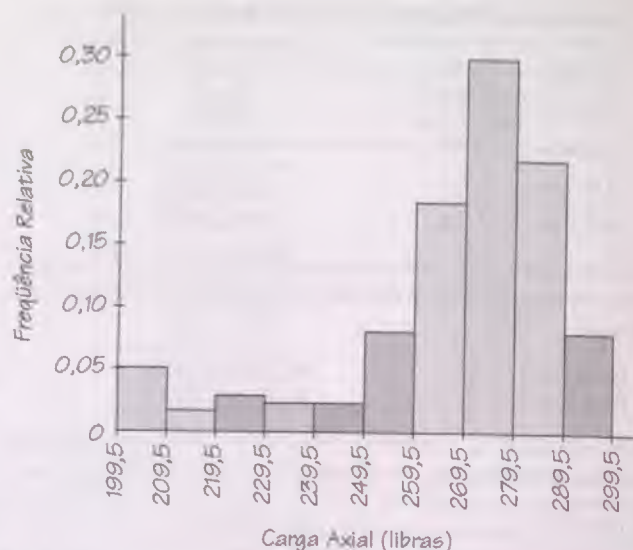
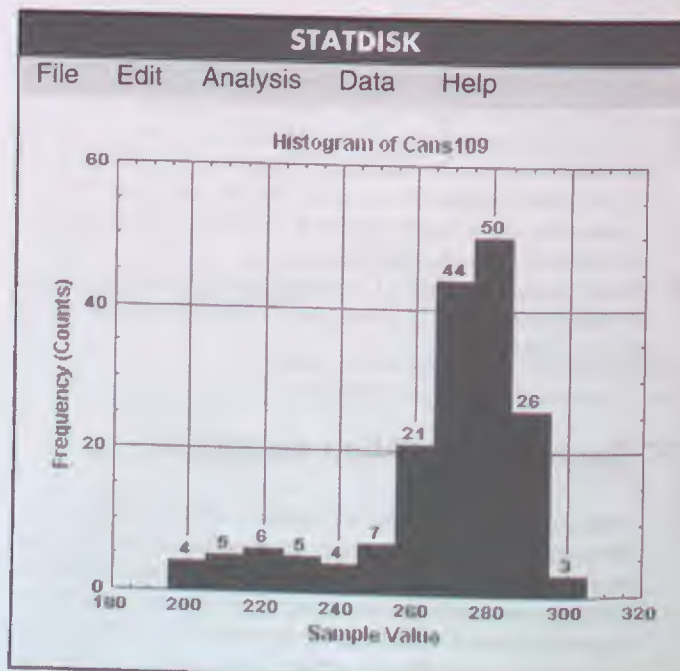


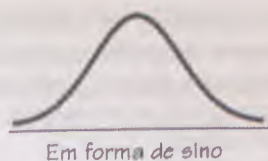
Fig. 2-2 Histograma das frequências relativas das cargas axiais de latas de alumínio.

designando a escala vertical como "frequência relativa" e modificando os valores respectivos para a escala de 0 a 0,300, conforme a Figura 2-2. (A maior frequência relativa para esse conjunto de dados é 0,297, de forma que tem sentido tomar 0,300 como valor máximo na escala vertical; o fato de a maior frequência relativa ser 0,297 e o maior valor ser 297 é mera coincidência.) Assim como o histograma da Figura 2-1 representa a tabela de frequências da Tabela 2-2, o histograma de frequências relativas da Figura 2-2 representa a tabela de frequências relativas da Tabela 2-3.

Geração de Histogramas com o Uso de Calculadoras e Computadores

Apresentamos a seguir um histograma, feito por STATDISK, das cargas axiais de latas de alumínio com que estamos trabalhando





neste capítulo. A apresentação STATDISK é obtida utilizando-se Data da barra principal de ferramentas e introduzindo-se os dados com auxílio da opção Sample Editor. Utilizam-se então os comandos Copy e Paste para usar os dados no programa Histogram, que também se encontra sob Data. (Os comandos copy e paste são comuns a muitos programas Windows.) A apresentação do histograma pode ser obtida da versão Windows de Minitab, introduzindo primeiro os dados sob a coluna C1 na grade de dados. Utilizam-se então as opções Graph e Histogram. Pode-se gerar um histograma também em algumas calculadoras gráficas, como a TI-82 e a TI-83.

As tabelas de frequências e os gráficos tais como histogramas permitem-nos ver como se distribuem nossos dados; a distribuição dos dados é uma característica extremamente importante. As Figuras 2-3 e 2-4 são histogramas de dados reais (ver Conjuntos de Dados 12 e 13 no Apêndice B) com distribuições fundamentalmente diferentes.

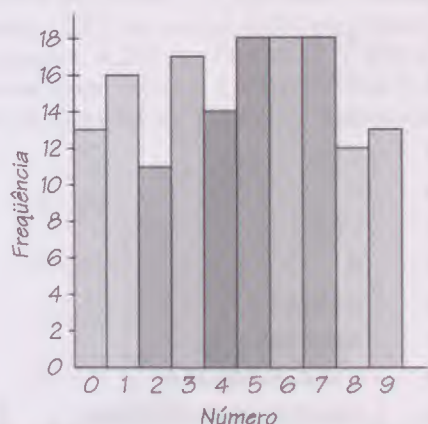


Fig. 2-3 Histograma dos resultados de uma loteria.

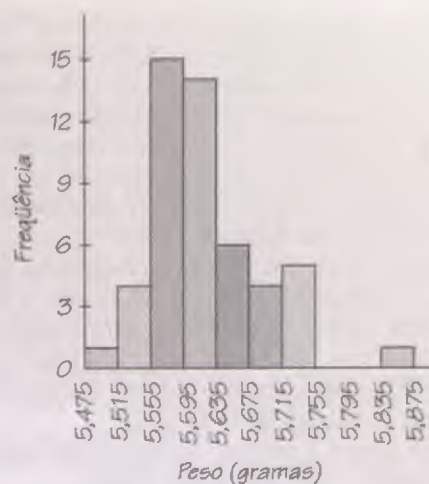


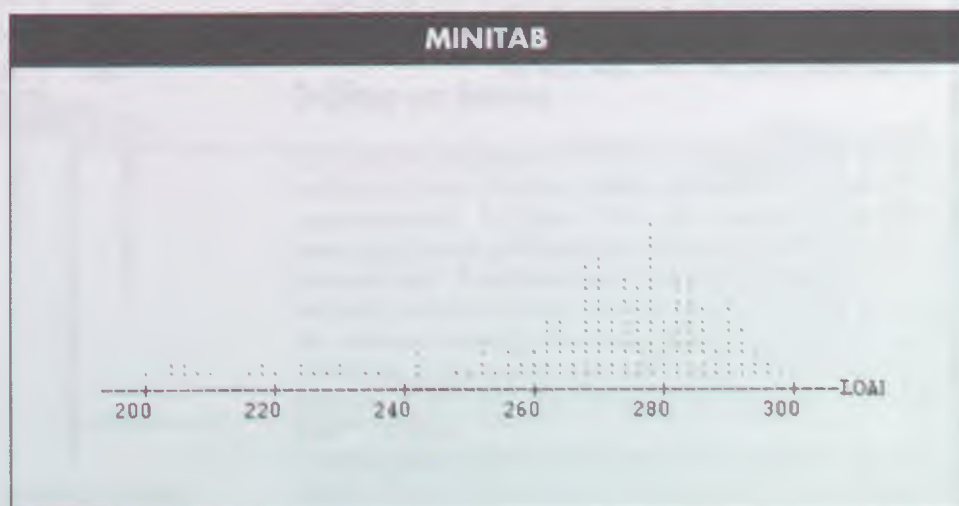
Fig. 2-4 Histograma dos pesos de moedas de 25 cents.

A Figura 2-3 é basicamente plana, ou uniforme, enquanto a Figura 2-4 tem aproximadamente a forma de um sino, no sentido de que se assemelha à segunda figura anterior sem número ilustrada aqui. Como a Figura 2-3 mostra algarismos selecionados da Loteria Pick Three de Maryland, é de se esperar que todos os algarismos sejam igualmente prováveis e que o histograma seja basicamente plano, como na Figura 2-3. Qualquer discrepância sensível da forma plana, ou uniforme, sugere que há algo errado com a loteria.

A forma de sino dos pesos das moedas de 25 centavos de dólar da Figura 2-4 é típica de uma ampla diversidade de circunstâncias, especialmente em processos de fabricação. Muitos processos estatísticos exigem que um conjunto de dados tenha uma distribuição em forma de sino análoga à apresentada na Figura 2-4, e uma maneira de verificar esse comportamento consiste em construir um histograma.

Gráficos por Pontos

A figura a seguir é um gráfico por pontos dos mesmos dados relativos a latas de alumínio relacionados na Tabela 2-1, obtido com o programa Minitab. (Com Minitab, introduzimos os dados



e selecionamos as opções *Graph*, *Character Graphs* e *Dotplot*.) Por esta ilustração, é muito fácil ver que um **gráfico por pontos** consiste em um gráfico em que cada observação é representada por um ponto ao longo da escala de valores. O ponto mais à esquerda, por exemplo, representa a carga axial de 200 lb. Quando os valores ocorrem mais de uma vez, são marcados como pontos em colunas verticais acima do valor correspondente na escala. Assim é que, nos dados da Tabela 2-1, a carga de 204 ocorre duas vezes, e esses valores são representados pelos dois pontos situados acima da locação correspondente a 204. (Este gráfico por pontos utiliza 10 intervalos para representar uma amplitude de 20 libras, de forma que cada traço na escala horizontal tracejada representa dois valores. O traço logo antes de 200 representa os valores de 199 e 200.)

O gráfico por pontos é análogo ao histograma pelo fato de permitir que vejamos a *distribuição dos dados*.



Gráficos Ramo-e-folhas

Já vimos que a construção de uma tabela de freqüências e do histograma correspondente nos dá informações valiosas sobre a natureza da distribuição dos dados, mas há a desvantagem de perdermos alguns detalhes sobre os mesmos. Em geral, não podemos recompor os dados originais a partir da tabela de freqüências ou do histograma. Vamos introduzir agora os gráficos do tipo **ramo-e-folhas**, que permitem vermos a distribuição dos dados *sem* perda de informação no processo.

Em um **gráfico ramo-e-folhas**, classificamos os dados segundo um padrão que revela a distribuição subjacente. O padrão consiste em separar um número (como 257) em duas partes — em geral, o primeiro ou os dois primeiros algarismos (25) e o outro algarismo (7). O ramo consiste nos algarismos mais à esquerda (25 neste caso), e as folhas consistem nos algarismos mais à direita (7, no caso). O método é ilustrado no exemplo seguinte.

EXEMPLO Construa um gráfico ramo-e-folhas com as cargas axiais de latas de alumínio da Tabela 2-1.

SOLUÇÃO Tomando os dois algarismos mais à esquerda como ramos, estes serão 20, 21, ..., 29. Traçamos então uma reta vertical e relacionamos as folhas conforme mostrado a seguir. O primeiro valor na Tabela 2-1 é 270; incluímos este valor registrando um 0 na linha (ramo) para 27. Continuamos a incluir todos os 175 valores, e compomos as folhas (os algarismos localizados à direita) de forma que os números se disponham em ordem crescente. A primeira linha representa os números 200, 201, 204, 204, 206 etc.

Ramo	Folhas
20	014466889
21	578
22	03358
23	0046
24	1228
25	01122466677899
26	0122223333345556677888888889999
27	000000011222223333344445556666777777788888899
28	00011112222233333444455566667789999
29	00011222334557

Deitando a página, podemos ver a distribuição desses dados. Eis a grande vantagem do gráfico ramo-e-folhas: Podemos visualizar a distribuição dos dados e, ainda assim, conservar toda a informação da lista original; se necessário, podemos recompor a relação original de valores.

O leitor notará que as linhas de algarismos em um gráfico ramo-e-folhas são análogas, em natureza, às barras de um histograma. Uma das diretrizes para a construção de histogramas é que o número de classes esteja entre 5 e 20; essa mesma orientação se aplica aos gráficos ramo-e-folhas, pelas mesmas razões. Tais gráficos podem ser *ampliados* de modo a incluir mais linhas, como podem também ser *condensados*, para reduzir o número de linhas. O gráfico ramo-e-folhas do exemplo precedente pode ser ampliado subdividindo-se as linhas entre aquelas com os algarismos 0 a 4 e as que contêm os algarismos 5 a 9. Mostramos aqui esse ramo-e-folhas ampliado. Quando se torna necessário *reduzir* o número de linhas, podemos *condensar* um gráfico ramo-e-folhas combinando linhas adjacentes, conforme ilustrado a seguir. Note que separamos por um asterisco os algarismos nas folhas associadas a cada ramo. Cada linha no gráfico condensado deve conter precisamente *um* asterisco, de modo que a forma do gráfico não sofra distorção.

Ramo	Folhas
20	0144
20	66889
21	
21	578
22	033
22	58
23	004
23	6
24	122
24	8
25	011224
25	66677899
26	012222333334
26	5556778888888889999
27	0000000112222233333334444
27	555566666777777778888888999
28	00011112222233333334444
28	555666677899999
29	00011222334
29	557

78-79	07*4	← Esta linha representa 780, 787, 794
80-81	*55	← Esta linha representa 815, 815.
82-83	9*	← Esta linha representa 829.
84-85	*	← Esta linha não tem dados.
86-87	79*0	← Esta linha representa 867, 869, 870.

Outra vantagem dos gráficos ramo-e-folhas é que sua construção constitui um processo rápido e fácil para ordenar os dados. A ordenação dos dados é necessária em vários processos estatísticos, como o cálculo da mediana (abordado na Seção 2-4) e a determinação de percentis ou quartis (Seção 2-6).

A Utilização de Computadores para Gráficos Ramo-e-folhas

O STATDISK não faz gráficos ramo-e-folhas, mas o Minitab os faz. Com o Minitab, introduza os dados na coluna C1 e utilize as opções Graph, Character Graphs e Stem-and-Leaf. A apresentação Minitab inclui uma coluna adicional de totais acumulados.



Diagramas de Pareto

Consideremos a afirmação: De 75.200 mortes por acidente nos EUA, em um ano recente, 43.500 foram causadas por veículos motorizados, 12.200 por quedas, 6.400 por envenenamento, 4.600 por afogamento, 4.200 por incêndios, 2.900 por ingestão de alimentos ou de um objeto, e 1.400 por armas de fogo (com base em dados do Conselho de Segurança Nacional). O ponto fraco

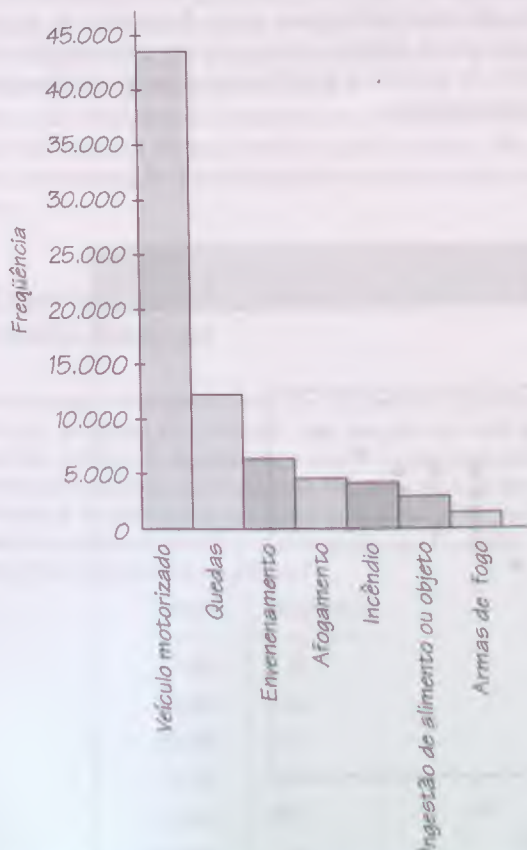



Fig. 2-5 Diagrama de Pareto: causas de mortes acidentais.

dessa afirmação escrita é não caracterizar bem um relacionamento entre categorias diferentes de dados qualitativos. Uma forma mais conveniente de indicar relações entre dados qualitativos é a construção de um diagrama de Pareto. (Recorde, da Seção 1-2, que os dados qualitativos representam uma característica não-numérica, como os tipos de morte acidental relacionados aqui.) Um **diagrama de Pareto** é um gráfico em barras para dados qualitativos, com as barras ordenadas de acordo com a frequência. Tal como no caso dos histogramas, as escalas verticais em um diagrama de Pareto podem representar frequências absolutas ou frequências relativas. A barra mais alta fica à esquerda, e as barras menores na extrema direita, conforme a Figura 2-5. Dispondo as barras por ordem de frequência, o diagrama de Pareto focaliza a atenção sobre as categorias mais importantes. Pela Figura 2-5, podemos ver que as mortes acidentais causadas por veículos motorizados representam um problema muito mais sério do que as outras categorias. Embora as mortes acidentais causadas por armas de fogo mereçam considerável atenção dos jornais, elas constituem um problema relativamente pequeno quando comparadas com as outras categorias.



Florence Nightingale

Florence Nightingale (1820-1910) é conhecida por muitos como a fundadora da profissão de enfermeira, mas ela também salvou milhares de vidas utilizando a estatística. Ao encontrar um hospital em más condições sanitárias e sem suprimentos, tratou de melhorar essas condições e passou a utilizar a estatística para convencer as autoridades da necessidade de uma reforma médica mais ampla. Elaborou gráficos originais para mostrar que, durante a guerra da Crimeia, morreram mais soldados em consequência de más condições sanitárias do que em combate. Florence Nightingale foi a pioneira na utilização não só da estatística social como das técnicas de gráficos.

Gráficos em Setores

Tal como os diagramas de Pareto, os gráficos em setores são utilizados para ilustrar dados qualitativos de modo mais compreensível. A Figura 2-6 é um exemplo de **gráfico em setor**, que ilustra graficamente dados qualitativos como fatias de uma torta. A construção de tal gráfico exige a divisão da torta em pedaços com as devidas proporções. Se a categoria de veículos motorizados responde por 57,8% do total de acidentes, então o setor que representa veículos motorizados deve ser $57,8\%$ do total. (O ângulo central deve ser $0,578 \times 360^\circ = 208^\circ$.)

O diagrama de Pareto da Figura 2-5 e o gráfico em setores da Figura 2-6 representam os mesmos dados, mas uma comparação



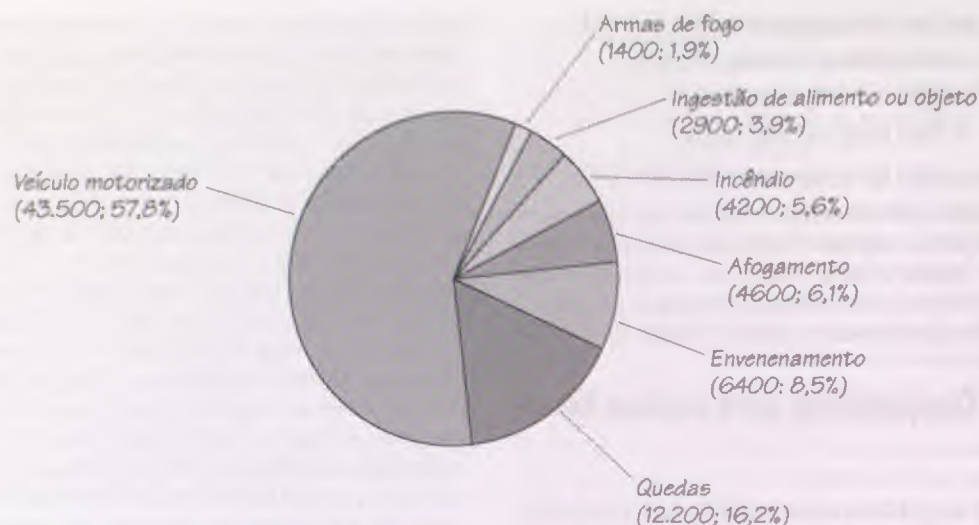


Fig. 2-6 Gráfico em setores: causas de mortes acidentais.

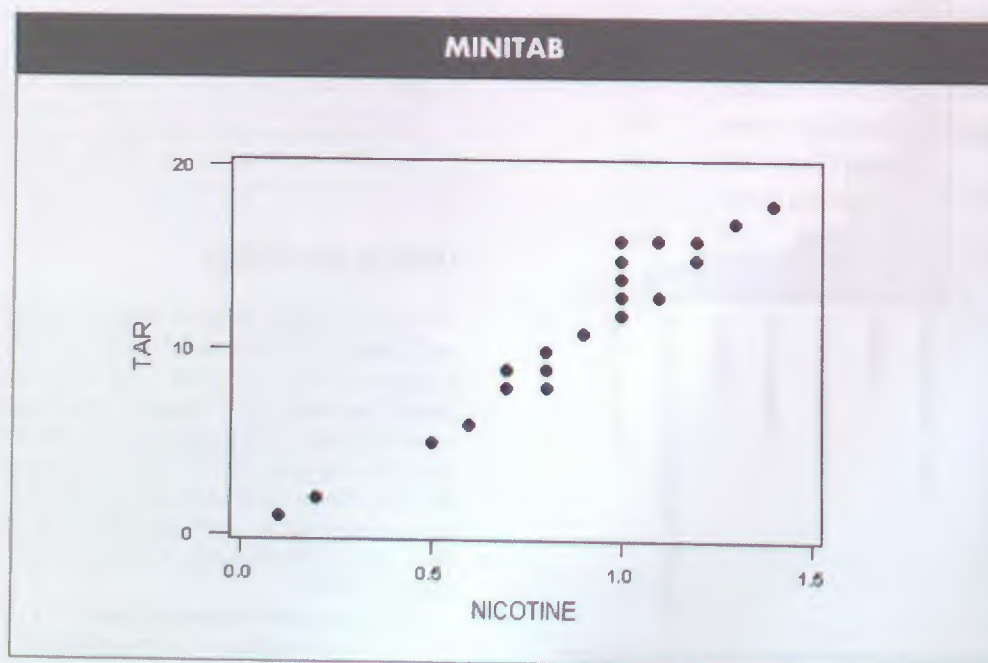
dos dois mostrará provavelmente um melhor desempenho do diagrama de Pareto para evidenciar os tamanhos relativos das diversas componentes.



Diagramas de Dispersão

Às vezes temos dados emparelhados de uma forma que associa cada valor de um conjunto a um determinado valor de um segundo conjunto. Um **diagrama de dispersão** é um gráfico dos dados emparelhados (x, y), com um eixo x horizontal e um eixo y vertical. Para construir manualmente um diagrama de dispersão, traçamos um eixo horizontal para os valores da primeira variável e um eixo vertical para os valores da segunda variável e marcamos os pontos. O padrão dos pontos assim

marcados costuma ajudar a determinar se existe algum relacionamento entre as duas variáveis. (Esse tópico será abordado extensamente quando tratarmos da correlação na Seção 9-2.) Utilizando os dados referentes à nicotina e alcatrão presentes em cigarros (Conjunto de Dados 4 do Apêndice B), geramos, com o Minitab, o diagrama de dispersão mostrado na figura. (Para obter esse gráfico, começamos introduzindo ou recuperando os dois conjuntos de dados emparelhados, de forma que eles apareçam nas colunas C1 e C2. Recorremos então às opções Graph e Plot. O STATDISK e a calculadora TI-83 também são planejados para gerar diagramas de dispersão.) Com base nesse gráfico, parece haver uma relação entre os conteúdos de alcatrão e nicotina nos cigarros, evidenciada pelo padrão dos pontos.



Tinta Invisível

O *National Observer* certa vez contratou uma firma para fazer uma pesquisa confidencial através do correio. O editor Henry Gemmill assegurou em uma circular que "cada resposta individual seria considerada confidencial, mas que, combinada a sua resposta com as outras em todo o país, teríamos um perfil de nossos assinantes". Um assinante sagaz utilizou um raio ultravioleta para detectar um código escrito na pesquisa com tinta invisível. Esse código poderia ser utilizado para identificar o autor da resposta. Gemmill não sabia que esse processo estava sendo usado, e desculpou-se publicamente. O caráter confidencial foi mantido, conforme prometido, mas a anonimidade não havia sido prometida diretamente, de forma que não foi mantida.

Outros Gráficos

Há inúmeros outros recursos pictóricos, além dos que acabamos de indicar, para representar dados de forma interessante e eficiente. O Exercício 27 se refere a um polígono de frequência, que é uma variante do histograma. Na Seção 2-7 são apresentados diagramas em caixas (*boxplots*), muito úteis para visualizar uma distribuição de dados. Os pictogramas ilustram dados por meio de figuras de objetos ou pessoas, como soldados, tanques, aviões, pilhas de moedas ou sacos de dinheiro. No Capítulo 12, diversos gráficos ilustram padrões de dados ao longo do tempo.

Considere a figura no encarte, tida talvez como "o melhor gráfico estatístico jamais traçado". A figura inclui seis variáveis diferentes relativas à marcha do exército de Napoleão sobre Moscou em 1812. A faixa grossa à esquerda ilustra o tamanho do exército quando começou a invasão da Rússia a partir da Polônia. A faixa inferior descreve a retirada de Napoleão, com as correspondentes temperaturas e datas. Embora elaborado em 1861 por Charles Joseph Minard, esse gráfico é considerado engenhoso mesmo pelos padrões atuais.

Nesta seção focalizamos a natureza ou a forma da distribuição de dados e os métodos de representá-los graficamente. Nas seções seguintes abordaremos outras maneiras de avaliar características de dados.

2-3 Exercícios A: Habilidades e Conceitos Básicos

- Os visitantes do Parque Nacional de Yellowstone consideram uma erupção do gêiser Old Faithful uma atração que não pode ser perdida. A tabela de frequências a seguir resume uma amostra de tempos (em minutos) entre erupções. Construa um histograma para a tabela de frequências dada. Se um guia turístico deseja garantir que seus turistas presenciem uma erupção, qual o tempo mínimo que devem permanecer no parque?

Tempo	Frequência
40-49	8
50-59	44
60-69	23
70-79	6
80-89	107
90-99	11
100-109	1

- Obtiveram-se na faculdade do autor os dados ao lado referentes aos carros de estudantes e aos de professores e funcionários. Construa um histograma de frequências relativas para cada conjunto de dados. Com base nos resultados, quais são as diferenças perceptíveis entre as duas amostras?

Idade	Estudantes	Funcionários e Professores
0-2	23	30
3-5	33	47
6-8	63	36
9-11	68	30
12-14	19	8
15-17	10	0
18-20	1	0
21-23	0	1

- A tabela de frequências a seguir dá as velocidades de motoristas multados pela polícia da cidade de Poughkeepsie. Esses motoristas estavam dirigindo em um trecho da zona de 30 mi/h, em Creek Road, que passa pela faculdade do autor. Construa um histograma para essa tabela de frequências. O que essa distribuição sugere sobre o limite fixado comparado com o limite de velocidade constatado?

Velocidade	Frequência
42-43	14
44-45	11
46-47	8
48-49	6
50-51	4
52-53	3
54-55	1
56-57	2
58-59	0
60-61	1

- As companhias de seguro pesquisam continuamente as idades e as causas de morte. Construa um histograma de frequências relativas correspondente à tabela de frequências ao lado. Os dados se baseiam em um estudo da revista *Time* sobre vítimas fatais de armas de fogo na América durante uma semana. O que o histograma sugere quanto às idades dessas vítimas fatais?

Idade na Morte	Frequência
16-25	22
26-35	10
36-45	6
46-55	2
56-65	4
66-75	5
76-85	1

Nos Exercícios 5 e 6, relacione os valores originais nos conjuntos de dados representados pelos dois gráficos ramo-e-folhas.

5. Ramos	Folhas	6. Ramos	Folhas
57	017	10	21 45 91
58	13349	11	11 32 77 83
59	456678	12	04 12 49
60	23	13	69

Nos Exercícios 7 e 8, construa o gráfico por pontos para os dados representados pelo ramo-e-folhas dos exercícios indicados.

7. Exercício 5

8. Exercício 6

Nos Exercícios 9-12, construa os gráficos ramo-e-folhas para os conjuntos de dados constantes do Apêndice B.

9. Os comprimentos (em polegadas) de ursos do Conjunto de Dados 3. (Sugestão: Inicialmente, arredonde os comprimentos para a polegada mais próxima.)
10. As taxas de pulsação das alunas de estatística do Conjunto 8.
11. Pesos (em gramas) das 50 moedas de 25 centavos de dólar relacionados no Conjunto de Dados 13. (Utilize um gráfico ramo-e-folhas ampliado com cerca de 8 linhas.)
12. Pesos (em libras) de artigos de plástico descartados por 62 residências. Recorra aos Dados 1 e arredonde inicialmente os pesos relacionados para o próximo décimo de libra (uma casa decimal). (Use um gráfico ramo-e-folhas ampliado com cerca de 11 linhas.)
13. Foi feito um estudo para determinar como as pessoas obtêm empregos. A tabela que segue relaciona dados de 400 pessoas escolhidas aleatoriamente. Os dados se baseiam em resultados do National Center for Career Strategies (Centro Nacional de Estratégias de Carreiras). Construa um diagrama de Pareto que corresponda aos dados em questão. Qual seria a abordagem mais eficiente para uma pessoa que deseje um emprego?

Fontes de Trabalho dos que Respondem à Pesquisa	Frequência
Anúncios tipo "Procura-se"	56
Firmas de pesquisas	44
Rádio e televisão	280
Envio de correspondência em massa	20

14. Construa um gráfico em setores para os dados do Exercício 13. Compare o gráfico em setores com o diagrama de Pareto e indique qual deles melhor apresenta a importância relativa das fontes de trabalho.
15. Uma análise de descarrilamentos de trens mostrou que 23 descarrilamentos foram causados por más condições da linha, 9 foram devidos a falhas no equipamento, 12 foram atribuídos a erro humano e 6 tiveram outras causas. [Fonte: Dados da Federal Railroad Administration (Departamento Federal de Administração de Ferrovias).] Construa um gráfico em setores para representar os dados em questão.
16. Construa um diagrama de Pareto para os dados do Exercício 15. Compare o diagrama de Pareto com o gráfico em setores, e determine qual dos gráficos mostra com maior eficiência a importância relativa das causas de descarrilamentos de trens.

Nos Exercícios 17-18, use os dados emparelhados do Apêndice B para construir um diagrama de dispersão.

17. No Conjunto de Dados 4, utilize a escala horizontal para o alcatrão e a escala vertical para o monóxido de carbono. Com base no resultado, parece haver uma relação entre o alcatrão e o monóxido de carbono nos cigarros? Em caso afirmativo, descreva esse relacionamento.
18. No Conjunto de Dados 3, use a escala horizontal para os perímetros dos pescuços dos ursos e a escala vertical para os pesos dos animais. Com base no resultado, qual é a relação entre o tamanho do peçoço de um urso e o seu peso?

Nos Exercícios 19-22, recorra aos conjuntos de dados do Apêndice B.

- a. Construa um diagrama.
- b. Descreva a forma geral da distribuição, como forma de sino, uniforme ou assimétrica.

19. Conjunto de Dados 3 do Apêndice B: pesos de ursos. (Tome 11 classes com amplitude de 50 e comece com $-0,5$ como limite inferior de classe.)
20. Conjunto de Dados 11 do Apêndice B: pesos de 100 M&Ms. (Utilize 12 classes com amplitude de 0,017 e tome 0,8375 como limite inferior de classe.)
21. Conjunto de Dados 1 do Apêndice B: pesos de papel descartado por 62 residências em uma semana. (Tome 10 classes.)
22. Conjunto de Dados 12 do Apêndice B: os 300 números sorteados na loteria de Maryland (não é a loteria Pick Three).

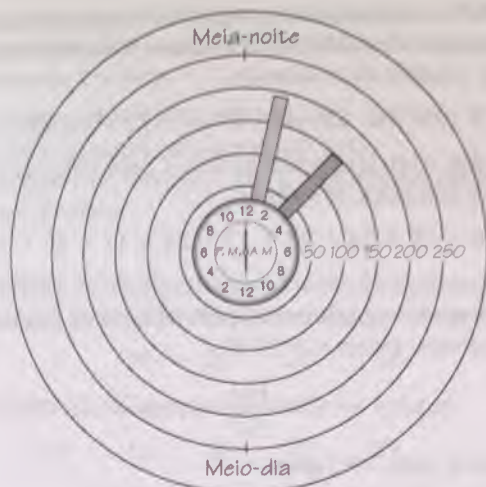
Nos Exercícios 23-26, recorra à figura do encarte, que descreve a campanha de Napoleão na Rússia em 1812. A faixa grossa à esquerda ilustra o tamanho do exército quando ele começou a invadir a Rússia a partir da Polônia, e a faixa inferior descreve a retirada de Napoleão.

23. Determine a porcentagem dos combatentes que sobreviveram a toda a campanha.
24. Determine o número e a porcentagem dos que morreram cruzando o rio Berezina.
25. Quantos morreram, no retorno de Moscou, no intervalo de tempo em que a temperatura caiu de 16°F para -6°F ?
26. Dos que chegaram a Moscou, quantos morreram no percurso de volta entre Moscou e Botr? (Observe que 33.000 homens não foram a Moscou, mas se juntaram aos que voltavam de lá.)

2-3 Exercícios B: Além do Básico

27. Um polígono de frequência é uma variante de um histograma que utiliza segmentos de retas ligando pontos em lugar de barras. Construa um polígono de frequências modificando o histograma da Figura 2-1 como segue: Inicialmente, substitua as fronteiras de classe na escala horizontal pelos pontos médios das classes. Em seguida, substitua as barras por pontos localizados acima de cada ponto médio a uma altura igual à frequência da classe. Terceiro, ligue os pontos e prolongue o gráfico à direita e à esquerda, de modo que comece e termine com uma frequência 0.
28. São fornecidas tabelas de frequência dos 100 primeiros algarismos na representação decimal do número π e dos 100 primeiros algarismos da representação decimal de $22/7$.
 - a. Construa histogramas que representem as tabelas de frequências, e assinale quaisquer diferenças.
 - b. Os números π e $22/7$ são ambos reais; mas diferem fundamentalmente um do outro; como?

π		$22/7$	
x	f	x	f
0	8	0	0
1	8	1	17
2	12	2	17
3	11	3	1
4	10	4	17
5	8	5	16
6	9	6	0
7	8	7	16
8	12	8	16
9	14	9	0



29. Com uma coleção de dados amostrais, construímos uma tabela de freqüências com 10 classes e, em seguida, construímos o histograma correspondente. Indique como o histograma é afetado se se duplica o número de classes mas se mantém a mesma escala vertical.
30. Em um estudo de seguro de acidentes com veículos motorizados no estado de Nova York, classificam-se as colisões fatais de acordo com a hora do dia, com os resultados constantes da tabela a seguir. [Fonte: Dados do New York State Department of Motor Vehicles (Departamento de Veículos Motorizados do Estado de Nova York).]
- Complete o gráfico circular e construa um histograma.
 - Qual dos dois ilustra melhor os dados? Por quê?
 - Como o período de 4 às 6 horas da manhã é o que acusa menor número de colisões fatais, podemos concluir que esse período é o mais seguro para dirigir? Por que sim ou por que não?

Hora	Número de Acidentes Fatais
Manhã 12-2	194
2-4	149
4-6	100
6-8	131
8-10	119
10-12	160
Tarde 12-2	152
2-4	221
4-6	230
6-8	211
8-10	223
10-12	178

31. No artigo "Idades dos Atores e Atrizes Ganhadores do Oscar" (revista *Mathematics Teacher*), de Richard Brown e Gretchen Davis, utilizam-se gráficos ramo-e-folha para comparar as idades de atores e de atrizes no momento da premiação. Eis os resultados para os 34 últimos vencedores recentes de cada categoria.

Atores:	32	37	36	32	51	53	33	61	35	45	55	39
	76	37	42	40	32	60	38	56	48	48	40	
	43	62	43	42	44	41	56	39	46	31	47	
Atrizes:	50	44	35	80	26	28	41	21	61	38	49	33
	74	30	33	41	31	35	41	42	37	26	34	
	34	35	26	61	60	34	24	30	37	31	27	

- a. Construa um ramo-e-folhas conjugado para esses dados. Os dois primeiros valores de cada grupo foram registrados a seguir.

Idade dos Atores	Ramo	Idade das Atrizes
	2	
72	3	
	4	4
	5	0
	6	
	7	
	8	

- b. Utilizando os resultados da parte a, compare os dois conjuntos distintos de dados e explique quaisquer diferenças.

2-4 Medidas de Tendência Central



O objetivo fundamental desta seção é apresentar as medidas de tendência central importantes.

DEFINIÇÃO

Uma medida de tendência central é um valor no centro ou no meio de um conjunto de dados.

Enquanto as Seções 2-2 e 2-3 trataram de tabelas de freqüência e gráficos que revelam a natureza ou a forma da distribuição de um conjunto de dados, esta seção focaliza a determinação de valores típicos ou representativos de um conjunto de dados. Há diferentes maneiras de definir o centro e, assim, há diferentes definições de medidas de tendência central, inclusive a média, a mediana, a moda e o ponto médio. Começemos com a média.

O Paradoxo do Tamanho de uma Turma

Há ao menos duas maneiras de obter o tamanho médio de uma turma, que podem ter resultados muito diferentes. Em uma faculdade, se tomarmos o número de alunos em 737 turmas, obtemos uma média de 40 alunos. Mas se formos compilar uma lista dos tamanhos de turma para cada estudante e utilizar essa lista, obteremos um tamanho médio de turma de 147. Essa grande discrepância é devida ao fato de que há muitos alunos em turmas grandes, mas poucos alunos em turmas pequenas. Sem alterar o número de turmas ou a faculdade, poderíamos reduzir o tamanho médio de turma, formando turmas com aproximadamente o mesmo tamanho. Isso melhoraria também o acompanhamento das aulas, que é melhor em turmas menores.

A Média

A média (aritmética) é, de modo geral, a mais importante de todas as mensurações numéricas descritivas. Na Figura 2-7 ilustramos a propriedade da média como centro do conjunto de dados, no sentido de que é um ponto de equilíbrio dos mesmos.

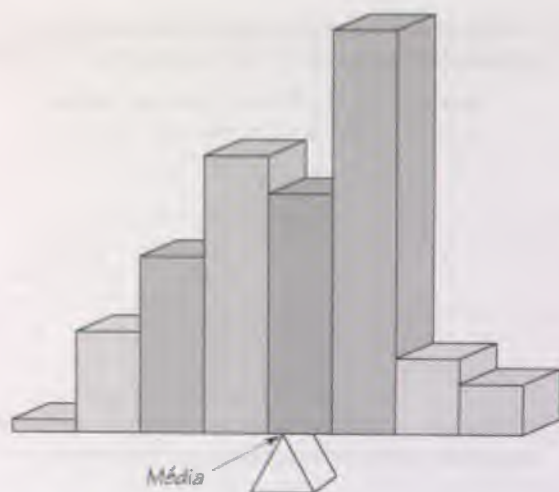


Fig. 2-7 A média como ponto de equilíbrio. Um fulcro, colocado na posição da média, equilibrará o histograma.

DEFINIÇÃO

A **média aritmética** de um conjunto de valores é o valor obtido somando-se todos eles e dividindo-se o total pelo número de valores. Essa medida particular de tendência central será utilizada freqüentemente em todo o resto deste texto, e será designada simplesmente como **média**.

Esta definição pode expressar-se como na Fórmula 2-1, onde a letra grega Σ (sigma maiúsculo) indica um **somatório** de valores, de forma que Σx representa a soma de todos os valores. O símbolo n denota o **tamanho da amostra**, que é o número de valores em consideração.

Fórmula 2-1
$$\text{média} = \frac{\Sigma x}{n}$$

A média pode denotar-se por \bar{x} (leia-se "x barra") se o conjunto de valores de que dispomos é uma amostra extraída de uma população maior; se todos os valores da população foram considerados, denotamos por μ (minúscula grega mu) a média calculada. (As estatísticas amostrais são em geral representadas por letras do alfabeto latino, como \bar{x} , ao passo que os parâmetros populacionais costumam representar-se por letras gregas, como μ .) Muitas calculadoras podem calcular a média de um conjunto de dados: introduzem-se os dados e aciona-se uma tecla \bar{x} . A introdução dos dados varia de uma calculadora para outra, de forma que é necessário consultar o respectivo manual.

EXEMPLO Relacionam-se a seguir os tempos (em anos) que os 10 primeiros presidentes americanos sobreviveram à posse. Calcule a média desta amostra.

10 29 26 28 15 23 17 25 0 20

SOLUÇÃO Aplica-se a Fórmula 2-1 para calcular a média. Primeiro somamos os valores.

$$\Sigma x = 10 + 29 + 26 + 28 + 15 + 23 + 17 + 25 + 0 + 20 = 193$$

Dividimos em seguida o total pelo número de valores. Como há 10 valores, temos $n = 10$ e

$$\bar{x} = \frac{193}{10} = 19,3$$

A média é, pois, 19,3 anos.

Para os 10 valores do exemplo precedente, 19,3 está no centro, de acordo com a definição de média. Outras definições de uma medida de tendência central envolvem diferentes percepções de como se determina o centro.

Seis Graus de Separação

Os psicólogos sociais, os historiadores, os cientistas políticos e os especialistas em comunicações estão entre os que se interessam pelo "Problema do Pequeno Mundo": Dadas duas pessoas quaisquer no mundo, quantas ligações intermediárias são necessárias para ligar as duas pessoas originais? O psicólogo social Stanley Milgram fez um experimento utilizando o sistema postal dos EUA. As pessoas foram instruídas a procurar contactar outras pessoas-alvo enviando um formulário a uma pessoa conhecida que julgassem estar próximo do alvo. Das 160 cadeias iniciais, apenas 44 foram completadas. O número de relacionamentos intermediários variou de 2 a 10, com uma mediana de 5. Utilizou-se um modelo matemático para mostrar que, se essas cadeias que faltavam fossem completadas, a mediana seria ligeiramente superior a 5. (Ver "The Small World Problem", de Stanley Milgram, *Psychology Today*, maio de 1967.)

A Mediana

DEFINIÇÃO

A **mediana** de um conjunto de valores é o valor do meio desse conjunto, quando os valores estão dispostos em ordem crescente (ou decrescente). A mediana é representada geralmente por \tilde{x} (lê-se: "x til").

Notação

Σ	denota <i>somatório</i> de um conjunto de valores.
x	é a <i>variável</i> usada para representar valores individuais dos dados.
n	representa o <i>número de valores em uma amostra</i> .
N	representa o <i>número de valores em uma população</i> .
$\bar{x} = \frac{\Sigma x}{n}$	é a <i>média de um conjunto de valores amostrais</i> .
$\mu = \frac{\Sigma x}{N}$	denota a <i>média de todos os valores de uma população</i> .

Para calcular a mediana, disponha primeiro os valores em ordem (crescente ou decrescente); em seguida aplique um dos dois processos a seguir:

1. Se o número de valores é ímpar, a mediana é o número localizado exatamente no meio da lista.
2. Se o número de valores é par, a mediana é a média dos dois valores do meio.

EXEMPLO Calcule a mediana dos tempos de sobrevivência (em anos após a posse) dos cinco primeiros presidentes americanos.

10 29 26 28 15

SOLUÇÃO Inicialmente, ordenemos os valores:

10 15 26 28 29

O número de valores é 5, que é ímpar; assim, a mediana é precisamente o número do meio. Logo, a mediana deste conjunto de dados é 26.

EXEMPLO Os valores a seguir são os pagamentos (em dólares) feitos aos executantes de um concerto de rock. A média é \$8900. Calcule a mediana.

500 600 800 50.000 1000 500

SOLUÇÃO Ordenemos inicialmente os valores:

500 500 600 800 1000 50.000

O número de valores é 6, um número par; procuramos, pois, os dois valores do meio e obtemos a sua média. Os dois valores centrais são 600 e 800; a mediana é, pois, a soma desses valores dividida por 2, ou seja, \$700.

Neste conjunto, a média de \$8900 é fortemente afetada pelo valor atípico de \$50.000, o que não ocorre com a mediana de \$700.

Moda

DEFINIÇÃO

A **moda** de um conjunto de dados é o valor que ocorre com maior frequência. Quando dois valores ocorrem com a mesma frequência máxima, cada um deles é uma moda, e o conjunto se diz **bimodal**. Se mais de dois valores ocorrem com a mesma frequência máxima, cada um deles é uma moda, e o conjunto é **multimodal**. Quando nenhum valor é repetido, o conjunto não tem moda. Costuma-se denotar a moda por *M*.

Um Cidadão Médio

O homem americano "médio" se chama Robert. Tem 31 anos, altura de 1,75 m, pesa 78 kg, seu manequim é 48, calça sapatos tamanho 43 e tem 85 cm de cintura. Consome anualmente 5,6 kg de massa, 11,8 kg de bananas, 1,8 kg de batatas fritas, 8,15 kg de sorvete e 35,8 kg de carne. Em cada ano, vê televisão durante 2567 horas e recebe 585 cartas ou assemelhadas pelo correio. Após comer sua porção de batatas fritas, ler a correspondência e ver televisão, ele termina o dia com 7,7 horas de sono. O dia seguinte começa com 21 minutos de transporte para um emprego, onde trabalha 6,1 horas.

EXEMPLO Determine a moda dos seguintes conjuntos de dados.

- a. 5 5 5 3 1 5 1 4 3 5
- b. 1 2 2 2 3 4 5 6 6 6 7 9
- c. 1 2 3 6 7 8 9 10

SOLUÇÃO

- a. O número 5 é a moda, porque é o valor que ocorre com maior frequência.
- b. Os números 2 e 6 são ambos modas, porque ocorrem com a mesma frequência máxima. O conjunto de dados é bimodal.
- c. Não há moda, porque não há valor repetido.

Das diferentes medidas de tendência central que estamos considerando, a moda é a única que pode ser usada com dados em nível nominal de mensuração, conforme ilustrado no próximo exemplo.

EXEMPLO Um estudo sobre tempos de reação abrangeu 30 canhotos, 50 destros e 20 ambidestros. Embora não possamos tomar a média numérica dessas características, podemos afirmar que a moda é destra, que é a característica que ocorre com maior frequência.

Ponto Médio

DEFINIÇÃO

O **ponto médio** é o valor que está a meio caminho entre o maior e o menor valor. Para obtê-lo, somamos esses valores extremos e dividimos o resultado por 2, como na fórmula a seguir:

$$\text{ponto médio} = \frac{\text{maior valor} + \text{menor valor}}{2}$$

EXEMPLO Determine o ponto médio dos tempos de sobrevivência (após a posse) dos 10 primeiros presidentes americanos:

10 29 26 28 15 23 17 25 0 20

SOLUÇÃO Obtém-se como segue o ponto médio:

$$\frac{\text{maior valor} + \text{menor valor}}{2} = \frac{29 + 0}{2} = 14,5 \text{ anos}$$

Embora o ponto médio não seja muito usado, incluímo-lo aqui para enfatizar o fato de que há diferentes maneiras de definir o centro de um conjunto de dados. (Veja também Exercícios 20-22.)

Ao nos referirmos ao valor médio de um conjunto de dados, devemos ser precisos, mencionando o termo exato, como média, mediana, moda ou ponto médio.

EXEMPLO Para os 175 valores de cargas axiais de latas de alumínio, relacionados na Tabela 2-1, determine (a) a média, (b) a mediana, (c) a moda e (d) o ponto médio.



SOLUÇÃO

- a. Média: A soma dos 175 valores é 46.745; assim,

$$\bar{x} = \frac{46.745}{175} = 267,1 \text{ lb}$$

- b. Mediana: Dispostos os valores em ordem crescente, verificamos que o 88.^o valor, 273, está no meio exato, de modo que a mediana é 273,0. (Os valores podem facilmente ser dispostos em ordem crescente construindo-se um gráfico ramo-e-folhas, conforme vimos na Seção 2-3, ou utilizando-se um programa de computador como STATDISK ou Minitab.) Expressamos o resultado com mais uma casa decimal utilizando a regra do arredondamento que segue este exemplo.

- c. Moda: A carga axial mais freqüente é 268 lb, que ocorre 9 vezes. É, pois, a moda.

- d. Ponto médio: Obtemo-lo aplicando a fórmula

$$\begin{aligned} \text{intervalo} &= \frac{\text{maior valor} + \text{menor valor}}{2} = \\ \text{médio} &= \frac{297 + 200}{2} = 248,5 \text{ lb} \end{aligned}$$

Passamos a resumir os resultados acima.

média: 267,1 lb

mediana: 273,0 lb

moda: 268 lb

ponto médio: 248,5 lb

Já construímos uma tabela de freqüências e um histograma para os dados da Tabela 2-1, e vimos a distribuição dos dados. Temos agora informações importantes sobre o centro dos dados.

Regra do Arredondamento

Eis uma regra simples para arredondamento de respostas:

Tome uma decimal a mais, além das que aparecem nos dados.

Devemos arredondar apenas a resposta final, e não os valores intermediários. Por exemplo, a média de 2, 3, 5 é 3,3333333..., que pode ser arredondada para 3,3. Como os dados originais são expressos em números inteiros, arredondamos a resposta para o décimo mais próximo. Outro exemplo: a média de 2,1, 3,4 e 5,7 é arredondada para 3,73 com duas decimais (uma a mais em relação às que figuram nos valores originais).

**A Média de uma Tabela de Freqüências. A Média Ponderada**

Quando os dados estão resumidos em uma tabela de freqüências, podemos aproximar a média substituindo os limites de classe pelos pontos médios das classes e supondo que todos os elementos da classe se concentrem no respectivo ponto médio. Na Tabela 2-2, por exemplo, a primeira classe de 200-209 contém 9 valores que se situam em algum ponto entre os limites de classe, mas não sabemos os valores

específicos desses 9 números. A fim de possibilitar os cálculos, supomos que todos os 9 valores se concentrem no ponto médio 204,5. Com 9 valores de 204,5, temos um total de $9 \times 204,5 = 1840,5$ que contribui para o total geral de todos os valores. O número de valores é igual à soma das freqüências, e assim podemos aplicar a Fórmula 2-2 para achar a média de uma tabela de freqüências. Na realidade, a Fórmula 2-2 não envolve um conceito fundamentalmente diferente; é apenas uma variante da Fórmula 2-1.

$$\text{Fórmula 2-2 } \bar{x} = \frac{\sum(f \cdot x)}{\sum f} \text{ média de uma tabela de freqüências}$$

onde x = ponto médio da classe
 f = freqüência
 $\sum f = n$

As cargas axiais das latas de alumínio da Tabela de Freqüências 2-2 foram introduzidas na Tabela 2-5, onde aplicamos a Fórmula 2-2. (Podemos também calcular a média de uma tabela de freqüências com uma calculadora TI-83: Introduzimos os pontos médios em L1, introduzimos as freqüências em L2 e utilizamos STAT, CALC, e 1 = Var Stats e introduzimos então L1, L2.) Quando utilizamos a coleção original de dados para calcular a média diretamente, obtivemos o valor 267,1, de modo que o valor da média ponderada baseada na tabela de freqüências é apenas ligeiramente diferente.

Em certas situações, os valores têm graus de importância diferentes, o que nos leva a calcular uma **média ponderada**, que é uma média dos valores afetados de pesos diferentes. Em tais casos, calculamos a média ponderada atribuindo pesos diferentes aos diversos valores, como se vê na Fórmula 2-3.

$$\text{Fórmula 2-3 } \text{média ponderada: } \bar{x} = \frac{\sum(w \cdot x)}{\sum w}$$

Suponha, por exemplo, que queiramos a média de 5 notas de teste (85, 90, 75, 80, 95), com os quatro primeiros testes valendo 15% cada um, e o último valendo 40%. Basta atribuímos o peso 15 a cada uma das quatro primeiras notas, o peso 40 à última nota e calcularmos a média pela Fórmula 2-3, como segue:

$$\begin{aligned} \bar{x} &= \frac{\sum(w \cdot x)}{\sum w} \\ &= \frac{(15 \times 85) + (15 \times 90) + (15 \times 75) + (15 \times 80) + (40 \times 95)}{15 + 15 + 15 + 15 + 40} \\ &= \frac{8750}{100} = 87,5 \end{aligned}$$

TABELA 2-5 Determinação de $\sum f$ e $\sum (f \cdot x)$

Carga Axial	Freqüência f	Ponto Médio da Classe x	$f \cdot x$
200-209	9	204,5	1.840,5
210-219	3	214,5	643,5
220-229	5	224,5	1.122,5
230-239	4	234,5	938,0
240-249	4	244,5	978,0
250-259	14	254,5	3.563,0
260-269	32	264,5	8.464,0
270-279	52	274,5	14.274,0
280-289	38	284,5	10.811,0
290-299	14	294,5	4.123,0
Total	$\sum f = 175$		$\sum (f \cdot x) = 46.757,5$
$\bar{x} = \frac{\sum(f \cdot x)}{\sum f} = \frac{46.757,5}{175} = 267,2$			

Outro exemplo: As notas de provas podem ser calculadas atribuindo-se a cada conceito (literal) um certo número de pontos ($A = 4$, $B = 3$ etc.) e atribuindo-se então a cada número uma frequência igual ao número de horas de crédito. Um conceito C em um curso de 3 créditos seria equivalente a um ponto médio de classe 2 com frequência 3. Novamente aqui, podemos aplicar a Fórmula 2-3 para calcular esse tipo de média.

A Melhor Medida de Tendência Central

Vimos que, para os dados da Tabela 2-1, a média, a mediana, a moda e o ponto médio tinham os valores 267,1, 273,0, 268 e 248,5, respectivamente. Qual dessas medidas de tendência central é a melhor? Infelizmente, não há uma resposta única, porque não há critérios objetivos para determinar a medida mais representativa para todos os conjuntos de dados. As diversas medidas de tendência central têm diferentes vantagens e desvantagens, algumas das quais estão resumidas na Tabela 2-6. Uma vantagem importante da média é que leva em conta todos os valores, mas uma grande desvantagem é que às vezes pode ser seriamente afetada por alguns valores extremos. Essa desvantagem pode ser superada com o uso da média aparada, descrita no Exercício 25.

Assimetria

A comparação da média, mediana e moda pode nos dizer algo sobre a característica da assimetria, definida a seguir e ilustrada na Figura 2-8.

DEFINIÇÃO

Uma distribuição de dados é **assimétrica** quando não é simétrica, estendendo-se mais para um lado do que para o outro. (Uma distribuição de dados é **simétrica** quando a metade esquerda do seu histograma é aproximadamente a imagem-espelho da metade direita.)

Os dados assimétricos *para a esquerda* dizem-se **negativamente assimétricos**; a média e a mediana estão à esquerda da moda. Embora nem sempre previsíveis, os dados negativamente assimétricos têm em geral a média à esquerda da mediana. (Veja Figura 2-8(a).) Os dados assimétricos *para a direita* dizem-se **positivamente assimétricos**; a média e a mediana estão à direita da moda.

TABELA 2-6 Comparação entre Média, Mediana, Moda e Ponto Médio

Medida	Definição	Quão Frequente?	Existência	Leva em Conta todos os Valores?	Afetada pelos Valores Extremos?	Vantagens e Desvantagens
Média	$\bar{x} = \frac{\sum x}{n}$	"média" mais familiar	existe sempre	sim	sim	usada em todo este livro; funciona bem com muitos métodos estatísticos
Mediana	valor do meio	usada comumente	existe sempre	não	não	costuma ser uma boa escolha se há alguns valores extremos
Moda	valor mais frequente	usada às vezes	pode não existir; pode haver mais de uma moda	não	não	apropriada para dados ao nível nominal
Ponto médio	$\frac{\text{alto} + \text{baixo}}{2}$	raramente usada	existe sempre	não	sim	muito sensível a valores extremos

Comentários gerais:

- Para um conjunto de dados aproximadamente simétrico com uma moda, a média, a mediana, a moda e o ponto médio tendem a coincidir.
- Para um conjunto de dados obviamente assimétrico, convém levar em conta a média e a mediana.
- A média é relativamente *confiável*; ou seja, quando as amostras são extraídas da mesma população, as médias tendem a ser mais constantes do que outras medidas (constantes no sentido de que as médias amostrais extraídas da mesma população não variam tanto quanto as outras medidas).

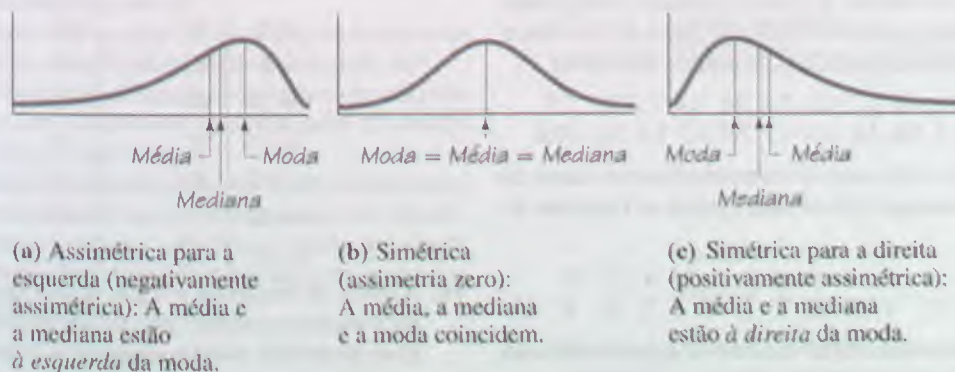


Fig. 2-8 Assimetria.

da moda. Novamente aqui, a maioria dos dados positivamente assimétricos tem a média à direita da mediana. (Veja Figura 2-8(c).)

Se examinarmos o histograma da Figura 2-1 para as cargas axiais de latas de alumínio que estamos considerando neste capítulo, veremos um gráfico que se apresenta assimétrico para a esquerda. Na prática, muitas distribuições de dados são simétricas. As distribuições assimétricas para a direita são mais comuns do que as assimétricas para a esquerda, porque em geral é mais fácil obter valores excepcionalmente grandes do que valores excepcionalmente pequenos. Com as rendas anuais, por exemplo, é impossível termos valores abaixo do limite inferior zero, mas há algumas pessoas que ganham milhões de dólares (ou reais) em um ano. As rendas anuais tendem, pois, a ser assimétricas para a direita, conforme a Figura 2-8(c).

2-4 Exercícios A: Habilidades e Conceitos Básicos

Nos Exercícios 1-4, determine (a) a média, (b) a mediana, (c) a moda e (d) o ponto médio.

- Os valores a seguir são os pesos (em onças) de bifes constantes do cardápio de um restaurante como "Bifes Porterhouse de 20 onças" (dados coletados por um aluno do autor). Supõe-se que o peso seja de 21 oz porque os filés perdem cerca de uma onça ao serem cozidos. Os pesos a seguir parecem razoáveis?

17 20 21 18 20 20 20 18 19 19
20 19 21 20 18 20 20 19 18 19

- Algarismos selecionados na Loteria Pick Three de Maryland:

0 7 3 6 2 7 6 6 6 3 8 1 7 8 7
1 6 8 6 9 5 2 1 5 0 3 9 9 0 7

- Depósitos de nitrato (em kg por hectare) como parte da chuva ácida no estado de Massachusetts de julho a setembro dos últimos anos (com base em dados do Ministério da Agricultura dos EUA):

6,40 5,21 4,66 5,24 6,96 5,53 8,23 6,80 5,78 6,00 5,41

- Concentrações sangue-álcool de 15 motoristas envolvidos em acidentes fatais e condenados à prisão (com base em dados do Ministério da Justiça dos EUA):

0,27 0,17 0,17 0,16 0,13 0,24 0,29 0,24
0,14 0,16 0,12 0,16 0,21 0,17 0,18

Nos Exercícios 5-8, determine a média, a mediana, a moda e o ponto médio de cada uma das duas amostras e compare os dois conjuntos de resultados.

- Tempos de espera de clientes no Banco Jefferson Valley (onde todos os clientes formam uma fila única) e no Banco de Providence (onde os clientes entram em três filas de guichês diferentes):

Jefferson Valley: 6,5 6,6 6,7 6,8 7,1 7,3 7,4 7,7 7,7 7,7

Providence: 4,2 5,4 5,8 6,2 6,7 7,7 7,7 8,5 9,3 10,0

- Amostras das idades (em anos) de carros de alunos e carros de professores e funcionários da faculdade, obtidas na faculdade do autor:

Alunos: 10 4 5 2 9 7 8 8 16 4 13 12

Prof. e Funcs.: 7 10 4 13 23 2 7 6 6 3 9 4

- Largura máxima de amostras de crânios de egípcios do sexo masculino, de 4000 aC a 150 aD (com base em dados de *Ancient Races of the Thebaid* por Thomson e Randall-Maciver):

4000 aC: 131 119 138 125 129 126 131 132 126 128 128 131
150 aD: 136 130 126 126 139 141 137 138 133 131 134 129

- Pesos (em libras) de papel e plástico descartado em residências durante uma semana (dados coletados para o Projeto do Lixo na Universidade do Arizona):

Papel: 9,55 6,38 2,80 6,98 6,33 6,16 10,00 12,29

Plástico: 2,19 2,10 1,41 0,63 0,92 1,40 1,74 2,87

Nos Exercícios 9-12, recorra ao conjunto de dados do Apêndice B e determine (a) a média, (b) a mediana, (c) a moda e (d) o ponto médio.

- Conjunto de Dados 2 do Apêndice B: Temperaturas do corpo às 8 horas da manhã no dia 1
- Conjunto de Dados 4 do Apêndice B: Conteúdo de nicotina de todos os cigarros relacionados
- Conjunto de Dados 3 do Apêndice B: Pesos dos ursos
- Conjunto de Dados 11 do Apêndice B: Pesos dos bombons M&M vermelhos.

Nos Exercícios 13-16, ache a média dos dados resumidos na tabela de frequências dada.

- Os visitantes do Parque Nacional de Yellowstone consideram uma erupção do Old Faithful uma atração que não deve ser perdida. A tabela de frequências resume uma amostra de tempos (em minutos) decorridos entre as erupções.

Tempo	Frequência
40-49	8
50-59	44
60-69	23
70-79	6
80-89	107
90-99	11
100-109	1

- Na faculdade do autor obtiveram-se amostras de carros de estudantes e carros dos professores e funcionários da faculdade, com as respectivas idades (em anos). Essas idades estão resumidas na tabela de frequência a seguir. Ache a idade média de ambos os grupos de carros. Com base nos resultados, percebe-se alguma diferença significativa entre as duas amostras? Em caso afirmativo, quais são elas?

Idade	Estudantes	Prof. e Funcs.
0-2	23	30
3-5	33	47
6-8	63	36
9-11	68	30
12-14	19	8
15-17	10	0
18-20	1	0
21-23	0	1

- A tabela de frequência a seguir dá as velocidades desenvolvidas por motoristas multados pela polícia da cidade de Poughkeepsie. Esses motoristas estavam dirigindo em uma zona de Creek Road com limite de velocidade de 30 mi/h. Compare a velocidade média observada com o limite de 30 mi/h.

Velocidade	Frequência
42-43	14
44-45	11
46-47	8
48-49	6
50-51	4
52-53	3
54-55	1
56-57	2
58-59	0
60-61	1

16. As companhias de seguro pesquisam continuamente as idades na morte e as respectivas causas. Os dados se baseiam em um estudo da revista *Time* sobre as mortes causadas por armas de fogo na América durante uma semana. Que podemos concluir do resultado?

Idade na morte	Frequência
16-25	22
26-35	10
36-45	6
46-55	2
56-65	4
66-75	5
76-85	1

2-4 Exercícios B: Além do Básico

17. Um estudante obtém as notas 60, 84 e 90 em testes, e 88 no exame final. Calcule a média ponderada das notas se cada teste corresponde a 20% e o exame final corresponde a 40% da nota final.
18. O boletim de um estudante acusa A em um curso de 4 créditos, A em um curso de 3 créditos, C em um curso de 3 créditos e D em um curso de 2 créditos. Atribuem-se pontos aos conceitos como segue: A = 4, B = 3, C = 2, D = 1, F = 0. Se as notas são ponderadas de acordo com as horas de crédito, determine a média ponderada arredondada para três decimais.
19. a. Calcule a média, a mediana, a moda e o ponto médio das seguintes rendas anuais (em dólares) de médicos autônomos (com base em dados da American Medical Association):
108.000 236.000 179.000 206.000 236.000
- b. Se se adiciona um valor constante k a cada renda, como são afetados os resultados da parte (a)?
- c. Se os valores das rendas na parte (a) são multiplicados por uma constante k , como são afetados os resultados da parte (a)?
- d. Às vezes os dados são transformados, substituindo-se cada valor x por $\log x$. Para os valores dados de x , determine se a média dos valores de $\log x$ é igual a $\log \bar{x}$.
20. A **média harmônica** costuma ser usada como medida de tendência central para conjuntos de dados que consistem em taxas de variação, como por exemplo velocidades. Obtém-se a média harmônica dividindo-se o número n de valores pela soma dos *inversos* de todos os valores. Expressa-se como:

$$\frac{n}{\sum \frac{1}{x}}$$

(Nenhum valor pode ser zero.) Por exemplo, a média harmônica de 2, 4, 10 é

$$\frac{n}{\sum \frac{1}{x}} = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{10}} = \frac{3}{0,85} = 3,5$$

- a. Quatro estudantes dirigem de Nova York à Flórida (1200 milhas) a uma velocidade de 40 mi/h (sim, é verdade!) e voltam à velocidade de 60 mi/h. Qual é sua velocidade média para a viagem de ida e volta? (Usa-se a média harmônica para calcular médias de velocidades.)
- b. Um despachante da Kramden Bus Company calcula a velocidade média, em mi/h, do percurso de ida e volta de Boston a Providence. Dão-se a seguir os resultados obtidos em 14 viagens diferentes. Com base nesses dados, qual é a velocidade média de um ônibus nesse percurso?

42,6	41,3	38,2	42,9	43,4	43,7	40,8
34,2	40,1	41,2	40,5	41,7	39,8	39,6

21. A **média geométrica** é usada em administração e economia para achar taxas médias de variação, de crescimento, ou razões médias. Dados n valores (todos positivos), a média geométrica é a raiz n^{ma} do seu produto. Por exemplo, determina-se a média geométrica de 2, 4, 10 multiplicando-se os três valores — o que dá 80, e tomando-se a raiz cúbica do resultado (porque há três valores). O resultado é 4,3. O *fator de crescimento médio* para o dinheiro, composto às taxas anuais de juro de 10%, 8%, 9%, 12% e 7% pode ser determinado calculando-se a média geométrica de 1,10, 1,08, 1,09, 1,12 e 1,07. Calcule esse fator médio de crescimento.
22. A **média quadrática** é utilizada em geral em experimentos físicos. Em sistemas de distribuição de energia, por exemplo, as tensões e correntes são em geral dadas em termos de sua média quadrática. Obtém-se a média quadrática de um conjunto de valores elevando-se cada um ao quadrado, somando-se os resultados, dividindo-se o total pelo número n de valores e tomando-se a raiz quadrada do resultado. Por exemplo, a média quadrática de 2, 4, 10 é

$$\sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{4 + 16 + 100}{3}} = \sqrt{\frac{120}{3}} = \sqrt{40} = 6,3$$

Calcule a média quadrática dos seguintes valores de fornecimento de energia (em volts): 151, 162, 0, 81, -68.

23. As tabelas de frequência costumam apresentar classes com intervalo aberto, como a tabela a seguir, que resume os tempos gastos em estudo por calouros (com base em dados de *The American Freshman*, em *USA Today*). Não se pode aplicar diretamente a Fórmula 2-2, porque o ponto médio da classe "mais de 20" não está definido. Calcule a média supondo que esta última classe seja realmente (a) 21-25, (b) 21-30, (c) 21-40. O que se pode concluir?

Horas de estudo por semana	Frequência
0	5
1-5	96
6-10	57
11-15	25
16-20	11
Mais de 20	6

24. Quando os dados são resumidos em uma tabela de frequências, pode-se achar a mediana identificando primeiro a *classe mediana* (a classe que contém a mediana). Supomos então que os valores se

distribuíam uniformemente nessa classe, e interpolamos. Esse processo é descrito por

$$\left(\frac{(n+1)}{2} - (m+1) \right) \frac{\text{amplitude da classe}}{\text{frequência da classe mediana}}$$

(limite inferior da classe mediana) +

onde n é a soma de todas as frequências de classe e m é a soma das frequências das classes que precedem a classe mediana. Utilize este processo e os dados da Tabela de Frequências 2-2 para achar a carga axial mediana.

25. Como a média é muito sensível a valores extremos, é acusada de não ser uma medida *robusta* de tendência central. A **média aparada** é mais robusta. Para achar a média aparada em 10% de um conjunto de dados, primeiro ordenamos os dados, em seguida eliminamos 10% dos valores superiores e 10% dos valores inferiores, e calculamos a média dos valores restantes. Para os pesos de ursos do Conjunto de Dados 3 do Apêndice B, determine (a) a média; (b) a média aparada em 10%; (c) a média aparada em 20%. Compare os resultados.
26. Consultando um almanaque, um pesquisador determina o salário médio dos professores para cada estado americano. Soma esses 50 valores e divide o total por 50, para obter a média. O resultado é igual ao salário médio nacional dos professores? Por quê?



2-5 Medidas de Variação

Esta seção aborda a característica da variação, de grande importância para a estatística, sendo, por isso, uma das principais de todo o livro. O leitor deve dominar os seguintes conceitos-chave: (1) a variação se refere a quanto os valores podem diferir entre si e pode ser medida por números específicos; (2) os números relativamente próximos uns dos outros têm baixas medidas de variação, enquanto os valores mais dispersos têm maior medida de variação; (3) o desvio-padrão é uma medida de variação particularmente importante, e devemos saber calculá-lo para um conjunto de valores; (4) os valores dos desvios-padrão devem ser interpretados corretamente.

Muitos bancos costumavam exigir que os clientes formassem filas separadas para os diversos guichês, mas recentemente passaram a adotar fila única. Qual o motivo dessa modificação? O tempo médio de espera não se modifica, porque a fila de espera não afeta a eficiência dos caixas. A adoção de fila única se deveu ao fato de os clientes preferirem tempos de espera mais *consistentes* com menor variação. Assim é que milhares de bancos efetuaram uma modificação que resultou em uma variação menor (e clientes mais satisfeitos), mesmo que a média não tenha sido afetada. Consideremos agora a mesma amostra de dados bancários usada no Exercício 5 da seção precedente. Os valores relacionados são tempos de espera (em minutos) de clientes.

Um Bom Conselho aos Jornalistas

O colunista Max Frankel escreveu no *The New York Times*: "As escolas de jornalismo não dão a devida importância à estatística, e algumas permitem que seus estudantes se formem sem qualquer treinamento com números. Como podem tais repórteres escrever conscientemente sobre comércio, bem-estar social, crime, ou tarifas aéreas, saúde e nutrição? O uso descuidado pela mídia de números sobre a incidência de

acidentes ou de doenças assusta o povo, deixando-o vulnerável aos truques jornalísticos, à demagogia política, e à fraude comercial." O colunista cita diversos casos, inclusive o exemplo de um artigo de página inteira sobre o déficit da cidade de Nova York, com uma promessa do prefeito daquela cidade de cobrir um déficit orçamentário de \$2,7 bilhões; mas em todo o artigo não se menciona uma vez sequer o *total* do orçamento, de modo que a cifra de \$2,7 bilhões por si só pouco significa.

Banco Jefferson Valley (Fila única)	6,5	6,6	6,7	6,8	7,1	7,3	7,4	7,7	7,7	7,7
Banco da Providência (Fila múltipla)	4,2	5,4	5,8	6,2	6,7	7,7	7,7	8,5	9,3	10,0

Os clientes do Jefferson Valley Bank entram em uma fila única que é atendida por três caixas. Os clientes do Bank of Providence podem entrar em qualquer uma de três filas que conduzem a três guichês. Se fizermos o Exercício 5 da Seção 2-4, veremos que ambos os bancos têm a mesma média de 7,15, a mesma mediana de 7,20, a mesma moda de 7,7 e o mesmo ponto médio de 7,10. Com base apenas nestas medidas de tendência central, poderíamos admitir que os tempos de espera nos dois bancos fossem praticamente os mesmos. Todavia, esquadrinhando os tempos de espera originais, constataríamos uma diferença fundamental: O Jefferson Valley Bank tem tempos de espera com muito menos *variação* do que o Bank of Providence. Mantidas todas as outras características, os clientes provavelmente preferirão o Jefferson Valley Bank, onde não correm o risco de entrar em uma fila muito mais lenta do que as outras.

Fazendo uma comparação subjetiva dos tempos de espera nos dois bancos, podemos ver a característica da variação. Passemos agora a algumas formas específicas de *medir* efetivamente a variação. Começaremos com a amplitude.

Amplitude

A **amplitude** de um conjunto de dados é a diferença entre o maior valor e o menor valor. Para calculá-lo, basta subtrairmos o menor valor do maior. Para o caso do Jefferson Valley Bank, a amplitude é de $7,7 - 6,5 = 1,2$ min. Os tempos de espera no Bank of Providence têm uma amplitude de 5,8 min, o que sugere maior variação.

O cálculo da amplitude é bastante fácil, mas como ele depende apenas do menor e do maior valor, em geral não é tão bom quanto outras medidas de variação que levam em conta todos os valores. (Veja no Exercício 25 um exemplo em que a amplitude é enganosa.)

Desvio-Padrão e Variância

De modo geral, o desvio-padrão é a mais importante e mais útil medida de variação. Ao contrário da amplitude, o desvio-padrão leva em conta todos os valores, mas essa vantagem torna o cálculo mais difícil. Definimos a seguir o desvio-padrão, mas para entender perfeitamente esse conceito, é preciso lermos cuidadosamente o restante desta seção.

DEFINIÇÃO

O **desvio-padrão** de um conjunto de valores amostrais é uma medida da variação dos valores em relação à média. Calcula-se com o auxílio da Fórmula 2-4.

Fórmula 2-4 $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$ desvio padrão amostral

Quase todas as calculadoras científicas e pacotes estatísticos são programados para calcular automaticamente o desvio-padrão. Na Seção 2-6 discutimos a utilização de calculadoras e computadores, mas é interessante o leitor consultar logo o manual de sua calculadora para ver o processo de cálculo que dá o desvio-padrão.

Por que definir uma medida de variação da maneira indicada na Fórmula 2-4? Ao medir a variação em um conjunto de dados amostrais, é razoável começarmos com os desvios dos valores em relação à média. Para determinado valor x , o valor do desvio é $x - \bar{x}$, que é a diferença entre o valor e a média. Mas a soma de todos esses desvios é sempre zero, o que na verdade nada significa para nós. Para termos uma estatística que realmente meça a variação (em lugar de ser sempre zero), poderíamos tomar a soma de valores absolutos, como em $\sum |x - \bar{x}|$. Determinando a média deste somatório, obtemos o **desvio médio** (ou **desvio absoluto**), dado pela seguinte expressão:

$$\text{Desvio médio} = \frac{\sum |x - \bar{x}|}{n}$$

Em vez de utilizar valores absolutos, podemos obter uma medida de variação ainda melhor, tomando os quadrados dos desvios ($x - \bar{x}$), que são não-negativos. Resulta que o desvio-padrão tem a mesma unidade de medida que os valores originais. Por exemplo, se os tempos de espera dos clientes são medidos em minutos, o desvio-padrão será expresso também em minutos. Com base na Fórmula 2-4, podemos estabelecer como se segue o processo de cálculo do desvio-padrão.

Processo para Determinar o Desvio-Padrão com a Fórmula 2-4

- Passo 1: Achar a média \bar{x} dos valores.
 Passo 2: Subtrair a média de cada valor individual ($x - \bar{x}$).
 Passo 3: Elevar ao quadrado cada uma das diferenças obtidas no Passo 2. [Este processo produz números da forma $(x - \bar{x})^2$.]
 Passo 4: Somar todos os quadrados obtidos no Passo 3, obtendo $\sum (x - \bar{x})^2$.
 Passo 5: Dividir o total do Passo 4 pelo número ($n - 1$); isto é, 1 menos que o número total de observações.
 Passo 6: Extrair a raiz quadrada do resultado do Passo 5.

Mais Ações, Menor Risco

Em seu livro *Investments*, os autores Zvi Bodie, Alex Kane e Alan Marcus afirmam que "o desvio-padrão médio dos ganhos proporcionados por uma carteira composta apenas de ações de uma única companhia é de 0,554. O risco médio de uma carteira diminui rapidamente na medida em que se diversificam as ações da carteira". Os autores observam que, com ações de 32 companhias, o desvio-padrão é de 0,325, indicando muito menor variação e risco. Salientam que com apenas uns poucos tipos de ações uma carteira tem um elevado grau de "risco específico", o que significa que o risco é atribuído ao pequeno número de ações em jogo. Com mais de 30 tipos de ação, há um risco específico muito pequeno; quase todo o risco é um "risco de mercado", atribuído ao mercado de ações como um todo. Os autores observam que esses princípios nada mais são do que a aplicação da bem conhecida lei das médias.

EXEMPLO Determine o desvio-padrão dos tempos de espera em guichês dos clientes do Jefferson Valley Bank. Esses tempos de espera (em minutos) são dados a seguir:

6,5 6,6 6,7 6,8 7,1 7,3 7,4 7,7 7,7 7,7

SOLUÇÃO Muitos estudantes acham fácil utilizar a função desvio-padrão embutida em suas calculadoras, mas recomendamos que o processo seja realmente entendido, seguindo os passos detalhados para o cálculo. (Ver Tabela 2-7, onde se executam os seguintes passos.)

Passo 1: Obtenha a média de 7,15, somando os valores e dividindo o total pelo número de valores:

$$\bar{x} = \frac{\sum x}{n} = \frac{71,5}{10} = 7,15 \text{ min}$$

Passo 2: Subtraia de cada valor a média 7,15, obtendo os seguintes valores de $(x - \bar{x})$: -0,65, -0,55, ..., 0,55.

Passo 3: Eleve ao quadrado cada valor do Passo 2, obtendo os valores $(x - \bar{x})^2$: 0,4225; 0,3025; ...; 0,3025.

Passo 4: Some todos os valores precedentes, obtendo

$$\sum (x - \bar{x})^2 = 2,0450$$

Passo 5: Há $n = 10$ valores; divida, pois, por 9 (= 10 - 1):

$$2,0450 \div 9 = 0,2272$$

Passo 6: Determine a raiz quadrada de 0,2272. O desvio-padrão é

$$\sqrt{0,2272} = 0,48 \text{ min}$$

Teoricamente, deveríamos dar aqui uma interpretação do desvio-padrão de 0,48 min, mas essa interpretação será dada mais adiante. Por ora, o leitor deve exercitar-se no cálculo de um desvio-padrão utilizando os tempos de espera no Bank of Providence. Com esses dados, verificará que o desvio-padrão é de 1,82 min. Embora a interpretação desses desvios-padrão seja dada mais adiante, podemos compará-los; verificaremos que o desvio-padrão dos tempos de espera no Jefferson Valley Bank (0,48 min) é muito menor do que o do caso do Bank of

TABELA 2-7 Cálculo do Desvio-Padrão para os Clientes do Banco Jefferson Valley

x	$x - \bar{x}$	$(x - \bar{x})^2$
6,5	-0,65	0,4225
6,6	-0,55	0,3025
6,7	-0,45	0,2025
6,8	-0,35	0,1225
7,1	-0,05	0,0025
7,3	0,15	0,0225
7,4	0,25	0,0625
7,7	0,55	0,3025
7,7	0,55	0,3025
7,7	0,55	0,3025
Totais: 71,5		2,0450
$\bar{x} = \frac{71,5}{10} = 7,15 \text{ min}$		
$s = \sqrt{\frac{2,0450}{10 - 1}} = \sqrt{0,2272} = 0,48 \text{ min}$		

Providence (1,82 min). Isso reforça a nossa conclusão subjetiva, de que os tempos de espera no Jefferson Valley Bank têm variação muito menor do que os do Bank of Providence.

Em nossa definição, referimo-nos ao desvio-padrão de dados amostrais. Para o cálculo do desvio-padrão σ (minúscula grega sigma) de uma população, vale uma fórmula ligeiramente diferente: em lugar de dividirmos por $n - 1$, dividimos por N , tamanho da população, como se vê na expressão seguinte.

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} \quad \text{desvio-padrão populacional}$$

Por exemplo, se os 10 valores da Tabela 2-7 constituem uma população, o desvio-padrão é:

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} = \sqrt{\frac{2,0450}{10}} = 0,45 \text{ min}$$

Como em geral lidamos com dados amostrais, vamos utilizar a Fórmula 2-4, dividindo por $n - 1$. Muitas calculadoras dão o desvio-padrão, com a divisão por $n - 1$ correspondendo a uma tecla σ_{n-1} ou s , enquanto que a tecla σ , ou σ corresponde a uma divisão por N . Por alguma razão, engenhosa mas estranha, as calculadoras utilizam diversas notações; as que seguem, entretanto, são as mais comuns em estatística. Essas notações compreendem referências à variância de um conjunto de valores; passamos agora a descrever essa medida de variação.

Notação

s denota o desvio-padrão de um conjunto de dados amostrais

σ denota o desvio-padrão de um conjunto de dados populacionais

s^2 é a variância de um conjunto de dados amostrais

σ^2 é a variância de um conjunto de dados populacionais

Nota: Em artigos de revistas e relatórios profissionais, costuma-se indicar o desvio-padrão por SD (*standard deviation*) e a variância por Var.

Omitindo a Etapa 6 (tomar a raiz quadrada) no processo de cálculo do desvio-padrão, obtemos a **variância**, definida na Fórmula 2-5.

$$\text{Fórmula 2-5} \quad s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \quad \text{variância amostral}$$

Analogamente, podemos expressar a variância populacional como

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} \quad \text{variância populacional}$$

Comparando as Fórmulas 2-4 e 2-5, vemos que a variância é o quadrado do desvio-padrão. Embora a variância venha a ser usada mais adiante, devemos inicialmente concentrar-nos no conceito de desvio-padrão, para bem apreender o significado dessa estatística. Uma dificuldade com a variância é que ela não é expressa nas mesmas unidades dos dados originais. Assim é que um conjunto de dados pode ter um desvio-padrão de \$3,00 e uma variância de 9,00 dólares quadrados. Como dólar quadrado é um conceito abstrato que não atingimos diretamente, a variância se nos afigura difícil de ser compreendida.

Regra do Arredondamento

Tal como na Seção 2-4, utilizamos a regra seguinte para arredondar resultados finais:

Tomar uma casa decimal a mais, em relação às que constam dos dados originais.

Devemos arredondar apenas o resultado final, e não resultados intermediários. Se, por alguma razão, tivermos de arredondar resultados intermediários, devemos trabalhar com pelo menos duas casas decimais além das que devem constar do resultado final.

Fórmula Abreviada e Dados Agrupados

Damos a seguir duas outras fórmulas para o desvio-padrão. Essas fórmulas não envolvem qualquer conceito diferente; são apenas versões distintas da Fórmula 2-4. Primeiro, a Fórmula 2-4 pode expressar-se na forma equivalente:

$$\text{Fórmula 2-6} \quad s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n - 1)}} \quad \text{Fórmula abreviada para o desvio padrão}$$

As Fórmulas 2-4 e 2-6 são equivalentes no sentido de que sempre dão os mesmos resultados. Pouparamos ao leitor o trabalho algébrico para mostrar essa igualdade. A Fórmula 2-6 é chamada *fórmula abreviada*, porque tende a ser mais conveniente para uso com números extensos ou com grandes conjuntos de valores. A Fórmula 2-6 é usada em geral em calculadoras e programas de computador, porque exige apenas três registros de memória (para n , $\sum x$ e $\sum x^2$), em lugar de um registro de memória separado para cada valor individual. A Fórmula 2-6 também elimina erros de arredondamentos intermediários, originados quando não se utiliza o valor exato da média. Não obstante, muitos professores preferem utilizar apenas a Fórmula 2-4 para o cálculo do desvio-padrão. Argumentam que a Fórmula 2-4 reforça o conceito de que o desvio-padrão é um tipo de desvio médio, enquanto a Fórmula 2-6 obscurece essa idéia. Outros professores não fazem qualquer objeção à Fórmula 2-6. Incluímos a fórmula abreviada para aqueles que desejem utilizá-la. Já apresentamos um exemplo de cálculo do desvio-padrão com a Fórmula 2-4; ilustraremos a seguir a aplicação da Fórmula 2-6.

EXEMPLO Calcule o desvio-padrão dos seguintes tempos de espera (em minutos) de clientes do Jefferson Valley Bank, aplicando a Fórmula 2-6:

6,5 6,6 6,7 6,8 7,1 7,3 7,4 7,7 7,7 7,7

SOLUÇÃO A Fórmula 2-6 exige a determinação dos valores de n , $\sum x$ e $\sum x^2$. Como há 10 valores, temos $n = 10$. A soma dos 10 valores é 71,5 e, assim, $\sum x = 71,5$. Calcula-se como se segue a terceira componente necessária:

$$\begin{aligned} \sum x^2 &= 6,5^2 + 6,6^2 + 6,7^2 + \dots + 7,7^2 \\ &= 42,25 + 43,56 + 44,89 + \dots + 59,29 \\ &= 513,27 \end{aligned}$$

Estamos em condições de aplicar a Fórmula 2-6 para calcular o valor do desvio-padrão.

$$\begin{aligned} s &= \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n - 1)}} = \sqrt{\frac{10(513,27) - (71,5)^2}{10(10 - 1)}} \\ &= \sqrt{\frac{20,45}{90}} = 0,4766783 = 0,48 \text{ min (arredondado)} \end{aligned}$$

Pode-se estabelecer uma fórmula para o desvio-padrão quando os dados se apresentam resumidos em uma tabela de frequências. O resultado é:

$$s = \sqrt{\frac{\sum f \cdot (x - \bar{x})^2}{n - 1}}$$

Daremos a esta fórmula uma expressão equivalente, que em geral simplifica os cálculos.

Fórmula 2-7

$$s = \sqrt{\frac{n[\sum (f \cdot x^2)] - [\sum (f \cdot x)]^2}{n(n - 1)}} \quad \text{desvio-padrão para tabela de frequência}$$

com: x = ponto médio da classe

f = frequência da classe

n = tamanho da amostra (ou $\sum f$ = soma das frequências)

EXEMPLO Aplique a Fórmula 2-7 para estimar o desvio-padrão das 175 cargas axiais das latas de alumínio da Tabela de Frequências 2-2.

SOLUÇÃO A aplicação da Fórmula 2-7 exige a determinação dos valores de n , $\sum (f \cdot x)$ e $\sum (f \cdot x^2)$. Determinados esses valores, pela Tabela 2-8, podemos aplicar a Fórmula 2-7, como segue:

$$\begin{aligned} s &= \sqrt{\frac{n[\sum (f \cdot x^2)] - [\sum (f \cdot x)]^2}{n(n - 1)}} = \\ &= \sqrt{\frac{175(12.579.173,75) - (46.757,5)^2}{175(175 - 1)}} = \\ &= \sqrt{\frac{15.091.600}{30.450}} = \sqrt{495,6190476} = 22,3 \text{ lb} \end{aligned}$$

As 175 cargas axiais têm um desvio-padrão estimado em 22,3 lb. (O valor exato calculado com base no conjunto original de dados é 22,1 lb; a aproximação é, pois, bastante satisfatória.)



Podemos também utilizar uma calculadora TI-83 para calcular o desvio-padrão de dados condensados em uma tabela de frequências. Introduzimos primeiro os pontos médios em L1, em seguida as frequências em L2; utilizamos então STAT, CALC e

1-VarStats e introduzimos L1 e L2 para obter os resultados que incluem a média e o desvio-padrão.

Para Entender o Desvio-padrão

Procuraremos aqui atribuir um sentido intuitivo ao desvio-padrão. De início, devemos ter em mente que o desvio-padrão mede a variação entre valores. Valores próximos uns dos outros originam desvios-padrão menores, enquanto valores muito afastados uns dos outros dão um desvio-padrão maior. Interrompamos a leitura e devotemos um momento ao estudo da Figura 2-9. Veremos que, quando os dados se dispersam, o valor do desvio-padrão aumenta.

Como a variação é um conceito relevante, e como o desvio-padrão tem grande importância na sua medida, abordaremos três maneiras diferentes de atribuir um sentido ao desvio-padrão. A primeira é uma regra prática que utiliza a amplitude para obter uma estimativa bastante rudimentar do desvio padrão. (Poderíamos melhorar a precisão dessa regra levando em conta fatores como o tamanho da amostra e a natureza da distribuição, mas, por ora, preferimos sacrificar a precisão em favor da simplicidade. Queremos uma regra simples que nos permita interpretar o valor do desvio-padrão; mais adiante estudaremos métodos que produzam resultados mais precisos.)

Regra Prática (desvio-padrão em termos da amplitude)

Para conjuntos de dados típicos, a amplitude mede aproximadamente 4 desvios-padrão ($4s$), de forma que podemos aproximar como segue o desvio-padrão:

$$\text{desvio-padrão} \approx \frac{\text{amplitude}}{4} \quad \text{regra prática}$$

Esta expressão dá uma estimativa razoável para o desvio-padrão, quando conhecemos os valores mínimo e máximo. Desde que conheçamos o desvio-padrão, podemos utilizá-lo para entender melhor os dados, fazendo estimativas dos valores mínimo e máximo como se segue:

$$\text{mínimo} \approx (\text{média}) - 2 \times (\text{desvio-padrão})$$

$$\text{máximo} \approx (\text{média}) + 2 \times (\text{desvio-padrão})$$

TABELA 2-8 Cálculo do Desvio-Padrão para uma Tabela de Frequências

Carga Axial	Frequência f	Ponto Médio da Classe x	$f \cdot x$	$f \cdot x^2$
200-209	9	204,5	1.840,5	376.382,25
210-219	3	214,5	643,5	138.030,75
220-229	5	224,5	1.122,5	252.001,25
230-239	4	234,5	938,0	219.961,00
240-249	4	244,5	978,0	239.121,00
250-259	14	254,5	3.563,0	906.783,50
260-269	32	264,5	8.464,0	2.238.728,00
270-279	52	274,5	14.274,0	3.918.213,00
280-289	38	284,5	10.811,0	3.075.729,50
290-299	14	294,5	4.123,0	1.214.223,50
Total	$\sum f = 175$		$\sum (f \cdot x) = 46.757,5$	$\sum (f \cdot x^2) = 12.579.173,75$

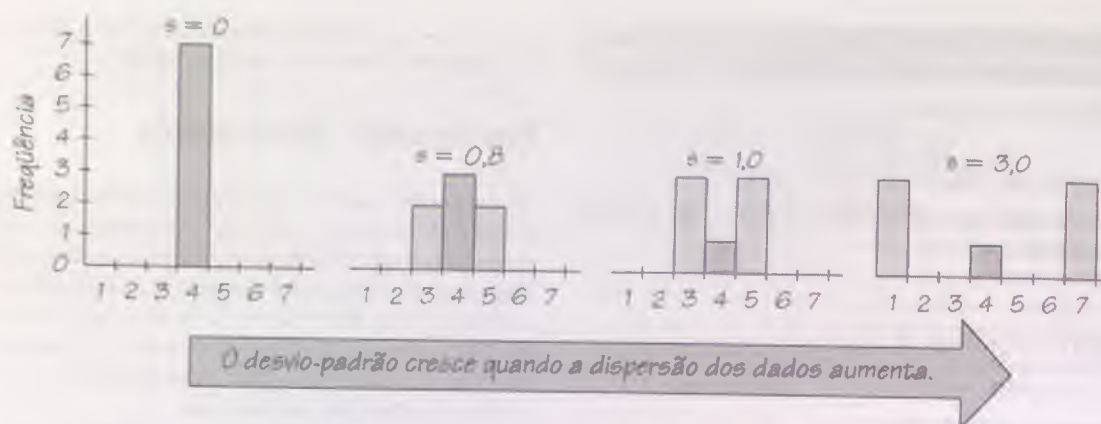


Fig. 2-9 Média idêntica, desvios-padrão diferentes.

Ao calcularmos um desvio-padrão com uma das Fórmulas 2-4 ou 2-6, podemos utilizar a regra prática como uma verificação do resultado obtido, mas não devemos esquecer que, embora a aproximação leve a uma vizinhança da resposta, ainda assim pode acusar grande diferença. Para os tempos de espera dos clientes do Jefferson Valley Bank (6,5; 6,6; 6,7; 6,8; 7,1; 7,3; 7,4; 7,7; 7,7; 7,7) calculamos o desvio-padrão pela Fórmula 2-6, obtendo $s = 0,48$. A amplitude desses valores é $7,7 - 6,5 = 1,2$, o que nos permite aplicar a regra prática para obter uma estimativa de s como segue:

$$s \approx \frac{\text{amplitude}}{4} = \frac{1,2}{4} = 0,3 \text{ min}$$

Ora, acabamos de ver que o desvio-padrão é realmente 0,48, de modo que a estimativa obtida pela regra prática (0,3) parece demasiadamente pequena. Todavia, nossa estimativa confirma que, de modo geral, estamos bem próximos do valor correto; sem dúvida, um valor como 7 para s se afiguraria incorreto.

Consistência no Correio

Pesquisa recente feita com 29.000 pessoas que utilizam o serviço postal dos EUA revelou que elas gostariam de maior consistência no tempo que uma carta leva para ser entregue. Ora, uma carta local pode levar um dia ou vários dias para ser entregue. O jornal USA Today registrou uma queixa comum: "Por favor, diga-me com quantos dias de antecedência eu devo postar um cartão de aniversário para minha mãe."

O nível de consistência pode ser medido pelo desvio-padrão dos tempos de entrega. Um desvio-padrão mais baixo revela maior consistência. O desvio-padrão é em geral um recurso criticamente importante para controlar a qualidade de bens e serviços.

EXEMPLO Com auxílio da regra prática, estime o desvio-padrão da amostra de 175 cargas axiais de latas de alumínio da Tabela 2-1.

SOLUÇÃO Utilizando a regra prática para estimar o desvio-padrão de dados amostrais, calculamos a amplitude e a dividimos por 4. Percorrendo a lista de valores, vemos que o menor é 200 e o maior é 297, de forma que a amplitude é $297 - 200 = 97$. O desvio-padrão s é estimado como segue:

$$s \approx \frac{\text{intervalo}}{4} = \frac{97}{4} = 24,3 \text{ lb}$$

Esse resultado está próximo do valor correto de 22,1, obtido com o cálculo do valor exato do desvio-padrão pela Fórmula 2-4 ou 2-6.

Como as cargas axiais das latas de alumínio da Tabela 2-1 têm uma média de 267,1, um desvio-padrão de 22,1 e uma distribuição como a da Figura 2-1, concluímos que essas latas podem facilmente suportar as pressões de 158 lb-165 lb aplicadas ao se fixarem as tampas no lugar. Recordemos, do enunciado do Problema do Capítulo, que essas latas têm uma espessura de 0,0109 in., que é inferior à espessura comumente adotada. Com base em nosso conhecimento das características importantes do conjunto de dados da Tabela 2-1, concluímos que é possível economizar utilizando essas latas menos espessas.

O exemplo precedente ilustra como utilizar dados sobre a amplitude, para estimar o desvio-padrão. O exemplo que se segue constitui uma ilustração particularmente importante de uma interpretação do desvio-padrão.

EXEMPLO A Gates Electronics Company fabrica barbeadores recarregáveis, sem fio, que têm vida média de 8,0 anos, com desvio-padrão de 3,0 anos. Utilizando a regra prática, estime a vida mais longa e a mais breve desses barbeadores.

SOLUÇÃO Estimamos a maior e a menor duração de vida pela regra prática, como se segue:

$$\begin{aligned} \text{mínimo} &\approx (\text{média}) - 2 \times (\text{desvio-padrão}) \\ &= 8,00 - 2(3,0) = 2,0 \text{ anos} \\ \text{máximo} &\approx (\text{média}) + 2 \times (\text{desvio-padrão}) \\ &= 8,0 + 2(3,0) = 14,0 \text{ anos} \end{aligned}$$

Podemos, pois, esperar que a maioria dos barbeadores em questão dure de 2,0 a 14,0 anos. Tenha em mente que esses resultados são estimativas grosseiras, mas, com o conhecimento da média e do desvio-padrão, estamos em condições de obter aproximações do menor e do maior valor, passando a entender melhor como os dados variam.

Regra Empírica (ou Regra 68-95-99) para os Dados

Outra regra que auxilia a interpretação do valor de um desvio-padrão é a **regra empírica**, aplicável somente a conjuntos de dados com distribuição aproximadamente em forma de sino, conforme a Figura 2-10. Essa figura mostra como a média e o desvio-padrão estão relacionados com a proporção dos dados que se enquadram em determinados

limites. Assim é que, com uma distribuição em forma de sino, temos 95% dos seus valores a menos de dois desvios-padrão da média. A regra empírica costuma ser designada abreviadamente como a **regra 68-95-99**.

A Regra 68-95-99 para Dados com Distribuição em Forma de Sino

- Cerca de 68% dos valores estão a menos de 1 desvio-padrão a contar da média.
- Cerca de 95% dos valores estão a menos de 2 desvios-padrão a contar da média.
- Cerca de 99,7% dos valores estão a menos de 3 desvios-padrão a contar da média.

EXEMPLO Os QIs de um grupo de adultos apresentam distribuição em forma de sino com média 100 e desvio-padrão 15. Aplique a regra empírica para achar a percentagem de adultos com QI entre 55 e 145.

SOLUÇÃO A chave para a resolução deste problema consiste em reconhecer que 55 e 145 estão, cada um, exatamente a três desvios-padrão da média. (Como o desvio-padrão é $s = 15$, decorre que $3s = 45$, de modo que 3 desvios-padrão abaixo da média são $100 - 45 = 55$, e 3 desvios-padrão acima da média são $100 + 45 = 145$.) A regra empírica afirma que 99,7% de todos os valores estão a menos de 3 desvios-padrão a contar da média, donde decorre que 99,7% dos adultos devem ter QI entre 55 e 145. Como os valores fora deste intervalo são bastante raros, uma pessoa com QI acima de 145 ou abaixo de 55 deve ser considerada excepcional.

Um terceiro conceito importante para compreendermos e interpretarmos o valor do desvio-padrão é o **teorema de Tchebichev**. A regra empírica precedente se aplica apenas a

conjuntos de dados com distribuição em forma de sino. O teorema de Tchebichev se aplica a qualquer conjunto de dados, mas seus resultados são muito aproximados.

Teorema de Tchebichev

A proporção (ou fração) de *qualquer* conjunto de dados a menos de K desvios-padrão a contar da média é sempre *ao menos* $1 - 1/K^2$, onde K é um número positivo maior do que 1. Para $K = 2$ e $K = 3$, temos os seguintes resultados específicos:

- Ao menos 3/4 (ou 75%) de todos os valores estão no intervalo que vai de 2 desvios-padrão abaixo da média a 2 desvios-padrão acima da média ($\bar{x} - 2s$ a $\bar{x} + 2s$).
- Ao menos 8/9 (ou 89%) de todos os valores estão no intervalo que vai de 3 desvios-padrão abaixo da média até 3 desvios-padrão acima da média ($\bar{x} - 3s$ a $\bar{x} + 3s$).

Utilizando valores de QI com média 100 e desvio-padrão 15, o teorema de Tchebichev afirma que ao menos 75% dos valores estarão entre 70 e 130, e ao menos 89% dos valores estarão entre 55 e 145.

Após o estudo desta seção, deve estar claro para o leitor que o desvio-padrão é uma medida da variação entre os valores. O leitor deve ainda estar em condições de calcular o desvio-padrão para um conjunto de dados, interpretar os valores do desvio-padrão e reconhecer que, para um conjunto típico, é raro um valor do mesmo diferir da média por mais de 2 ou 3 desvios-padrão.

2-5 Exercícios A: Habilidades e Conceitos Básicos

Nos Exercícios 1-4, determine a amplitude, a variância e o desvio-padrão do conjunto de dados. (Os dados são os mesmos utilizados na Seção 2-4, onde determinamos medidas de tendência central. Aqui, trata-se de medidas de variação.)

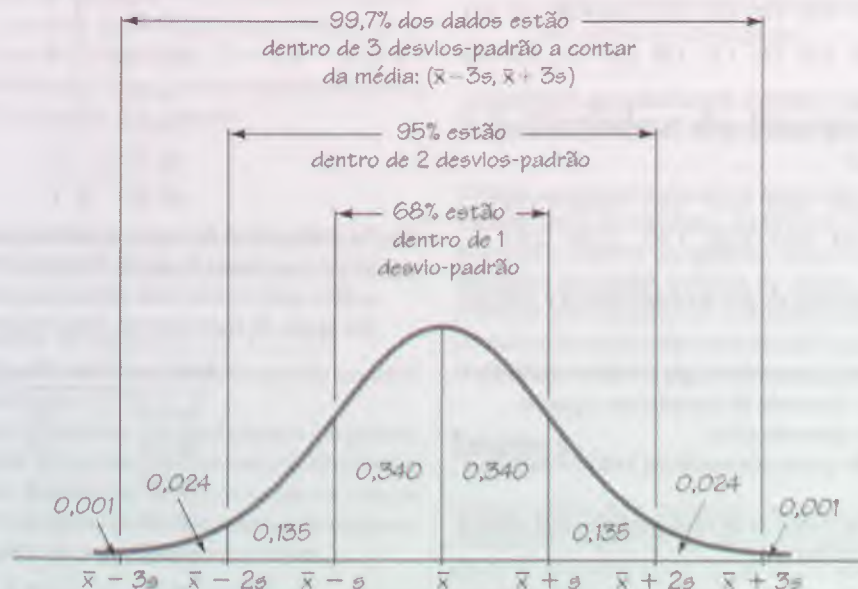


Fig. 2-10 A regra empírica.

1. Os valores a seguir são os pesos (em onças) de bifes constantes do cardápio de um restaurante como "bifes Porterhouse de 20 oz" (com base em dados coletados por um aluno do autor).

17 20 21 18 20 20 20 18 19 19
20 19 21 20 18 20 20 19 18 19

2. Algarismos escolhidos na loteria Pick Three de Maryland:

0 7 3 6 2 7 6 6 6 3 8 1 7 8 7
1 6 8 6 9 5 2 1 5 0 3 9 9 0 7

3. Resíduos de nitrato (em kg por hectare) como parte da chuva ácida em Massachusetts, de julho a setembro dos últimos anos (com base em dados do Ministério da Agricultura dos EUA):

6,40 5,21 4,66 5,24 6,96 5,53 8,23 6,80 5,78 6,00 5,41

4. Concentrações sangue-álcool de 15 motoristas envolvidos em acidentes fatais e condenados à prisão (Fonte: Dados do Ministério da Justiça dos EUA):

0,27 0,17 0,17 0,16 0,13 0,24 0,29 0,24
0,14 0,16 0,12 0,16 0,21 0,17 0,18

Nos Exercícios 5-8, determine a amplitude, a variância e o desvio-padrão para cada uma das duas amostras, e compare os dois conjuntos de resultados. (Na Seção 2-4 utilizamos esses mesmos dados.)

5. Tempos de espera de clientes no Jefferson Valley Bank (onde todos os clientes formam uma fila única) e no Bank of Providence (onde os clientes formam filas separadas para cada um dos três guichês). Esses conjuntos de dados já foram estudados nesta seção.

Jefferson Valley: 6,5 6,6 6,7 6,8 7,1 7,3 7,4 7,7 7,7 7,7
Providência: 4,2 5,4 5,8 6,2 6,7 7,7 7,7 8,5 9,3 10,0

6. Amostras das idades (em anos) de carros de alunos e de professores e funcionários de uma faculdade, obtidos na faculdade do autor.

Estudantes: 10 4 5 2 9 7 8 8 16 4 13 12
Profs. e func.: 7 10 4 13 23 2 7 6 6 3 9 4

7. Largura máxima de crânios de homens egípcios de 4000 aC a 150 aD (Fonte: Dados de *Ancient Races of the Thebaid*, por Thomson and Randall-Maciver):

4000 a.C.: 131 119 138 125 129 126 131 132 126 128 128 131
150 A.D.: 136 130 126 126 139 141 137 138 133 131 134 129

8. Pesos (em libras) de papel e plástico descartados em residências durante uma semana [Dados coletados no Projeto do Lixo da Universidade de Arizona]:

Papel: 9,55 6,38 2,80 6,98 6,33 6,16 10,00 12,29
Plástico: 2,19 2,10 1,41 0,63 0,92 1,40 1,74 2,87

Nos Exercícios 9-12, recorra aos dados do Apêndice B e calcule o desvio-padrão.

9. Conjunto 2, Apêndice B: temperaturas do corpo às 8 da manhã do dia 1
10. Conjunto 4, Apêndice B: conteúdo de nicotina em cigarros
11. Conjunto 3, Apêndice B: pesos de ursos
12. Conjunto 11, Apêndice B: pesos dos bombons M&M vermelhos

Nos Exercícios 13-16, determine o desvio-padrão dos dados resumidos na tabela de frequências.

13. Os visitantes do Parque Nacional de Yellowstone (EUA) consideraram uma erupção do gêiser Old Faithful uma atração imperdível.

A tabela de frequências resume os intervalos de tempo (em minutos) entre as erupções.

Tempo	Frequência
40-49	8
50-59	44
60-69	23
70-79	6
80-89	107
90-99	11
100-109	1

14. Dê-se a seguir, numa tabela de frequências, um resumo das idades de carros de alunos e de professores e funcionários da faculdade do autor. Determine o desvio-padrão de cada conjunto de dados. Com base nos resultados, há diferenças sensíveis entre as duas amostras? Em caso afirmativo, quais?

Idade	Estudantes	Professores/ Funcionários
0-2	23	30
3-5	33	47
6-8	63	36
9-11	68	30
12-14	19	8
15-17	10	0
18-20	1	0
21-23	0	1

15. A tabela de frequências a seguir dá as velocidades desenvolvidas por motoristas multados na cidade de Poughkeepsie em um trecho onde a velocidade máxima é de 30 mi/h.

Velocidade	Frequência
42-43	14
44-45	11
46-47	8
48-49	6
50-51	4
52-53	3
54-55	1
56-57	2
58-59	0
60-61	1

16. As companhias de seguro pesquisam continuamente as idades na morte e as causas de morte. Os dados se baseiam no estudo levado a efeito pela revista *Time* sobre as pessoas que morreram vitimadas por armas de fogo durante uma semana.

Idade na Morte	Frequência
16-25	22
26-35	10
36-45	6
46-55	2
56-65	4
66-75	5
76-85	1

17. Se o leitor vai comprar uma bateria para substituir a do seu carro, preferirá uma que venha de uma população com $\sigma = 1$ mês ou uma que venha de uma população com $\sigma = 1$ ano? (Suponha que ambas as populações tenham mesma média e mesmo preço.) Justifique sua escolha.
18. Como administrador, o leitor deve comprar lâmpadas para um hospital. Escolheria as lâmpadas Ultralight, que têm vida média $\mu = 3000$ h e $\sigma = 200$ h, ou as lâmpadas Electrolyte, com $\mu = 3000$ h e $\sigma = 250$ h? Explique.
19. Aplique a regra prática para estimar o desvio-padrão das alturas de seus colegas da turma de estatística.
20. Aplique a regra prática para estimar o desvio-padrão das notas do último exame final de estatística.

2-5 Exercícios B: Além do Básico

21. Um teste de datilografia acusa notas com $\bar{x} = 80,0$ e $s = 10,0$, e um histograma mostra que a distribuição das notas tem a forma aproximada de um sino. Aplique a regra empírica para responder:
 - a. Qual a porcentagem das notas entre 70 e 90?
 - b. Qual a porcentagem das notas a menos de 20 pontos da média?
 - c. Entre quais valores devem estar 99,7% das notas? (A média 80 deve estar a meio caminho entre esses dois valores.)
22. As alturas de mulheres adultas acusam média de 63,6 in. e desvio-padrão de 2,5 in. O que nos afirma o teorema de Tchebichev sobre as mulheres com altura entre 58,6 in. e 68,6 in.? Entre 56,1 in. e 77,1 in.?
23. a. Determine a amplitude e o desvio-padrão s da amostra seguinte de rendas (em dólares) de médicos autônomos (com base em dados da American Medical Association):

108.000	236.000	179.000	206.000	236.000
---------	---------	---------	---------	---------

 - b. Como são afetados os resultados da parte (a) se se adiciona um valor constante k a cada renda?
 - c. Se cada renda da parte (a) é multiplicada por uma constante k , como são afetados os resultados de (a)?
 - d. Por vezes, os dados são transformados, substituindo-se cada valor x por $\log x$. Para os valores dados de x , determine se o desvio-padrão dos valores de $\log x$ é igual a $\log s$.
 - e. Para os dados relativos a temperaturas do Conjunto 2 do Apêndice B (12 horas do dia 2), $\bar{x} = 98,20^\circ\text{F}$ e $s = 0,62^\circ\text{F}$. Determine \bar{x} e s para os dados, após transformar cada temperatura para a escala Celsius. [Sug.: $C = 5(F - 32)/9$.]
24. Se considerarmos os valores 1, 2, 3, ..., n como uma população, o desvio-padrão pode ser calculado pela fórmula

$$\sigma = \sqrt{\frac{n^2 - 1}{12}}$$

Esta fórmula é equivalente à Fórmula 2-4, modificada pela divisão por n em lugar de $n - 1$, onde o conjunto de dados consiste nos valores 1, 2, 3, ..., n .

- a. Calcule o desvio-padrão da população 1, 2, 3, ..., 100.
- b. Ache uma expressão para o cálculo do desvio-padrão amostral s para os valores amostrais 1, 2, 3, ..., n .
- c. Os computadores e as calculadoras em geral utilizam um gerador de números aleatórios que produz valores entre 0,00000000 e 0,99999999. Com o decorrer do processo, todos os valores tendem a ocorrer com a mesma frequência relativa. Determine a média e o desvio-padrão da população desses valores.
25. Dois grupos diferentes de uma turma de estatística fazem o mesmo teste-surpresa, com as notas relacionadas a seguir. Ache a amplitude

e o desvio-padrão para cada grupo. Que conclusões sobre a variação nos dois grupos os valores da amplitude sugerem? Por que razão a amplitude é enganosa neste caso? Que conclusões sobre a variação nos dois grupos o desvio-padrão sugere?

Grupo 1: 1 20 20 20 20 20 20 20 20 20 20

Grupo 2: 2 3 4 5 6 14 15 16 17 18 19

26. a. Utiliza-se o **coeficiente de variação**, expresso como porcentagem, para descrever o desvio-padrão em relação à média. Esse coeficiente permite-nos comparar a variabilidade de conjuntos de dados com diferentes unidades de medida (como pés versus minutos), e se calcula como se segue:

$$\frac{s}{\bar{x}} \cdot 100 \quad \text{ou} \quad \frac{\sigma}{\mu} \cdot 100$$

Determine o coeficiente de variação para as seguintes idades de carros (em anos):

0 1 3 3 5 6 6 6 6 8 12

- b. Genichi Taguchi desenvolveu um processo de melhoria de qualidade e redução de custo de fabricação mediante uma combinação de engenharia e estatística. Um elemento fundamental no processo de Taguchi é a **razão sinal-para-ruído**. A maneira mais simples de calcular essa razão consiste em dividir a média pelo desvio-padrão. Determine a razão sinal-para-ruído para os dados amostrais da parte (a).
27. Na Seção 2-4, introduzimos o conceito geral de assimetria. A assimetria pode ser medida pelo **índice de assimetria de Pearson**:

$$I = \frac{3(\bar{x} - \text{mediana})}{s}$$

Se $I \geq 1,00$ ou $I \leq -1,00$, os dados podem ser considerados **significativamente assimétricos**. Ache o índice de assimetria de Pearson para as cargas axiais de latas de alumínio da Tabela 2-1, e determine então se existe assimetria significativa.

28. a. Uma amostra consiste em 6 valores que se situam entre 1 e 9 inclusive. Qual o maior valor possível do desvio-padrão?
- b. Para qualquer conjunto de n valores com desvio-padrão s , todo valor deve estar a menos de $s\sqrt{n-1}$ da média. Uma professora de estatística afirma que as notas de um teste em sua turma de 17 alunos tiveram média 75,0 e desvio-padrão 5,0. Kelly, que se julga a melhor aluna da turma, alega ter obtido nota 97. Pode ser verdadeira tal alegação?

2-6 Medidas de Posição

Vamos agora introduzir os **escores z** , que permitem comparar valores mais facilmente, através de sua padronização. Introduziremos também os quartis, percentis e decis, com os quais podemos entender melhor os dados, focalizando sua posição relativa em relação ao conjunto como um todo. Os quartis introduzidos aqui serão também utilizados nos diagramas em caixa (*boxplots*), a serem abordados na seção seguinte.

Escores z

Quase todos nós estamos familiarizados com os QIs, e reconhecemos que um QI de 102 é bastante comum, enquanto um QI de 170 é raro. Esse QI de 102 é bastante comum porque está próximo da média de 100, mas o QI de 170 é raro porque está bem acima de 100. Esta circunstância pode sugerir uma

diferença entre os valores típicos e os valores raros, com base em sua diferença em relação à média ($x - \bar{x}$). Mas o vulto, ou tamanho, dessa diferença depende da escala que estamos utilizando. Com valores de QI, uma diferença de 2 pontos é insignificante, mas para médias de notas de uma faculdade uma diferença de 2 pontos entre 2,00 e 4,00 é altamente significativa, sobretudo para os pais dos alunos. Seria muito melhor se dispuséssemos de um padrão que não levasse em conta a escala utilizada. Com o valor, ou escore, padronizado, dividimos a diferença $x - \bar{x}$ (ou $x - \mu$) pelo desvio-padrão para chegarmos a esse resultado.

DEFINIÇÃO

O escore padronizado, ou escore z , é o número de desvios-padrão pelo qual um valor x dista da média (para mais ou para menos). Obtém-se como segue:

Amostra		População
$z = \frac{x - \bar{x}}{s}$	ou	$z = \frac{x - \mu}{\sigma}$

(Arredondar z para duas decimais.)

EXEMPLO As alturas da população de homens adultos têm média $\mu = 69,0$ in., desvio-padrão $\sigma = 2,8$ in. e distribuição em forma de sino. O jogador de basquete Michael Jordan ganhou reputação de gigante por suas proezas no jogo, mas com 78 in., ele pode ser considerado excepcionalmente alto, comparado com a população geral de homens adultos? Determine o escore z para a altura de 78 in.

SOLUÇÃO Como estamos lidando com parâmetros populacionais, o escore z se calcula como segue:

$$z = \frac{x - \mu}{\sigma} = \frac{78 - 69,0}{2,8} = 3,21$$

Podemos interpretar este resultado dizendo que a altura de Michael Jordan, de 78 in., está 3,21 desvios-padrão acima da média.

A importância dos escores z na estatística reside no fato de que eles permitem distinguir entre valores usuais e valores raros, ou incomuns. Consideramos usuais os valores cujos escores padronizados estão entre $-2,00$ e $2,00$, e incomuns os valores com escore z inferior a $-2,00$ ou superior a $2,00$. (Veja Figura 2-11.) A altura de Michael Jordan corresponde a um escore z de 3,21, que consideramos incomum, por ser superior a 2,00. Em comparação com a população geral, Jordan é excepcionalmente alto.

Nosso critério para classificar um escore z como incomum decorre da regra empírica e do teorema de Techebichev. Recorde que, pela regra empírica, para dados com distribuição em forma de sino, cerca de 95% dos valores estão a menos de 2 desvios-padrão da média. (Veja Figura 2-10 da seção precedente.) Por outro lado, o teorema de Techebichev afirma que, para qualquer conjunto de dados, ao menos 75% dos valores estão dentro de 2 desvios-padrão a contar da média.

Já vimos que os escores z são úteis para comparar escores de diferentes populações com médias distintas e desvios-padrão diferentes. O exemplo que segue ilustra essa aplicação dos escores z .

EXEMPLO Uma professora de estatística aplica dois testes diferentes a duas turmas do seu curso. Os resultados foram

$$\text{Turma 1: } \bar{x} = 75 \text{ e } s = 14$$

$$\text{Turma 2: } \bar{x} = 40 \text{ e } s = 8$$

Que nota é relativamente melhor: 82 no teste da Turma 1, ou 46 no da Turma 2?

SOLUÇÃO Não podemos comparar diretamente as notas 82 e 46 porque provêm de escalas diferentes. Transformamo-las, portanto, em escores z . Para o valor 82 da Turma 1, obtemos o escore z 0,50, porque

$$z = \frac{x - \bar{x}}{s} = \frac{82 - 75}{14} = 0,50$$

Para a nota 46 da Turma 2, o escore z correspondente é 0,75, porque

$$z = \frac{x - \bar{x}}{s} = \frac{46 - 40}{8} = 0,75$$

Isso significa que a nota 82 do teste da Turma 1 está 0,5 desvio-padrão acima da média, enquanto a nota 46 do teste da Turma 2 está 0,75 desvio-padrão acima da média. Isso implica que o resultado 46 do teste da Turma 2 é melhor, relativamente. Embora inferior a 82, a nota 46 tem melhor posição relativa no contexto dos outros resultados do teste. Mais adiante vamos utilizar amplamente os escores z .

Compra de Carro

Para a aquisição de um carro novo ou usado, uma boa referência é o grau de confiabilidade compilado e reportado pela revista *Consumer Reports*. Os dados relativos à frequência de consertos se baseiam em 10 milhões de dados coletados de milhares de leitores. Os estatísticos analisam os dados em busca de padrões que conduzam a listas de carros confiáveis e carros que devem ser evitados. A presidente da Consumers Union, Rhoda Karpatkin, escreve: "Já que os números têm tanta importância em nosso trabalho, não é de surpreender que os estatísticos representem a chave desse processo."

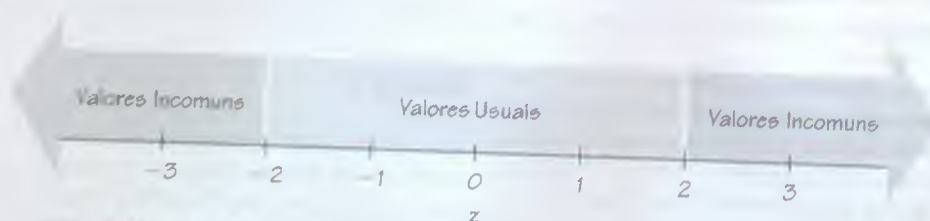


Fig. 2-11 Interpretação do escore z .

Valores com escores z inferiores a $z = -2,00$ ou superiores a $z = 2,00$ são considerados incomuns.

O exemplo precedente mostrou a eficácia dos escores z em medidas de comparação entre conjuntos diferentes de dados. Da mesma forma, os quartis, os decis e os percentis são medidas de posição convenientes para comparar valores dentro de um mesmo conjunto de dados, ou entre conjuntos diferentes.

Quartis, Decis e Percentis

Assim como a mediana divide os dados em duas partes iguais, os três **quartis**, denotados por Q_1 , Q_2 e Q_3 , dividem as observações ordenadas (dispostas em ordem crescente) em quatro partes iguais. Grosso modo, Q_1 separa os 25% inferiores dos 75% superiores dos valores ordenados; Q_2 é a mediana; e Q_3 separa os 75% inferiores dos 25% superiores dos dados. Mais precisamente, ao menos 25% dos dados serão no máximo iguais a Q_1 , e ao menos 75% dos dados serão no mínimo iguais a Q_1 . Ao menos 75% dos dados serão no máximo iguais a Q_3 , enquanto ao menos 25% serão, no mínimo, iguais a Q_3 .

Analogamente, há nove **decis**, denotados por $D_1, D_2, D_3, \dots, D_9$, que dividem os dados em 10 grupos com cerca de 10% deles em cada grupo. Há, finalmente, 99 **percentis**, que dividem os dados em 100 grupos com cerca de 1% em cada grupo. (Os quartis, decis e percentis são exemplos de *fractis*, que dividem os dados em partes aproximadamente iguais.) Um estudante que se submeteu ao vestibular para ingresso em uma faculdade é informado de que está no 92.º percentil. Isso não significa, entretanto, que ele tenha obtido 92% no exame; indica, apenas, que qualquer que tenha sido a nota obtida, ela foi superior a 92% (e inferior a 8%) das notas de toda a turma. O 92.º percentil é, pois, uma excelente classificação em relação aos outros que fizeram o exame.

O processo de determinação do percentil correspondente a um determinado valor x é bastante simples, como se pode ver na expressão seguinte.

$$\text{percentil do valor } x = \frac{\text{número de valores inferiores a } x}{\text{número total de valores}} \cdot 100$$

EXEMPLO A Tabela 2-9 relaciona as 175 cargas axiais das latas de alumínio, ordenadas da mais baixa até a mais elevada. Determine o percentil correspondente a 241.

SOLUÇÃO Pela Tabela 2-9, vemos que há 21 valores inferiores a 241, de forma que

$$\text{percentil de 241} = \frac{21}{175} \cdot 100 = 12$$

A carga axial de 241 é o 12.º percentil.

O exemplo precedente ilustra o processo de determinação do percentil correspondente a determinado valor. Para o processo inverso, há vários métodos diferentes para achar o valor correspondente a determinado percentil; o que vamos utilizar está esquematizado na Figura 2-12, em que é adotada a notação seguinte.

Notação

n	número de escores, ou valores, no conjunto de dados
k	percentil a ser utilizado
L	indicador que dá a posição de um escore
P_k	k^{mo} percentil

EXEMPLO Para as 175 cargas axiais de latas de alumínio da Tabela 2-9, determine o escore correspondente ao 25.º percentil; ou seja, determine o valor de P_{25} .

SOLUÇÃO Recorremos à Figura 2-12 e observamos que os dados já estão ordenados, do menor para o maior. Calculamos a seguir o indicador L como segue:

$$L = \left(\frac{k}{100} \right) n = \left(\frac{25}{100} \right) \cdot 175 = 43,75$$

Respondemos *não* à pergunta na Figura 2-12, se 43,75 é um número inteiro, e somos orientados a arredondar L para cima, ou seja, arredondar para 44. (Nesse processo em particular arredondamos L para o inteiro superior mais próximo, mas na maior parte das situações neste livro seguimos o processo geral de arredondamento.) O 25.º percentil, denotado por P_{25} , é o 44.º valor, ou escore, a contar do menor. Partindo, pois, do menor valor, 200, percorremos a lista até o 44.º valor, que é 262; assim, $P_{25} = 262$.

Suponha agora que queiramos achar o percentil correspondente a um escore de 262. Verificamos que há 41 valores abaixo de 262, não deixando de considerar cada valor individual, mesmo os que aparecem repetidos. Calculando o percentil correspondente a 262, obtemos $(41/175) \cdot 100 = 23$ (arredondado).

Custo do Riso

Há realmente um Índice de Custo do Riso (ICR) que leva em conta o custo de itens como óculos de Groucho Marx, entrada em clubes de comédia e 13 outros indicadores. Trata-se da mesma

TABELA 2-9 Valores Ordenados de Cargas Axiais de Latas de Alumínio

200	201	204	204	206	206	208	208	209	215	217	218	220	223	223
225	228	230	230	234	236	241	242	242	248	250	251	251	252	252
254	256	256	256	257	257	258	259	259	260	261	262	262	262	262
262	263	263	263	263	263	264	265	265	265	266	267	267	268	268
268	268	268	268	268	268	268	269	269	269	269	270	270	270	270
270	270	270	270	271	271	272	272	272	272	272	273	273	273	273
273	273	274	274	274	274	275	275	275	275	276	276	276	276	276
277	277	277	277	277	277	277	277	278	278	278	278	278	278	278
279	279	279	280	280	280	281	281	281	281	282	282	282	282	282
282	283	283	283	283	283	283	284	284	284	284	285	285	285	286
286	286	286	287	287	288	289	289	289	289	289	290	290	290	291
291	292	292	292	293	293	294	295	295	297					

abordagem básica usada para estabelecer o Índice de Preços ao Consumidor (IPC), que se baseia em uma média ponderada de bens e serviços adquiridos por um consumidor típico. Enquanto valores padronizados e percentis permitem comparar diferentes valores, eles ignoram o elemento tempo. Índices como ICR e IPC permitem-nos comparar o valor de uma variável com seu valor em uma época de referência. O valor de um índice é o valor atual, dividido pelo valor de referência e multiplicado por 100.

Há aqui uma pequena discrepância: no exemplo precedente encontramos 262 para o 25.º percentil, mas no processo inverso, 262 corresponde ao 23.º percentil. À medida que aumenta o número de dados, tais discrepâncias diminuem. Poderíamos eliminá-las utilizando um processo mais complicado, que aplica a interpolação em lugar do arredondamento.

Em razão do tamanho da amostra no exemplo precedente, o indicador L calculado foi inicialmente 43,75, valor que foi arredondado para 44, porque o valor original de L não era inteiro. No próximo exemplo ilustramos um caso em que o valor original de L é um número inteiro. Essa condição nos levará para o ramo direito no fluxograma da Figura 2-12.

EXEMPLO Determine o 40.º percentil P_{40} das cargas axiais da Tabela 2-9.

SOLUÇÃO Seguindo o processo delineado na Figura 2-12 e notando que os dados já estão ordenados do menor para o maior, calculamos

$$L = \left(\frac{k}{100}\right)n = \left(\frac{40}{100}\right) \cdot 175 = 70 \quad (\text{exatamente})$$

70 é um número inteiro, e a Figura 2-12 indica que P_{40} está a meio caminho entre os 70.º e 71.º valores. E como esses valores são ambos 269, concluímos que o 40.º percentil é 269.

Uma vez dominados os cálculos para os percentis, podemos seguir o mesmo processo para calcular os quartis e decis, levando em conta as relações indicadas na margem.

Utilizando essas relações, podemos ver que Q_1 é equivalente a P_{25} . Em um exemplo anterior, vimos que $P_{25} = 262$, e assim o primeiro quartil é $Q_1 = 262$. Se precisarmos achar o terceiro quartil, Q_3 , basta reformular o problema para determinar P_{75} e proceder como indicado na Figura 2-12.

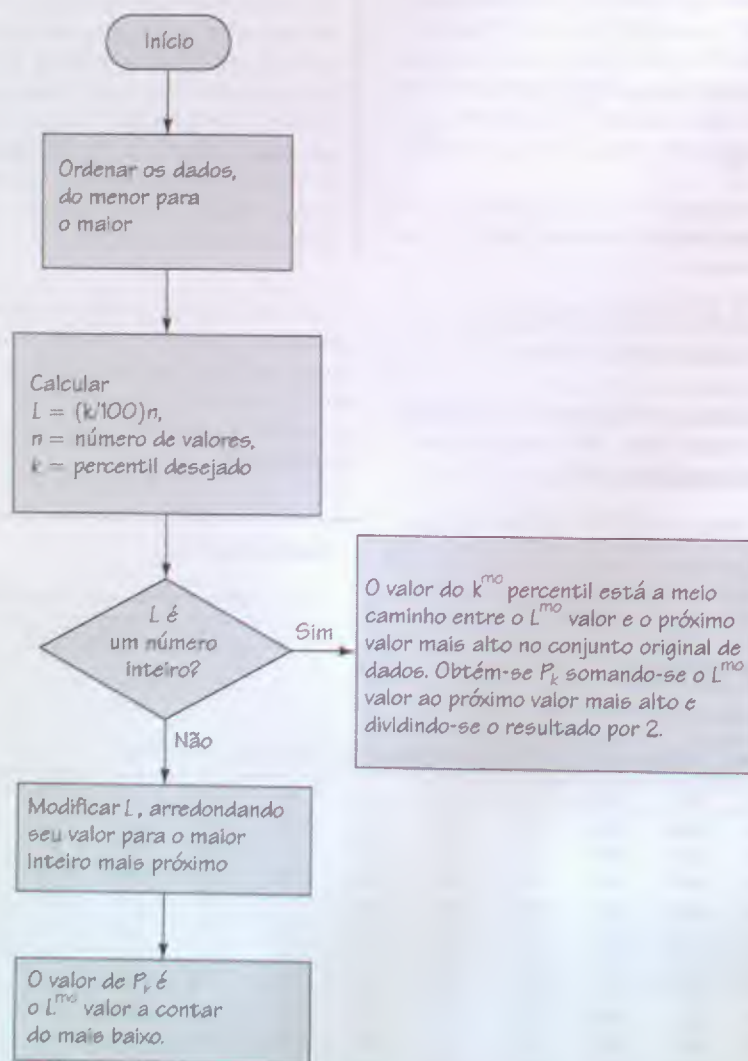


Fig. 2-12 Determinação do k -º percentil.

Quartis	Decis
$Q_1 = P_{25}$	$D_1 = P_{10}$
$Q_2 = P_{50}$	$D_2 = P_{20}$
$Q_3 = P_{75}$	\vdots
	$D_9 = P_{90}$

Além das medidas de tendência central e de variação já introduzidas, costumamos definir outras estatísticas utilizando quartis, decis ou percentis, como segue:

$$\begin{aligned}\text{intervalo interquartil} &= Q_3 - Q_1 \\ \text{intervalo semi-interquartil} &= (Q_3 - Q_1)/2 \\ \text{quartil médio} &= (Q_1 + Q_3)/2 \\ \text{amplitude de percentis 10-90} &= P_{90} - P_{10}\end{aligned}$$

Utilização de Calculadoras e Computadores na Estatística Descritiva

Ao lidarmos com grandes conjuntos de dados, é conveniente utilizarmos pacotes estatísticos a fim de obtermos resultados mais rápidos, fáceis e confiáveis. Os resultados que seguem, obtidos com STATDISK e Minitab, se baseiam nas 175 cargas axiais da Tabela 2-1; esses são exemplos de resultados que se obtêm quase com a mesma rapidez com que se introduzem os dados.

Podemos também utilizar as calculadoras para obter estatísticas descritivas. A maioria das calculadoras científicas dá pelo menos a média e o desvio-padrão. Com uma calculadora TI-83, devemos

utilizar STAT e Edit para introduzir um conjunto de dados em uma coluna, como L1; em seguida, aplicar STAT e CALC para obter a opção 1-Var Stats. Os resultados apresentados pela TI-83 incluem a média, a soma dos valores, a soma dos quadrados, o desvio-padrão, o número de valores (ou observações), o mínimo, o máximo, a mediana e os quartis. Como a TI-83 e o Minitab calculam os quartis de uma maneira ligeiramente diferente da adotada neste livro, pode haver algumas discrepâncias.

2-6 Exercícios A: Habilidades e Conceitos Básicos

Nos Exercícios 1-4, expresse todos os escores z com duas decimais.

- Os homens adultos (nos EUA) têm altura média de 69,0 polegadas, com desvio-padrão de 2,8 polegadas. Determine os escores z correspondentes a:
 - O jogador de basquete Mugsy Bogues que tem 5 pés e 3 in.
 - O jogador de basquete Shaquille O'Neal, que tem 7 pés e 1 polegada.
 - O autor, que é um jogador de golfe e tênis com 69,72 in.
- Os carros dos estudantes na faculdade do autor têm idade média de 7,90 anos, com desvio-padrão de 3,67 anos. Determine os escores z para os carros com as seguintes idades:
 - Um Corvette de 12 anos
 - Uma Ferrari de 2 anos
 - Um Porsche novo

STATDISK	
File	Edit Analysis Data Help
Cans109	
Sample Size, n	175
Mean, \bar{x}	267.11
Median	273.00
Midrange	248.50
RMS	268.02
Variance, s^2	488.95
St Dev, s	22.112
Mean Dev	16.019
Range	97.000
Minimum	200.00
1st Quartile	262.00
2nd Quartile	273.00
3rd Quartile	282.00
Maximum	297.00
$\sum x$	46745
$\sum x^2$	1257133

MINITAB						
Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
CANS109	175	267.11	273.00	269.15	22.11	1.67
Variable	Min	Max	Q1	Q3		
CANS109	200.00	297.00	262.00	282.00		

3. Os números de horas que os calouros passam estudando cada semana têm média de 7,06 h e desvio-padrão de 5,32 h (com base em dados de *The American Freshman*). Determine o escore z para um calouro que estuda 20 horas por semana.
4. Os tempos que os estudantes de curso secundário passam trabalhando em empregos cada semana têm média de 10,7 h e desvio-padrão de 11,2 h (com base em dados da National Federation of State High School Associations (Federação Nacional das Associações das Escolas Secundárias Estaduais)). Determine o escore z correspondente a um estudante que trabalha 8 horas por semana.

Nos Exercícios 5-8, expresse todos os escores z com duas decimais. Considere fora do comum um escore z inferior a $-2,00$ ou superior a $2,00$.

5. A admissão ao Beanstalk Club é limitada a mulheres e homens muito altos. A exigência de altura mínima para as mulheres é 70 in. As alturas das mulheres têm média de 63,6 in. e desvio-padrão de 2,5 in. Ache o escore z correspondente a uma mulher com 70 in. de altura e determine se se trata de uma altura fora do comum.
6. Uma mulher escreveu a *Dear Abby*, alegando ter dado à luz uma criança 308 dias após uma visita do seu marido, que estava servindo na Marinha. Os tempos de duração da gravidez acusam uma média de 268 dias, com desvio-padrão de 15 dias. Determine o escore z correspondente a 308 dias. Esse prazo pode ser considerado fora do comum? Que se pode concluir?
7. Certa máquina automática aceita moedas de 25 cents (de dólar) que não fujam ao padrão comum. Ache o escore z para uma moeda de 25 cents que pesa 5,50 g. Essa moeda será aceita pela máquina? (Os pesos das moedas de 25 cents têm média de 5,67 gramas, com desvio-padrão de 0,070 gramas.)
8. Para os homens com idades entre 18 e 24 anos, os níveis de colesterol (em mg/100 ml) têm média de 178,1 e desvio-padrão de 40,7 [com base em dados do National Health Survey (Serviço Nacional de Saúde dos EUA)]. Determine o escore z para um homem, com idade entre 18 e 24 anos, que tem um nível de colesterol de 275,2 mg/100 ml. Esse nível pode ser considerado excepcionalmente elevado?
9. Qual dos dois escores abaixo acusa melhor posição relativa?
 - a. Um escore de 60 em um teste com $\bar{x} = 50$ e $s = 5$
 - b. Um escore de 250 em um teste com $\bar{x} = 200$ e $s = 20$.
10. Dois grupos semelhantes de estudantes fazem testes equivalentes de facilidade de linguagem. Qual dos resultados seguintes indica maior facilidade relativa de linguagem?
 - a. Um escore de 65 em um teste com $\bar{x} = 70$ e $s = 10$
 - b. Um escore de 455 em um teste com $\bar{x} = 500$ e $s = 80$.
11. Três candidatos a um emprego fazem testes equivalentes de pensamento crítico. Qual dos escores abaixo corresponde à posição relativa mais elevada?
 - a. Um escore de 37 em um teste para o qual $\bar{x} = 28$ e $s = 6$
 - b. Um escore de 398 em um teste para o qual $\bar{x} = 312$ e $s = 56$
 - c. Um escore de 4,10 em um teste para o qual $\bar{x} = 2,75$ e $s = 0,92$
12. Três estudantes fazem testes equivalentes de senso de humor e, após terminada a risada, calculam-se seus escores. Qual é o escore relativo mais alto?
 - a. Um escore de 2,7 em um teste com $\bar{x} = 3,2$ e $s = 1,1$
 - b. Um escore de 27 em um teste em que $\bar{x} = 35$ e $s = 12$
 - c. Um escore de 850 em um teste em que $\bar{x} = 921$ e $s = 87$

Nos Exercícios 13-16, utilize as 175 cargas axiais ordenadas da Tabela 2-9. Ache o percentil correspondente ao valor dado.

13. 254 14. 265 15. 277 16. 288

Nos Exercícios 17-24, utilize as 175 cargas axiais da Tabela 2-9 para achar o percentil, quartil ou decil indicado.

17. P_{70} 18. P_{20} 19. D_6 20. D_3
21. Q_3 22. Q_1 23. D_1 24. P_1

Nos Exercícios 25-28, com base nos pesos (em libras) de ursos do Conjunto de Dados 3, do Apêndice B, determine o percentil correspondente ao peso indicado.

25. 144 26. 212 27. 316 28. 90

Nos Exercícios 29-36, com base nos pesos (em libras) de ursos do Conjunto de Dados 3, do Apêndice B, determine o percentil, o quartil ou o decil indicado.

29. P_{85} 30. P_{35} 31. Q_1 32. Q_3
33. D_9 34. D_1 35. P_{50} 36. P_{95}

2-6 Exercícios B: Além do Básico

37. Tome por base as cargas axiais ordenadas da Tabela 2-9.

- a. Determine o intervalo interquartil.
- b. Determine o quartil médio.
- c. Determine a amplitude de percentis 10-90.
- d. $P_{50} = Q_2$? Em caso afirmativo, isso ocorre sempre?
- e. $Q_2 = (Q_1 + Q_3)/2$? Em caso afirmativo, isto ocorre sempre?

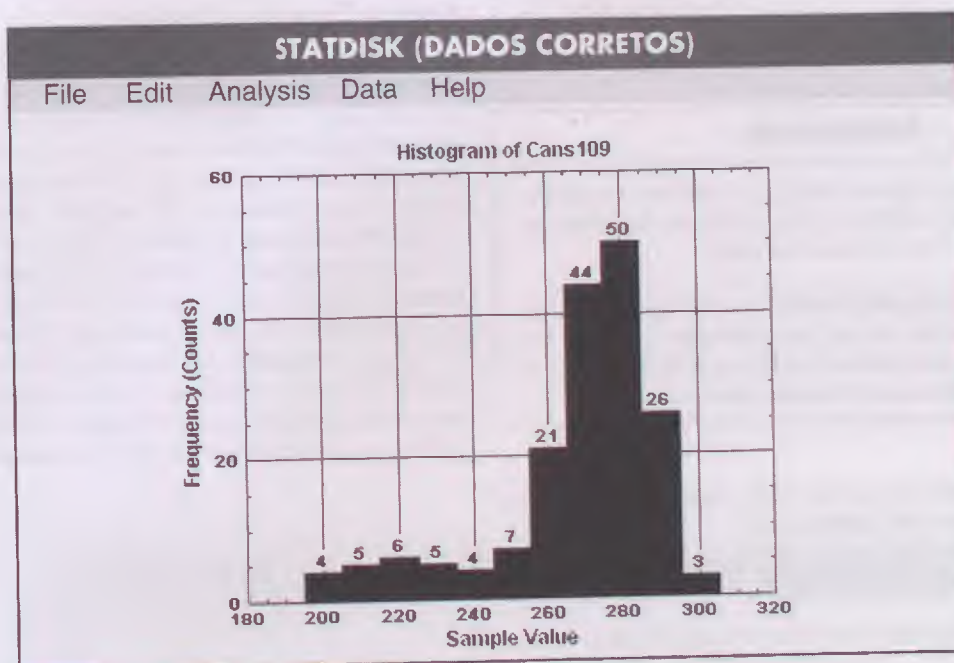
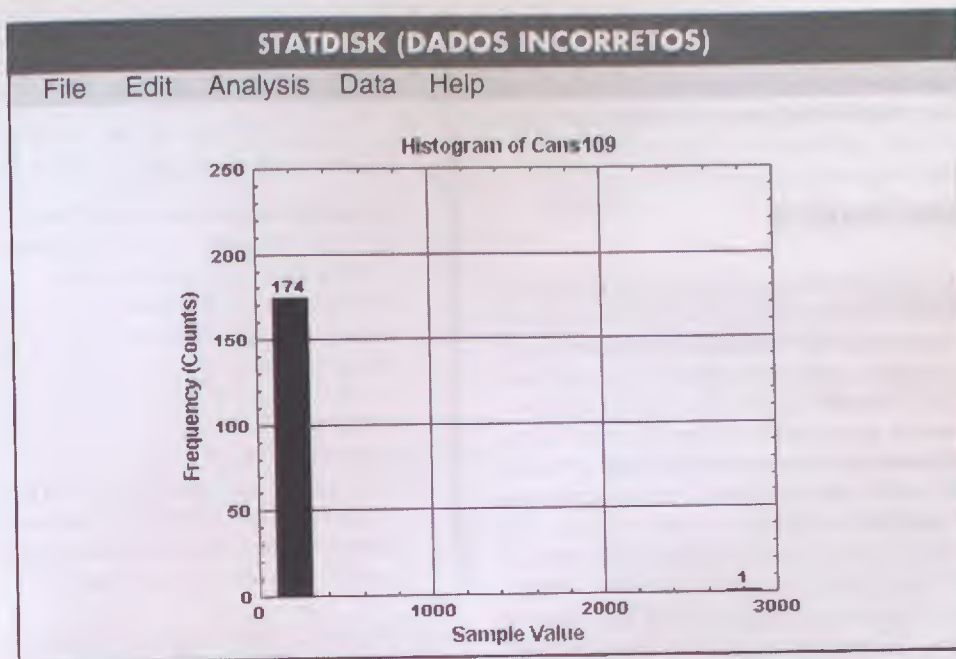
38. Ao determinar percentis utilizando a Figura 2-12, se o indicador L não é um número inteiro, arredondamo-lo para o maior inteiro mais próximo. Um processo alternativo consiste em interpolar, de modo que um indicador 23,75 conduza a um valor que está a 0,75 (ou $3/4$) no caminho entre os 23.º e 24.º escores. Utilize esse método de interpolação para calcular P_{35} , Q_1 e D_3 para os pesos relacionados no Conjunto de Dados 3 do Apêndice B.

39. Para as 175 cargas axiais das latas da Tabela 2-1, a média é 267,1 e o desvio-padrão é 22,1. Ache os dois valores fronteira que separam os valores ordinários dos valores incomuns.

40. Com os escores 2, 5, 8, 9 e 16, primeiro calcule \bar{x} e s ; em seguida, substitua cada valor pelo escore z correspondente. (Não arredonde os escores z ; tome tantas decimais quantas sua calculadora permitir.) Ache então a média e o desvio-padrão dos cinco escores z . Esses novos valores da média e do desvio-padrão serão obtidos para todo conjunto de escores z ?

2-7 Análise Exploratória de Dados (EDA)

Às vezes observamos ou coletamos dados com um objetivo específico em vista — por exemplo, verificar a eficiência de um novo tratamento de insônia. Outras vezes, não há qualquer objetivo específico; apenas desejamos explorar os dados para ver o que eles nos revelam. Na exploração de dados, podemos aplicar muitas das técnicas já apresentadas neste capítulo. Recorde que, na Seção 2-1, relacionamos três importantes características dos dados: (1) natureza ou forma da distribuição; (2) um valor representativo; e (3) uma medida de variação. É imprescindível considerar a distribuição dos dados, porque ela pode afetar não só os métodos estatísticos a ser usados, como também as conclusões a que chegarmos. No espírito da análise exploratória de dados, não devemos apenas visualizar o histograma e achar que entendemos a natureza da distribuição — é preciso explorar. A título de exemplo, mostramos dois histogramas obtidos com o



STATDISK das 175 cargas axiais da Tabela 2-1. O primeiro histograma representa os 175 valores com uma alteração: o primeiro valor, 270, é registrado incorretamente como 2700. O segundo histograma está correto. Note o efeito acentuado que um simples erro em um dos 175 valores tem sobre a forma do histograma. Nesse caso, o valor extremo, incorreto, de 2700 causa séria distorção no histograma. Em outros casos, tais valores extremos (chamados *outliers*) podem ser corretos, mas podem dar uma idéia errônea da verdadeira natureza da distribuição quando ilustrada por um histograma. Sem uma exploração mais aprofundada dos dados, podemos tirar conclusões seriamente errôneas dos histogramas.

Com EDA, dá-se ênfase à exploração original, com os objetivos de simplificar a descrição dos dados e obter uma visão

mais profunda da sua natureza. Adiante, nesta seção, fazemos uma comparação entre EDA e a estatística tradicional em três áreas principais da estatística.

Análise Exploratória de Dados

Explora os dados em um nível preliminar
Poucas (ou talvez nenhuma) hipóteses são feitas sobre os dados
Costuma exigir cálculos e gráficos relativamente simples

Estatística Tradicional

Confirma conclusões finais sobre os dados
Tipicamente, exige hipóteses muito importantes sobre os dados
Em geral, os cálculos são complexos e os gráficos desnecessários.

Na Seção 2-3, estudamos os gráficos do tipo ramo-e-folhas, um dos instrumentos comumente utilizados em EDA. Introduziremos agora os diagramas em caixa (*boxplots*) que não foram abordados antes porque exigem quartis só estudados na seção precedente.

Diagramas em Caixa (Boxplots)

Os diagramas em caixa são convenientes para revelar tendências centrais, dispersão, distribuição dos dados e a presença de *outliers* (valores extremos). A construção de um diagrama em caixa exige que tenhamos o valor mínimo, o primeiro quartil Q_1 , a mediana (ou segundo quartil Q_2), o terceiro quartil Q_3 e o valor máximo. Como as medianas revelam uma tendência central, ao passo que os quartis indicam a dispersão dos dados, os diagramas em caixa têm a vantagem de não serem tão sensíveis a valores extremos como outras medidas baseadas na média e no desvio-padrão. Por outro lado, os diagramas em caixa (*boxplots*) não dão informação tão detalhada quanto os histogramas ou os gráficos ramo-e-folhas, podendo não ser, assim, a melhor escolha quando lidamos com um único conjunto de dados. Os diagramas em caixa são, entretanto, mais convenientes na comparação de dois ou mais conjuntos de dados. Ao utilizarmos dois ou mais diagramas em caixa para comparar diferentes conjuntos de dados, é importante utilizarmos a mesma escala, de forma a possibilitar a comparação.



DEFINIÇÕES

O valor mínimo, o primeiro quartil Q_1 , a mediana, o terceiro quartil Q_3 e o valor máximo constituem um **resumo de cinco números** de um conjunto de dados.

Um **diagrama em caixas (boxplot)** é um gráfico de dados que consiste em uma reta que se prolonga do menor ao maior valor, e um retângulo com retas traçadas no primeiro quartil Q_1 , na mediana e no terceiro quartil Q_3 .

EXEMPLO Com base nos dados sobre pulsação de fumantes (Conjunto de Dados 8 do Apêndice B),

- Determine os valores que constituem o resumo de 5 números.
- Construa um diagrama em caixa para esses valores.

SOLUÇÃO

- O resumo de cinco números consiste no mínimo, Q_1 , mediana, Q_3 e no máximo. Para determinar esses valores, de-

vemos primeiro ordenar os dados do menor para o maior. Segue a lista ordenada dos 22 valores de pulsação de fumantes (Conjunto de Dados 8):

52 52 60 60 60 60 63 63 66 67 68
69 71 72 73 75 78 80 82 83 88 90

Nesta lista ordenada, é fácil identificar o mínimo 52 e o máximo 90. Com auxílio do fluxograma da Figura 2-12, vemos que o primeiro quartil Q_1 (ou P_{25}) é 60, que localizamos calculando $L = (25/100)22 = 5,5$, arredondado para 6. Q_1 é o sexto valor na lista ordenada, a saber, 60. A mediana é 68,5, que é o valor a meio caminho entre os 11.º e 12.º valores. Vemos também que $Q_3 = 78$, procurando na Figura 2-12 o 75.º percentil. O resumo de 5 números é, pois, 52, 60, 68,5, 78 e 90.

- Na Figura 2-13 temos o diagrama em caixas para os dados. Utilizemos o mínimo (52) e o máximo (90) para determinar uma escala de valores, e a seguir marcamos os valores com base no resumo de cinco números.

Na Figura 2-14 exibimos alguns diagramas em caixas genéricos, juntamente com as formas usuais de distribuição.

Valores Extremos (Outliers)

No decorrer da determinação de um resumo de 5 números e da construção de um diagrama em caixas, torna-se fácil identificar *outliers* (ou valores extremos), que são valores extremamente raros, no sentido de que estão muito afastados da maioria dos dados. Ao explorarmos um conjunto de dados, não podemos deixar de considerar os *outliers*, porque eles podem revelar informações importantes. Consideremos, por exemplo, a lista completa de pulsações do Conjunto de Dados 8. Basta ordenarmos os valores para ver que os valores 8 e 15 são *outliers*. Tratam-se de valores realmente excepcionais ou são valores errados? Embora haja alguns estudantes cujas condições físicas podem ser descritas como letárgicas, é extremamente improvável que alguém com uma pulsação de 8 ou 15 seja capaz de entrar em uma

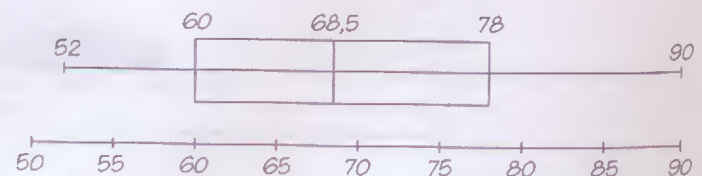


Fig. 2-13 Diagrama em caixas de pulsações (batidas por minuto) de fumantes.

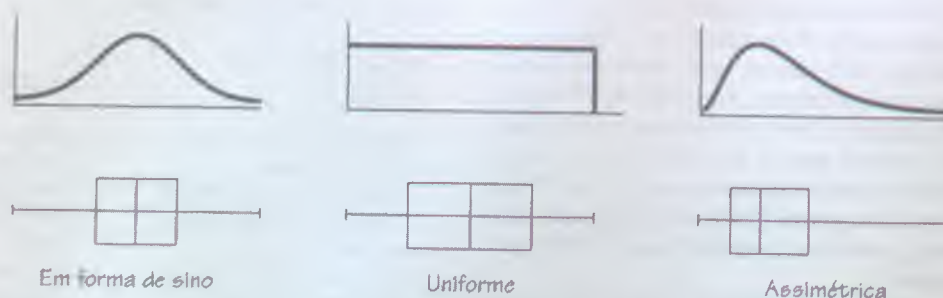


Fig. 2-14 Diagramas em caixas correspondentes às distribuições em forma de sino, uniforme e assimétrica.

sala de aula e sair dela por suas próprias forças. Concluímos, portanto, que 8 e 15 são erros, justificando-se a eliminação desses valores do conjunto. Devemos eliminar também a pulsação de 100? Não, porque esse valor não está demasiadamente distante dos outros, e provavelmente se refere a alguém excitado por estar em uma aula de estatística. De modo geral, devemos eliminar os *outliers* quando eles decorrem de erros óbvios; mas freqüentemente eles representam anomalias interessantes que merecem estudo mais detalhado. Na verdade, para alguns conjuntos de dados, os *outliers* são a característica mais importante. Um estudo sobre ovos e colesterol incluiu um homem que tinha consumido vários ovos por dia durante muitos anos. Sua taxa de consumo de ovos representava um *outlier*, mas o aspecto importante da questão é que o excesso de ovos não pareceu afetar seu nível de colesterol, que se manteve na média. Ao explorarmos dados, podemos estudar os efeitos dos *outliers* construindo gráficos e calculando medidas com eles e sem eles. (Veja Exercício 12, para uma forma de representar os *outliers* em diagramas em caixas.)

Utilização de Computadores e Calculadoras para Diagramas em Caixas

Podemos utilizar STATDISK, Minitab e a calculadora TI83 para criar diagramas em caixas. Com STATDISK, escolhemos o item Data do menu e utilizamos Sample Editor para introduzir os dados; clicamos COPY, escolhemos Data/Boxplot e clicamos PASTE; finalmente acionamos Evaluate. Com Minitab utilizamos as opções de File/New Worksheet/Graph/Boxplot. Os valores dos quartis calculados por Minitab e pela TI-83 podem diferir dos obtidos com a aplicação da Figura 2-12, de forma que os diagramas em caixas podem se apresentar ligeiramente diferentes.

Vimos que os *boxplots*, ou *diagramas em caixas*, são úteis para comparar conjuntos de dados; a figura a seguir apresenta os diagramas em caixa para as pulsações de fumantes e não-fumantes (Conjunto de Dados 8 do Apêndice B), feitos usando Minitab. Os *outliers* 8 e 15 foram excluídos.

Comparando os dois gráficos Minitab, vemos que não há diferenças substanciais. Os não-fumantes têm mais valores extremos, mas as medianas parecem coincidir, e a dispersão dos dados também é aproximadamente a mesma. Para o grupo de estudantes que faz estatística, parece que não há diferenças dignas de nota entre a pulsação dos fumantes e a dos não-fumantes.

2-7 Exercícios A: Habilidades e Conceitos Básicos

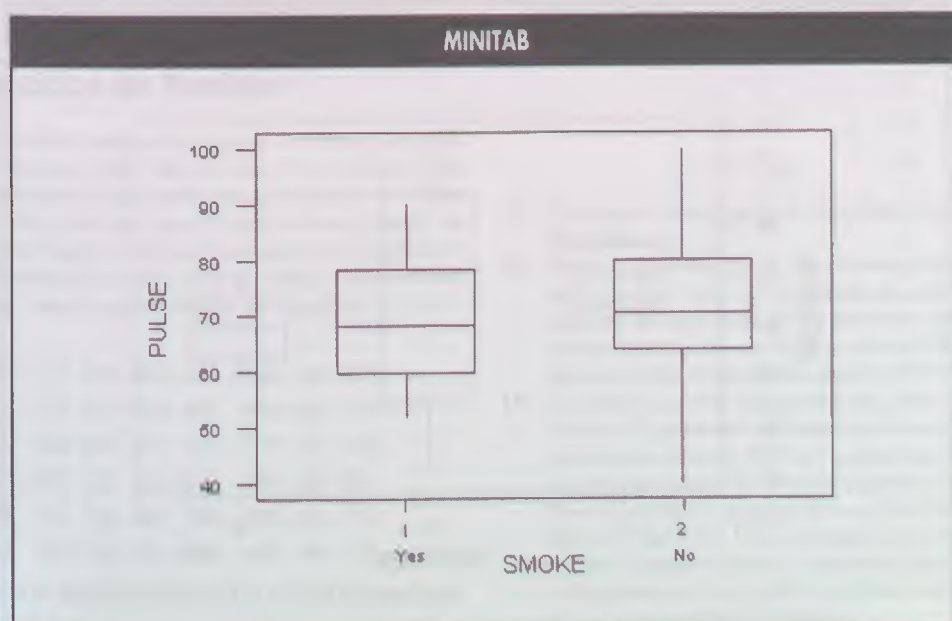
Inclua valores do resumo de 5 números em todos os diagramas em caixas.

1. Considere os dados do Conjunto 4 do Apêndice B e construa um diagrama em caixas para o conteúdo de nicotina de cigarros.
2. Com base nos dados do Conjunto 4 do Apêndice B, construa um diagrama em caixas para o conteúdo de alcatrão dos cigarros.
3. Em "Ages of Oscar-Winning Best Actors and Actresses" na revista *Mathematics Teacher*, por Richard Brown e Gretchen Davis, utilizam-se diagramas em caixas, ou *boxplots*, para comparar as idades dos atores e das atrizes na ocasião em que receberam o Oscar. Relacionam-se adiante os 34 vencedores recentes de cada categoria. Compare os dois conjuntos de dados com auxílio de um diagrama em caixas.



Atores: 32 37 36 32 51 53 33 61 35 45 55 39
76 37 42 40 32 60 38 56 48 48 40
43 62 43 42 44 41 56 39 46 31 47
Atrizes: 50 44 35 80 26 28 41 21 61 38 49 33
74 30 33 41 31 35 41 42 37 26 34
34 35 26 61 60 34 24 30 37 31 27

4. Considere o Conjunto 8 do Apêndice B para estes dois conjuntos de dados: pulsações dos fumantes e pulsações dos não-fumantes. Construa um diagrama em caixas para cada conjunto. Com base nos resultados, parece haver diferença de pulsação entre os dois grupos? Em caso afirmativo, quanto? É este o resultado esperado? (Exclua os valores 8 e 15, que devem ser erros.)

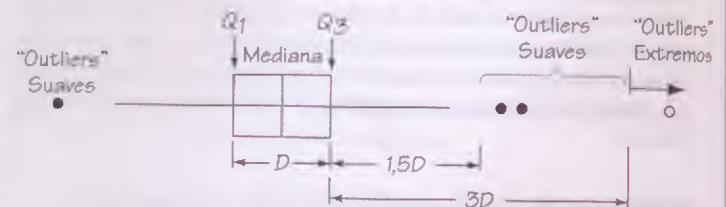


5. Considere o Conjunto 8 do Apêndice B para estes dois conjuntos de dados: taxas de pulsação para homens e para mulheres. Construa um diagrama em caixas para cada conjunto. Com base nos resultados, as taxas de pulsação dos dois conjuntos parecem ser diferentes? Em caso afirmativo, quanto? (Exclua os valores 8 e 15, que devem ser erros.)
6. Considere o Conjunto de Dados 10 do Apêndice B. Com auxílio de diagramas em caixas, compare os comprimentos dos filmes classificados R (restrito) com os dos filmes classificados não-R.
7. Considere o Conjunto de Dados 11 do Apêndice B. Com o auxílio de diagramas em caixas, compare os pesos dos bombons M&M vermelhos com os dos bombons M&M amarelos.
8. Considere o Conjunto de Dados 13 do Apêndice B. Construa um diagrama em caixas para os pesos das moedas de 25 cents. Compare a forma do gráfico resultante com as formas genéricas mostradas na Figura 2-14. Com base no diagrama em caixas, que podemos concluir sobre a natureza da distribuição?
9. Considere o Conjunto de Dados 12 do Apêndice B. Construa um diagrama em caixas para os 150 algarismos da Loteria "Pick Three" de Maryland. Compare a forma do gráfico resultante com as formas genéricas da Figura 2-14. Com base no gráfico, o resultado da loteria de Maryland parece estar de acordo com o resultado esperado?
10. Considere o Conjunto de Dados 1 do Apêndice B. Com auxílio de diagramas em caixas, compare os pesos do papel descartado com os pesos do plástico descartado.

- b. Desenhar normalmente a caixa com a mediana e os quartis, mas, ao prolongar as retas que se ramificam da caixa, caminhar apenas até os escores que estão a menos de $1,5D$ da mesma.
- c. Os **outliers suaves** são os valores que superam Q_3 em $1,5D$ a $3D$, ou estão $1,5D$ a $3D$ abaixo de Q_1 . Marque os **outliers suaves** com pontos cheios.
- d. Os **"outliers extremos"** são escores que excedem Q_3 em mais de $3D$ ou estão a mais de $3D$ abaixo de Q_1 . Marque os **outliers extremos** como pequenos círculos vazios.

A figura que acompanha é um exemplo do diagrama em caixas descrito aqui. Utilize esse processo para construir o diagrama em caixas para os valores dados, identificando os **outliers** extremos e suaves:

3 15 17 18 21 21 22 25 27 30 38 49 68



2-7 Exercícios B: Além do Básico

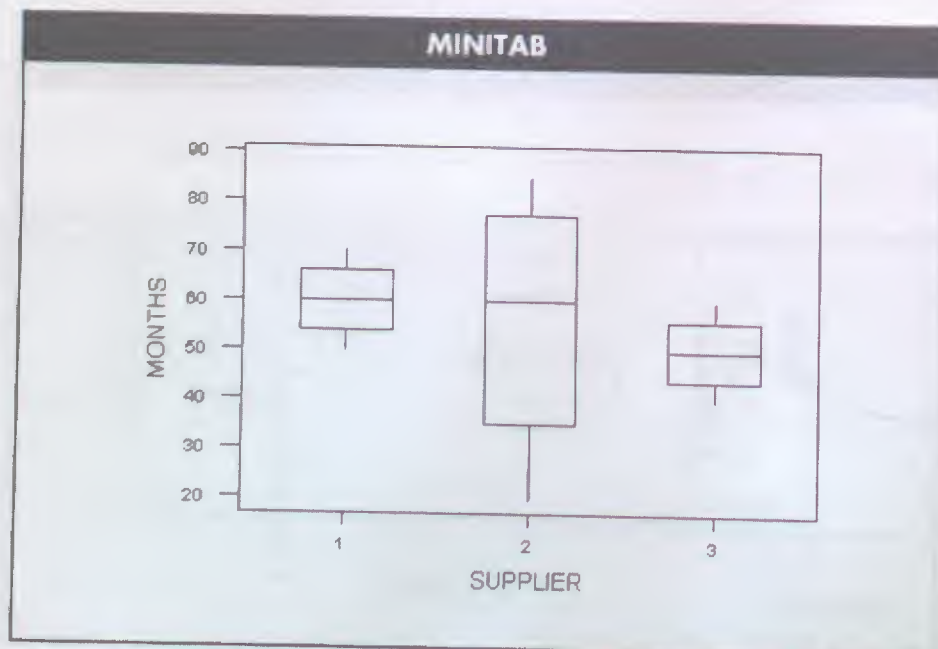
11. O supervisor de manutenção de uma frota de carros deve comprar baterias de substituição de um de três fornecedores. Para isto, testa a durabilidade de amostras de baterias desses três fornecedores, registrando as vidas (em meses), conforme resumo nos diagramas em caixa a seguir, obtidos com Minitab. Qual desses gráficos corresponde à marca que vai adquirir? Por quê?
12. Os diagramas em caixas, ou *boxplots*, discutidos nesta seção costumam chamar-se diagramas *esqueletais*. No estudo dos *outliers*, convém introduzir uma modificação na construção dos diagramas em caixas, como segue:
 - a. Calcular a diferença entre os quartis Q_3 e Q_1 , denotando-a por D : $D = Q_3 - Q_1$.



Vocabulário

estatística descritiva
estatística inferencial
tabela de frequência
frequência
limite inferior de classe
limite superior de classe
fronteiras de classe
pontos médios de classe
amplitude de classe
frequência relativa
tabela de frequência relativa
frequência acumulada
tabela de frequência acumulada
histograma

histograma de frequência relativa
gráfico por pontos
gráfico ramo-e-folhas
diagrama de Pareto
gráfico em setores
diagrama de dispersão
medida de tendência central
média aritmética
média
tamanho da amostra
mediana
moda
bimodal
multimodal



ponto médio	regra 68-95-99
média ponderada	teorema de Tchebichev
assimétrico	escore padronizado
simétrico	escore z
negativamente assimétrico	quartis
positivamente assimétrico	decis
amplitude	percentis
desvio-padrão	análise exploratória de dados (EDA)
desvio	resumo dos 5 números
desvio médio (ou absoluto)	diagrama em caixas
variância	boxplot
regra prática (desvio-padrão em termos da amplitude)	outlier
regra empírica	

Revisão

O Capítulo 2 abordou principalmente métodos e técnicas para resumir, descrever, explorar e comparar dados. Vimos as três características mais importantes dos dados, a saber, (1) natureza ou forma da distribuição, (2) valor representativo, e (3) medida de variação. Essas características podem ser estudadas e descritas com os recursos do Capítulo 2. Especificamente, para determinado conjunto de dados, devemos saber

- Resumir os dados, construindo uma tabela de frequências ou uma tabela de frequências relativas (Seção 2-2)
- Apresentar visualmente a natureza da distribuição, construindo um histograma, um gráfico por pontos, um ramo-e-folhas, um gráfico em setores, ou um diagrama de Pareto (Seção 2-3)
- Calcular medidas de tendência central: média, mediana, moda e ponto médio (Seção 2-4)
- Calcular medidas de variação: desvio-padrão, variância e amplitude (Seção 2-5)
- Comparar valores individuais, utilizando escores z, quartis, decis ou percentis (Seção 2-6)
- Investigar e explorar a dispersão de dados, o centro de dados e a amplitude de valores, com a construção de um diagrama em caixas, ou *boxplot* (Seção 2-7)

É preciso não só calcular as tabelas, gráficos e medidas, mas também compreender e interpretar esses resultados. Assim é que devemos entender com clareza que o desvio-padrão é uma medida da variação dos dados, e saber utilizá-lo para distinguir entre valores usuais e valores incomuns.

Exercícios de Revisão

1. A NCAA estava estudando meios de acelerar o término dos jogos universitários de basquetebol. Dão-se abaixo os tempos (em segundos) decorridos para jogar os dois últimos minutos do tempo regulamentar em 60 jogos das quatro primeiras rodadas do campeonato NCAA de basquetebol (com base em dados publicados no *USA Today*). Tomando o tempo mínimo como limite inferior da primeira classe, construa uma tabela de frequências com 9 classes.

756 587 929 871 378 503 564 1128 693 748
 448 670 1023 335 540 853 852 495 666 474
 443 325 514 404 820 915 793 778 627 483
 861 337 292 1070 625 457 676 494 420 862
 991 615 609 723 794 447 704 396 235 552
 626 688 506 700 240 363 860 670 396 345

2. Construa uma tabela de frequências relativas (com 9 classes) para os dados do Exercício 1.

3. Construa um histograma correspondente à tabela de frequências do Exercício 1.
4. Para os dados do Exercício 1, determine (a) Q_1 , (b) P_{45} , e (c) o percentil correspondente ao tempo de 335 s.
5. Aplique a regra prática para estimar o desvio-padrão dos dados do Exercício 1.
6. Utilize a tabela de frequências do Exercício 1 para achar a média e o desvio-padrão dos tempos.
7. Com os dados do Exercício 1, construa um gráfico ramo-e-folhas com 10 ramos.
8. Construa um diagrama em caixas (*boxplot*) para os dados do Exercício 1.
9. Dão-se a seguir os tempos (em segundos) decorridos entre a formulação do pedido e a entrega do prato em uma lanchonete McDonald's. Determine: (a) a média; (b) a mediana; (c) a moda; (d) o ponto médio; (e) a amplitude; (f) o desvio-padrão; (g) a variância:

135 90 85 121 83 69 87 159 177 135 227

10. Dão-se abaixo as idades de presidentes dos EUA na ocasião da posse. Calcule: (a) a média; (b) a mediana; (c) a moda; (d) o ponto médio; (e) o intervalo; (f) o desvio-padrão; (g) a variância; (h) Q_1 ; (i) P_{30} ; (j) D_7 .

57 61 57 57 58 57 61 54 68 51 49 64 50 48

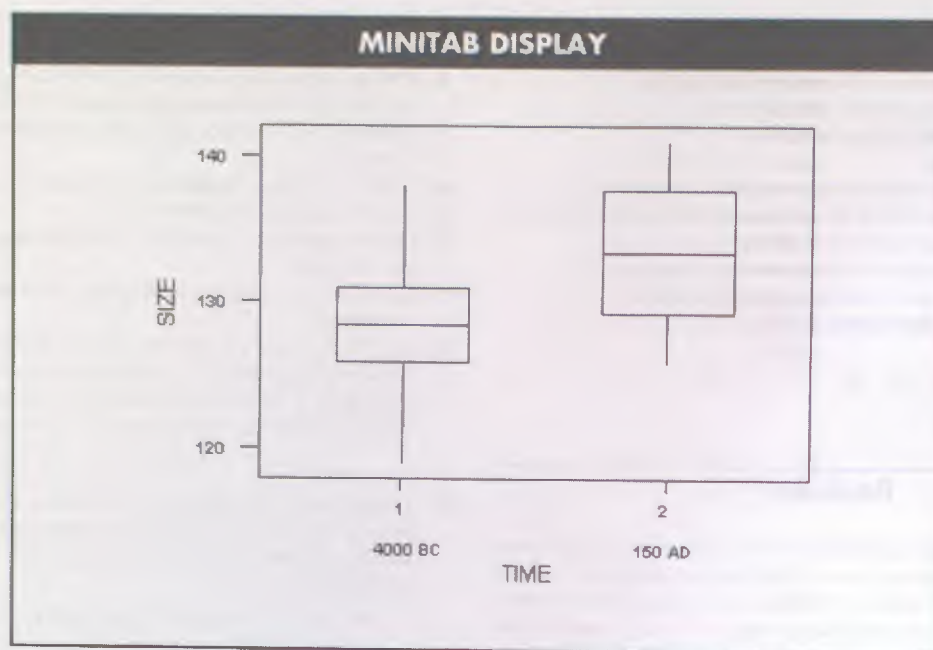
65 52 56 46 54 49 51 47 55 55 54 42 51 56

55 51 54 51 60 62 43 55 56 61 52 69 64 46

11. Os valores em um teste de percepção de profundidade acusam média 200 e desvio-padrão 40.
 - a. Um valor de 260 pode ser considerado excepcionalmente alto? Explique.
 - b. Qual o escore z correspondente a 185?
 - c. Supondo que os escores tenham uma distribuição em forma de sino, que nos informa a regra empírica sobre a porcentagem de escores entre 120 e 280?
 - d. Qual é a média, após adicionar 20 pontos a todos os escores?
 - e. Qual é o desvio-padrão na hipótese d?
12. A tabela a seguir dá os tempos (em anos) que os estudantes de certa faculdade levaram para obter o grau de bacharel (a partir de dados do National Center for Education Statistics). Com base na tabela, calcule a média e o desvio-padrão. Podemos considerar como incomum o fato de um estudante levar 8 anos para concluir o bacharelado? Explique.

Tempo (anos)	Número
4	147
5	81
6	27
7	15
7,5–11,5	30

13. Construa o histograma de frequências relativas para a tabela do Exercício 12.
14. Um psicólogo industrial deu a um empregado dois testes diferentes para medir o grau de satisfação no emprego. Qual resultado é melhor: um escore de 57 no primeiro teste, que teve média 72 e desvio-padrão 20, ou um escore de 450 no segundo teste, que acusou média 500 e desvio-padrão 80? Explique.
15. Considere os dois diagramas em caixas a seguir, obtidos com Minitab. O primeiro representa uma amostra de crânios de homens egípcios de cerca de 4000 a.C., enquanto o segundo representa uma amostra de crânios de homens egípcios de cerca de 150 a.D. (com base em dados de *Ancient Races of the Thebaid*, por Thomson and Randall-Maciver). Uma variação do tamanho das cabeças poderia sugerir mudanças sociais, como miscigenação com outras culturas. Comparando os dois gráficos, pode-se constatar variação na largura máxima dos crânios? Explique.



16. A Guarda Costeira dos EUA coletou dados sobre acidentes sérios com embarcações, categorizando-os conforme a seguir, com as respectivas frequências dadas entre parênteses. Construa um diagrama de Pareto resumindo os dados.

Colisão com outra embarcação (2203)

Colisão com um objeto fixo (839)

Encalhe (341)

Queda de pessoa no mar (431)

Soçobro (458)

estamos usando? (aleatória, estratificada, sistemática, por conglomerado, de conveniência)

- c. Faz-se uma pesquisa abordando todas as pessoas que saem da cabine eleitoral em 50 zonas eleitorais selecionadas aleatoriamente. Que tipo de amostragem estamos utilizando? (aleatória, estratificada, sistemática, por conglomerado, de conveniência)
3. Anualmente o Ministério da Energia dos EUA publica um *Annual Energy Review* que inclui o consumo de energia *per capita* (em milhões de Btu) para cada um dos 50 estados. Calculando-se a média desses 50 valores, o resultado é o consumo médio de energia *per capita* para todos os 50 estados combinados? Em caso negativo, explique como calcularia o consumo médio *per capita* para os 50 estados em conjunto.

Exercícios Cumulativos de Revisão

1. Dão-se a seguir os tempos (em horas) gastos em um dia com serviços de escritório por uma amostra de chefes de escritório (Fonte: Dados da Adia Personnel Services):

3,7 2,9 3,4 0,0 1,5 1,8 2,3 2,4 1,0 2,0

4,4 2,0 4,5 0,0 1,7 4,4 3,3 2,4 2,1 2,1

- Calcule a média, a mediana, a moda e o ponto médio.
 - Calcule o desvio-padrão, a variância e a amplitude.
 - Os dados provêm de uma população discreta ou contínua?
 - Qual é o nível de mensuração desses valores? (Nominal, ordinal, intervalar, razão)
2. a. Um conjunto de dados está no nível nominal de mensuração, e desejamos obter um valor representativo dos dados. Qual das medidas seguintes é mais adequada: média, mediana, moda ou ponto médio? Por quê?
- b. Obtém-se uma amostra telefonando para os 250 primeiros assinantes da lista telefônica local. Que tipo de amostragem

Projeto para Computador

Admite-se, de modo geral, que a temperatura média de um adulto sadio seja de 98,6°F. Com base no Conjunto de Dados 2 do Apêndice B, considere as temperaturas tomadas à meia-noite do segundo dia. Como o Conjunto de Dados 2 não está armazenado como um arquivo STATDISK ou Minitab, devemos utilizar STATDISK ou Minitab para introduzir as 106 temperaturas e salvá-las como um arquivo de nome BODYTEMP. Passamos então a obter um histograma, um diagrama em caixas, medidas de tendência central, medidas de variação, Q_1 , Q_3 , o mínimo e o máximo. Esses resultados permitam-nos descrever características importantes dos dados. Com base nessa amostra, que podemos concluir sobre a crença comum de que a temperatura média do corpo humano seja de 98,6°F? É este o resultado que esperávamos?

DOS DADOS PARA A DECISÃO

O Lixo e o Tamanho da População

Consideremos o Conjunto de Dados 1 do Apêndice B. Os dados se referem aos pesos de diferentes categorias de lixo de 62 residências, e foram coletados como parte do Garbage Project (Projeto do Lixo) na Universidade do Arizona. Há vários aspectos a considerar nesse conjunto de dados. No Capítulo 9 veremos se há alguma relação entre o tamanho da residência e a quantidade de lixo descartado, de forma que possamos prever o tamanho da população de uma região analisando o lixo descartado. Por ora, vamos trabalhar com estatística descritiva baseada nos dados.

- Construa um diagrama de Pareto e um gráfico em setores ilustrando os valores relativos dos pesos totais de resíduos de metal, papel, plástico, vidro, alimentos, jardinagem, tecidos e outros. (Em lugar de frequências, utilizamos os pesos totais.) Com base nos resultados, que categorias parecem ser as maiores componentes da quantidade total de resíduos? Há alguma categoria isolada que se destaque como a maior componente?
- Um gráfico em setores do *USA Today* mostra os resíduos de metal, papel, plástico, vidro, alimentos, jardinagem e

outros com os percentuais de 14%, 38%, 18%, 2%, 4%, 11% e 13%, respectivamente. Esses percentuais se afiguram compatíveis com o Conjunto de Dados 1 do Apêndice B?

- Determine, para cada categoria, a média e o desvio-padrão, e construa um histograma dos 62 pesos. Registre os resultados na tabela a seguir.
- As quantidades de lixo descartado são dadas por peso. Muitas regiões têm serviço de coleta de lixo residencial feito por caminhões que comprimem o lixo, e as taxas do serviço se baseiam no peso. Sob essas condições, o volume do lixo tem importância para o problema de coleta na comunidade? Há outros fatores importantes? Quais?
- Com base nos resultados precedentes, se fosse necessário desenvolver esforços de conservação ou reciclagem em virtude de a capacidade de coleta de resíduos em sua região estar quase esgotada, que providências tomaria?

	Metal	Papel	Plástico	Vidro	Alimentos	Jardinagem	Tecidos	Outros
Média								
Desvio-padrão								
Forma da distribuição								

ATIVIDADES EM GRUPO

- Atividade Extraclasse:** As Estimativas São Influenciadas por Números Âncora? No artigo "Weighing Anchors" na revista *Omni*, o autor John Rubin observou que, quando as pessoas estimam um valor, sua estimativa em geral é "âncorada" a (ou influenciada por) um número precedente, mesmo que esse número esteja totalmente desvinculado da grandeza que está sendo estimada. Para comprová-lo, pediu a diversas pessoas que estimassem rapidamente o valor de $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$. A resposta média foi 2250; mas, invertida a ordem dos números, a resposta média foi de 512. Rubin explica que, quando começamos nossos cálculos com números maiores (como $8 \times 7 \times 6$), nossas estimativas tendem a ser maiores. Observe que tanto 2250 como 512 estão muito abaixo do verdadeiro valor de 40.320. O artigo sugere que números irrelevantes podem influenciar

avaliações de propriedades imóveis, de automóveis, ou estimativas da probabilidade de uma guerra nuclear.

Realize um experimento para testar essa teoria. Escolha algumas pessoas e peça-lhes que estimem rapidamente o valor de

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

Selecione em seguida outro grupo de pessoas e peça-lhes que estimem rapidamente o valor de

$$1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

Registre as estimativas, juntamente com a ordem utilizada. Planeje cuidadosamente o experimento de modo que as condições sejam uniformes e os dois grupos amostrais sejam selecionados com o mínimo possível de tendenciosidade. Não revele a teoria aos indivíduos até que eles tenham feito suas

ATIVIDADES EM GRUPO (Continuação)

estimativas. Compare os dois conjuntos de resultados amostrais com auxílio de métodos deste capítulo. Elabore um relatório datilografado que inclua os dados coletados, os métodos usados, o método de análise, quaisquer gráficos e/ou estatísticas relevantes, e um resumo das conclusões. Inclua uma crítica das razões por que os resultados poderiam não ser corretos e indique maneiras de melhorar o experimento.

Uma variante do experimento precedente consiste em entrevistar pessoas sobre seus conhecimentos acerca da população do Quênia. Primeiro pergunte à metade das pessoas do grupo se elas acham que a população é superior ou inferior a 5 milhões; peça-lhes, em seguida, que estimem a população em um número efetivo. Pergunte à outra metade das pessoas se acham que a população é superior ou inferior a 80 milhões, e peça-lhes em seguida que estimem a população. (A população efetiva do Quênia é de 28 milhões.) Compare os dois conjuntos de resultados e identifique o efeito de "âncoragem" do número inicial mencionado aos indivíduos pesquisados.

2. *Atividade em Classe:* Em cada grupo de três ou quatro estudantes, ache o valor total das moedas em poder de cada um. Ache a média e o desvio-padrão do grupo, e permutue essas estatísticas com os outros grupos. Utilizando as médias grupais como um conjunto individual de dados, calcule a média, o desvio-padrão e a forma da distribuição. Compare esses resultados com a média e o desvio-padrão achados originalmente no grupo.

3. *Atividades em Classe:* A seguir temos as idades de motociclistas mortos em acidentes de trânsito (com base em dados do Ministério dos Transportes dos EUA). Se seu objetivo é dramatizar os perigos das motos para os jovens, que recurso seria mais eficiente: histograma, diagrama de Pareto, gráfico em setores, gráfico por pontos, média, mediana, ...? Construa o gráfico e ache a estatística que melhor atende ao objetivo. É correto distorcer deliberadamente os dados se o objetivo é salvar vidas de motociclistas?

17	38	27	14	18	34	16	42	28	24	40	20	23	31
37	21	30	25	17	28	33	25	23	19	51	18	29	

entrevista

Anthony DiUglio

Analista Nuclear, Probabilistic Risk Assessment, Consolidated Edison Company of New York, Inc.

Anthony DiUglio trabalha no Probabilistic Risk Assessment (PRA) Group (Grupo de Avaliação de Risco Probabilístico) da Unidade N° 2 (Indian Point) de geração nuclear da Consolidated Edison em Buchanan, Nova York. Em seu trabalho como Analista Nuclear, Tony estabelece probabilidades utilizadas para quantificar vários aspectos da avaliação do risco da usina. Tony é um ex-aluno do autor.

Quais são suas atribuições?

No PRA preocupamo-nos com três questões básicas sobre o risco: o que pode acontecer, qual é a chance de acontecer, e quais são as consequências caso aconteça. Essas questões sobre o risco se aplicam ao funcionamento seguro, confiável e contínuo de nossa usina. Quando quantificamos o risco, atribuímos números que são probabilidades. Se alguém sugere uma modificação no sistema de segurança da usina, analisamo-la do ponto de vista do risco. A modificação é melhor para o sistema? Afeta a operação da usina ou coloca em risco a saúde pública e a segurança?

Como o Sr. utiliza a probabilidade e/ou a estatística?

Usamos recursos fundamentais. Nossa PRA exige que quantifiquemos as taxas de reparo específicas da usina para todos os componentes ligados à segurança. Ao estabelecer taxas de reparo de componentes para bombas e válvulas, recorremos a dados da indústria em geral (genéricos) e a dados específicos de nossa usina. Combinamos essas informações, sob incerteza, e chegamos a probabilidades de reparo específicas para as diversas componentes.

Como utiliza a probabilidade e/ou a estatística em outros departamentos na Indian Point?

Nosso Performance Department calcula diversos parâmetros da usina, como taxa de aquecimento, geração em megawatts, custo de geração por kilowatt etc. Esses parâmetros são todos obtidos com o auxílio da estatística. Os recursos estatísticos que utilizamos são tendência de dados, curvas normais, desvios-padrão, histogramas etc. O Financial Planning utiliza amplamente a estatística ao projetar orçamentos e determinar suas restrições. Nossos previsores utilizam a teoria da probabilidade para prever a demanda em diferentes épocas do ano (por exemplo, inverno e verão para um, três e cinco anos para a frente). Há tanta gente utilizando a estatística em seu trabalho cotidiano que a estatística é hoje um instrumento poderoso para engenheiros, planejadores, previsores, e para nós da Avaliação do Risco.

Em termos de estatística, quais seriam suas recomendações aos candidatos a emprego?

Eles devem ter um bom conhecimento de probabilidade, estatística e suas aplicações. Como PRA é ainda uma área relativamente nova,

surtem frequentemente problemas que até então não havíamos encontrado; assim, muitos deles exigem criatividade. Uma vez de posse dos instrumentos básicos, seu tempo é empregado eficientemente. Não podemos nos comunicar com eficiência a menos que utilizemos uma linguagem comum — e esta linguagem é a estatística.

Seu trabalho conseguiu convencer a opinião pública de que sua usina é segura?

A segurança é sempre nossa primeira preocupação. No início da década de 1980, houve uma série de reuniões públicas realizadas pela Nuclear Regulatory Commission (NRC) para discutir se nossa usina devia, ou não, continuar em operação. A Consolidated Edison afirmava que sua usina era segura, justificando-se a continuação das operações da mesma através do nosso PRA. Ao término daquelas reuniões, a NRC concordou com nossa posição, e continuamos a operar.

Quem foi seu melhor professor de matemática?

O professor Mario Triola.

Sua utilização da probabilidade e da estatística tem aumentado, diminuído ou permanecido constante?

Tem aumentado continuamente. Estamos muito envolvidos com os indicadores de desempenho da usina como parâmetros da eficiência operacional da usina. Com a PRA temos agora um instrumento que nos permite focalizar a atenção sobre as componentes e funções mais importantes da usina. No caso de três componentes necessitarem todas de manutenção, a PRA permite-nos identificar qual componente deve voltar ao serviço em primeiro lugar. Em engenharia, se temos diversas componentes que devem ser melhoradas, a PRA permite identificar qual delas deve ser melhorada primeiro. Podemos quantificar os efeitos e, assim, dirigir melhor nossos recursos, tornando a usina mais segura.

Apêndice E: Respostas dos Exercícios de Número Ímpar (e de TODOS os Exercícios de Revisão e Exercícios Cumulativos de Revisão)

Seção 1-2

1. Contínua 3. Discreta 5. Discreta 7. Contínua
9. Ordinal 11. Nominal 13. Intervalar 15. Nominal
17. Ordinal
19. Intervalar. As diferenças entre os anos podem ser determinadas e são significativas, mas não há ponto de partida inerente, pois o tempo não começou no ano zero.

Seção 1-3

1. Excluem-se as pessoas com números não-listados e pessoas sem telefone.
3. Um estudo patrocinado pela indústria cítrica muito provavelmente chegará a conclusões favoráveis a essa indústria.
5. Como os pesquisados são auto-selecionados, os resultados da pesquisa não são válidos.
7. 62% de 8% de 1875 representam apenas 93.
9. Mães que comem lagostas tendem a ser mais ricas e portanto podem pagar por melhor atendimento médico.
11. Um fabricante de graxa para sapatos obviamente tem interesse na importância do seu produto, e há muitas maneiras de este fato afetar os resultados da pesquisa.
13. J. Douglas Carroll escreveu, em uma carta ao editor do *New York Times*, que a média de 69,5 para todos os homens é medida a contar do nascimento, enquanto os homens só se tornam regentes por volta dos 30 anos. Levando-se isto em conta, a média de 73,4 anos não é significativa.
15. O fraseado da questão é tendencioso; tende a encorajar respostas negativas. O tamanho da amostra (20) é muito pequeno. Os pesquisados são auto-selecionados, em lugar de serem selecionados pelo jornal. Se 20 leitores respondem, as percentagens devem ser múltiplos de 5; 87% e 13% não são resultados possíveis.

Seção 1-4

1. Estudo observacional 3. Experimento 5. Conveniência
7. Estratificada 9. Aleatória 11. Estratificada
13. Conglomerado 15. Aleatória
17. a. Uma vantagem das questões abertas é que proporcionam ao entrevistado e ao entrevistador muito maior diversidade de respostas; e uma desvantagem é que as questões abertas podem ser muito difíceis de analisar.
- b. Uma vantagem das questões fechadas é que reduzem a chance de uma interpretação errônea do tópico; uma desvantagem é que as questões fechadas impedem a inclusão de respostas válidas que o entrevistador pode não ter considerado.
- c. As questões fechadas são mais fáceis de analisar com processos estatísticos formais.

Seção 1-5

1. 2.636 3. 3.6055513 5. 1067,1111 7. 5005
9. STATDISK: 0,838 0,875 0,870
- Minitab: 3,22 1 1

Capítulo 1 Exercícios de Revisão

1. a. Contínua b. Razão c. Estratificada d. Estudo observacional
- e. Os produtos que utilizam as baterias podem sofrer danos.
2. a. Razão b. Ordinal c. Ordinal d. Intervalar e. Nominal
3. Como se trata de uma pesquisa pelo correio, os pesquisados são auto-selecionados e provavelmente consistirão naqueles que têm opiniões formadas sobre o assunto. Os pesquisados auto-selecionados não representam necessariamente o ponto de vista de todos os investidores.
4. a. Discreta b. Contínua c. Contínua
5. a. Sistemática b. Aleatória c. Conglomerado d. Estratificada
- e. Conveniência

6. Os pesquisados tendem a arredondar para um número par "simpático", como 50.
7. A amostra poderia ser tendenciosa, ao excluir os que trabalham, os que não fazem refeições na escola, os que viajam etc.
8. A cifra pode parecer muito precisa, mas provavelmente não é muito exata. Um número com tal grau de precisão pode sugerir, talvez incorretamente, que seja também exato.

Capítulo 1 Exercícios Cumulativos de Revisão

1. A segunda versão da questão é substancialmente menos confusa porque não inclui uma dupla negativa. Uma possibilidade para uma pergunta melhor:
"Com qual das duas afirmações o leitor concorda?"
• A exterminação de judeus pelos nazistas nunca ocorreu.
• A exterminação de judeus pelos nazistas ocorreu efetivamente."
2. A resposta varia

Seção 2-2

1. Amplitude de classe 6. Pontos médios: 2,5; 8,5; 14,5; 20,5; 26,5. Fronteiras de classe: -0,5; 5,5; 11,5; 17,5; 23,5; 29,5.
3. Amplitude de classe: 2,0. Pontos médios: 0,95; 2,95; 4,95; 6,95; 8,95. Fronteiras de classe: -0,05; 1,95; 3,95; 5,95; 7,95; 9,95.

5. Ausências	Frequência Relativa	7. Peso (kg)	Frequência Relativa
Menos de 6	0,195	Menos de 2,0	0,133
Menos de 12	0,205	Menos de 4,0	0,213
Menos de 18	0,190	Menos de 6,0	0,327
Menos de 24	0,200	Menos de 8,0	0,207
Menos de 30	0,210	Menos de 10,0	0,120

9. Ausências	Frequência Acumulada	11. Peso (kg)	Frequência Acumulada
Menos de 6	39	Menos de 2,0	20
Menos de 12	80	Menos de 4,0	52
Menos de 18	118	Menos de 6,0	101
Menos de 24	158	Menos de 8,0	132
Menos de 30	200	Menos de 10,0	150

13. No Exercício 1, os números de ausências se distribuem de maneira aproximadamente equitativa pelas cinco classes, mas as ausências no Exercício 2 apresentam frequências relativamente baixas no início, crescendo para um máximo na classe média, e decrescendo novamente em direção à última classe.
15. 0,26-0,75

17. Peso (kg)	Frequência	19. Tempo (min)	Frequência
0-49	6	56-63	8
50-99	10	64-71	3
100-149	10	72-79	9
150-199	7	80-87	17
200-249	8	88-95	8
250-299	2	96-103	4
300-349	4	104-111	1
350-399	3		
400-449	3		
450-499	0		
500-549	1		

21. Frequências relativas para os homens: 0,019; 0,071; 0,118; 0,171; 0,087; 0,273; 0,142; 0,118. Frequências relativas para as mulheres: 0,010; 0,072; 0,173; 0,265; 0,042; 0,279; 0,060; 0,100. As distribuições são muito parecidas, exceto quanto ao fato de haver desproporcionalmente mais mulheres na classe 3,0-3,9 e menos na classe 10,0-14,9.