capítulo 1

Estatística descritiva

Objetivos do estudo

Ao final deste capítulo, você deverá ser capaz de:

- Reconhecer tipos diferentes de dados e usar métodos apropriados para sintetizá-los e analisá-los.
- Usar técnicas gráficas para fazer uma síntese visual de séries de dados.
- Usar técnicas numéricas para sintetizar séries de dados.
- Reconhecer os pontos fortes e as limitações de tais métodos.
- Perceber a utilidade da variação como fonte de informação e conhecimento adicional sobre conjuntos de dados.

O objetivo dos métodos de estatística descritiva é simples: apresentar informações de maneira clara, concisa e precisa. A dificuldade na análise de muitos fenômenos, sejam eles econômicos, sociais ou de outra espécie, está no fato de que simplesmente há informação demais para a nossa mente assimilar. A tarefa dos métodos descritivos é, portanto, sintetizar toda essa informação e salientar os aspectos principais, sem que a figura fique distorcida.

Consideremos, por exemplo, o problema de apresentar informações sobre a riqueza dos cidadãos britânicos (o que será discutido mais adiante neste capítulo). Existem aproximadamente 17 milhões de unidades familiares para as quais há dados disponíveis, e apresentar os dados na forma bruta (isto é, o patrimônio de cada uma delas) não seria útil ou informativo (daria um livro de 30 mil páginas!). É mais útil ter menos informação, desde que ela seja representativa dos dados originais. Ao fazer isso, boa parte da informação original é deliberadamente descartada; na verdade, é possível definir estatística descritiva como a arte de descartar de maneira construtiva a maior parte dos dados!

Há muitos modos de sintetizar dados e poucas regras claras e simples a respeito de como fazê-lo. Jornais e revistas geralmente oferecem maneiras inovadoras (mas nem sempre bem-sucedidas) de apresentar dados. Existem, entretanto, algumas técnicas comprovadas, que serão abordadas neste capítulo. Elas são eficazes (a) porque nos dizem algo sobre os dados subjacentes e (b) por serem razoavelmente familiares a um grande

número de pessoas, o que nos permite falar numa linguagem comum. Por exemplo, a média revela alguma coisa sobre a posição dos dados e é um conceito conhecido da maioria das pessoas. Muitos pais dizem, por exemplo, que o dia de seu filho na escola foi "médio".

A definição do método apropriado de análise dos dados depende de diversos fatores: o tipo de informação a ser considerada, o nível cultural do público e a "mensagem" que se pretende transmitir. Métodos diferentes seriam usados para persuadir acadêmicos da validade de uma teoria sobre a inflação ou para convencer os consumidores de que o sabão em pó X lava mais branco do que o Y.

Para demonstrar o uso dos diversos métodos, são abordados três tópicos neste capítulo. Inicialmente, examinamos a relação entre desempenho educacional e perspectivas de emprego. Uma formação de nível superior aumenta as chances de se conseguir emprego? Os dados foram obtidos com indivíduos consultados em 2003. Isso significa que temos uma amostra de dados em ***cross section*** que dá uma visão da situação em certo momento. Verificamos a distribuição de índices de desempenho educacional entre as pessoas pesquisadas, bem como a sua relação com os resultados relativos a emprego.

Em segundo lugar, examinamos a distribuição de riqueza no Reino Unido em 2001. Os dados, mais uma vez, são do tipo *cross section*, mas dessa vez podemos utilizar métodos mais sofisticados, pois a riqueza é medida numa **escala razão**. Uma pessoa com patrimônio de £200.000 é duas vezes mais rica do que alguém com patrimônio de £100.000, por exemplo, e essa relação tem um significado. No caso da formação, não se pode dizer com precisão que uma pessoa é duas vezes mais bem formada do que outra (daí resulta o eterno debate sobre padrões de ensino). Os níveis de ensino podem ser ordenados (para que se diga se uma pessoa tem formação melhor do que outra), mas não é possível medir a "distância" entre eles. Dizemos que a escolaridade se mede numa escala **ordinal**. Em contraposição, não existe um ordenamento natural para as três categorias de emprego (empregado, desempregado, inativo), o que significa que essa variável é medida numa escala **nominal**.

Em terceiro lugar, examinamos o investimento no período de 1970 a 2002. Para tanto utilizamos dados em **série temporal**, pois dispomos de um número de observações da variável medida em diferentes momentos. Nesse caso, é importante levar em conta a dimensão temporal dos dados: as coisas pareceriam diferentes se as observações fossem feitas na ordem 1970, 1983, 1977... e não na ordem temporal correta. Também verificamos a relação entre duas variáveis – investimento e produto – nesse mesmo período e identificamos métodos apropriados para apresentar essa relação.

Nos três casos, usamos métodos tanto gráficos quanto numéricos para sintetizar os dados. Embora haja algumas diferenças entre os métodos empregados nos três casos, não existem compartimentos estanques: os métodos utilizados num caso poderiam ser adequados em outro, talvez com ligeiras modificações. Parte da competência do estatístico consiste em determinar quais são os métodos de análise e apresentação mais apropriados a cada problema específico.

Síntese de dados com o uso de técnicas gráficas

Formação e emprego – ou, depois de tudo isso, você conseguirá um emprego?

Começamos com a discussão de uma questão que deve ser importante para você: de que modo a educação influi nas suas chances de obter emprego? Já que por todo o mundo o desemprego atinge níveis elevados, tanto nos países em desenvolvimento quanto nos desenvolvidos, um dos possíveis benefícios do investimento em educação é a diminuição das chances de ficar sem trabalho. Mas em quanto ele reduz essa possibilidade? Utilizaremos várias técnicas gráficas para investigar esse tema.

Os dados brutos dessa investigação vêm de *Education and training statistics for the U.K. 2003*. Alguns desses dados são apresentados na tabela 1.1 e mostram o número de pessoas por categoria de emprego (em atividade, desempregadas ou inativas, ou seja, que não estão em busca de trabalho) e por nível de escolaridade (ensino superior, *A-level**, outro grau de escolaridade ou sem escolaridade). A tabela apresenta uma **tabulação cruzada** do *status* quanto a emprego por escolaridade, que é simplesmente uma contagem (**freqüência**) do número de pessoas situadas em cada uma das 12 células da tabela. Por exemplo, 8.224.000 pessoas com ensino superior estavam trabalhando, parte de pouco mais de 37 milhões de pessoas na faixa de idade de trabalho.

Gráfico de barras

A primeira técnica gráfica que utilizaremos é o **gráfico de barras**, mostrado na figura 1.1. Ele sintetiza a formação de pessoas em atividade, isto é, os dados da primeira linha da tabela. Os quatro graus de escolaridade são dispostos no eixo horizontal (x), enquanto as freqüências são medidas no eixo vertical (y). A altura de cada barra representa o número de pessoas em atividade na respectiva categoria.

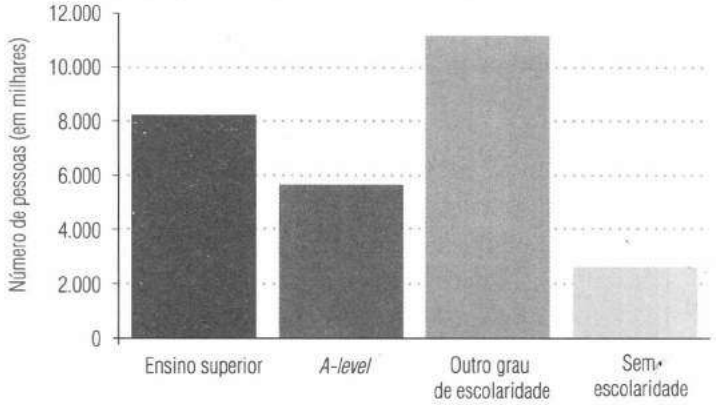
TABELA 1.1 Situação de emprego e escolaridade, 2003 (números em milhares)

	Ensino superior	*A-level*	Outro grau de escolaridade	Sem escolaridade	Total
Em atividade	8.224	5.654	11.167	2.583	27.628
Desempregados	217	231	693	303	1.444
Inativos	956	1.354	3.107	2.549	7.966
Total	9.397	7.239	14.967	5.435	37.038

* *A-level*, ou *advanced level*, é o certificado obtido por alunos da Inglaterra, País de Gales e Irlanda do Norte aprovados nos exames relativos aos dois últimos anos do ensino secundário optativo, aos 17 e 18 anos de idade. É exigido por muitas universidades e no mercado de trabalho. (N. do e.)

FIGURA 1.1

Grau de escolaridade das pessoas em atividade no Reino Unido, 2003



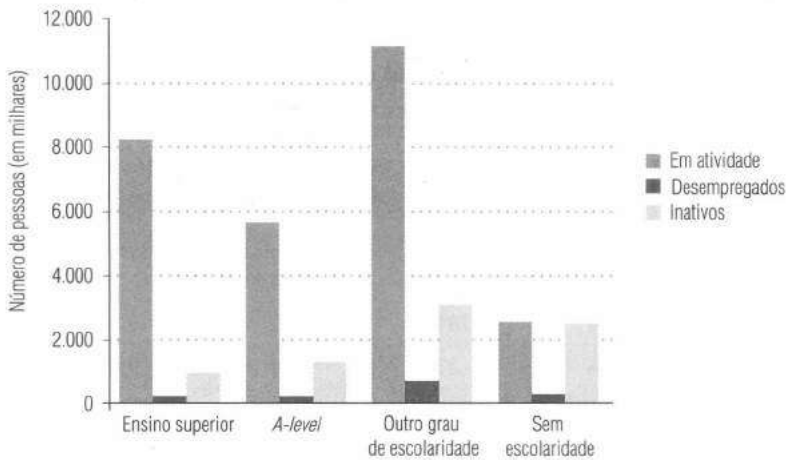
Nota: A altura de cada barra é determinada pela freqüência correspondente. A altura da primeira barra é de 8.224 unidades; a da segunda, de 5.654 unidades, e assim por diante. A ordem das barras poderia ser alterada (colocando-se a categoria "sem escolaridade" em primeiro lugar) sem que a mensagem fosse modificada.

Vê-se que o maior grupo é formado por pessoas com "outro grau de escolaridade", que é quase tão grande quanto os de "ensino superior" e "*A-level*" juntos. A categoria "sem escolaridade" é a menor de todas, embora represente uma proporção substancial do número de pessoas em atividade.

Seria interessante comparar essa distribuição com a de pessoas desempregadas e inativas. Isso é feito na figura 1.2, que acrescenta as barras relativas a essas duas outras categorias.

FIGURA 1.2

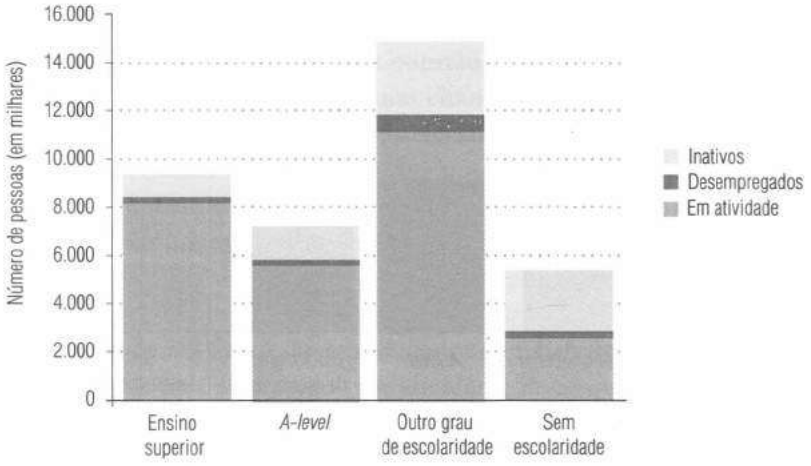
Grau de escolaridade por situação de emprego



Nota: As barras das categorias "desempregados" e "inativos" são construídas da mesma forma que as da categoria "em atividade"; a altura de cada barra é determinada pela freqüência.

FIGURA 1.3

Gráfico de barras empilhadas de escolaridade e situação de emprego



Nota: A altura total de cada barra é determinada pela soma das freqüências da categoria, fornecida na última linha da tabela 1.1.

Esse **gráfico de barras múltiplas** mostra que quanto mais baixo for o nível de formação, maior será o tamanho das categorias "desempregados" e "inativos". A categoria "sem escolaridade" é numericamente menos importante do que as demais, o que dificulta as comparações diretas, mas as categorias "desempregados" e "inativos" são grandes em relação ao número de pessoas "em atividade".

A figura 1.3 mostra um método alternativo de apresentação: o **gráfico de barras empilhadas**. Nesse caso, as barras são empilhadas, em vez de serem dispostas lado a lado.

Obtém-se uma visão mais clara no caso de os dados serem **transformados** em porcentagens por coluna, isto é, cada coluna sendo expressa como porcentagem do total. Isso facilita a realização de comparações diretas entre os vários níveis de ensino. Podemos assim ver, dentre as pessoas com ensino superior, qual proporção está "em atividade" (88%), e assim por diante. Esses números são apresentados na tabela 1.2.

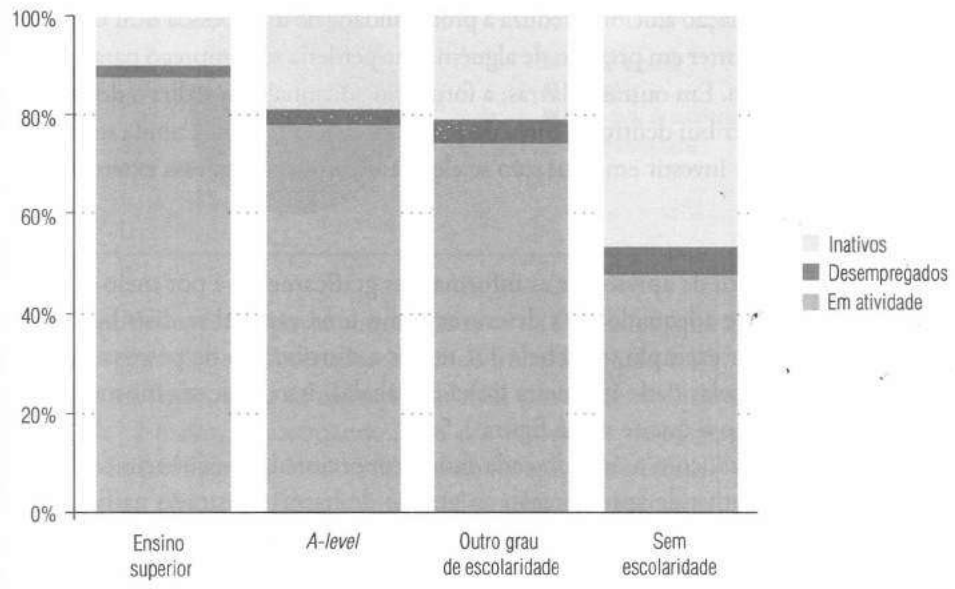
TABELA 1.2 Situação de emprego e escolaridade: porcentagens das colunas

	Ensino superior	*A-level*	Outro grau de escolaridade	Sem escolaridade	Total
Em atividade	88%	78%	74%	48%	75%
Desempregados	2%	3%	5%	6%	4%
Inativos	10%	19%	21%	47%	21%

Nota: As porcentagens das colunas são obtidas dividindo-se cada freqüência pelo total da coluna. Por exemplo, 88% é igual a 8.224 dividido por 9.397; 78% é igual a 5.654 dividido por 7.239, etc. As porcentagens podem não somar 100% em razão de arredondamento.

FIGURA 1.4

Porcentagens em cada situação de emprego, por grau de escolaridade



Depois disso, fica mais fácil fazer uma comparação direta entre os diferentes níveis de ensino (colunas). Isso aparece na figura 1.4, na qual todas as barras têm a mesma altura (correspondendo a 100%) e os componentes de cada uma delas indicam as *proporções* de pessoas em atividade, desempregadas ou inativas, em cada grau de escolaridade.

Fica claro agora como o *status* econômico varia de acordo com o nível de escolaridade, num resultado bastante dramático. Em particular:

- A probabilidade de desemprego aumenta drasticamente com a queda do grau de escolaridade (nesse caso, interpretam-se as proporções como probabilidades, ou seja, se 10% estão desempregados, então a probabilidade de que um indivíduo escolhido ao acaso não esteja trabalhando é igual a 10%).
- A maior diferença ocorre entre a categoria "sem escolaridade" e as outras três, em que as diferenças são relativamente pequenas.

Portanto, podemos concluir com segurança que a probabilidade de estar desempregado é reduzida significativamente pela formação? Poderíamos ir adiante e argumentar que o caminho para um desemprego menor é, sem dúvida, o investimento em educação? A resposta *pode* ser "sim" às duas perguntas, mas não provamos essa relação. Observe estas duas importantes considerações:

- Ignoramos a capacidade inata. As pessoas mais habilidosas tenderiam a estar empregadas e a investir mais em formação. Idealmente, gostaríamos de fazer a comparação entre indivíduos com capacidade semelhante, mas graus de escolaridade diferentes; entretanto, é difícil obter tais dados.
- Mesmo que a formação adicional reduza a probabilidade de uma pessoa ficar desempregada, isso pode ocorrer em prejuízo de alguém, que perderia seu emprego para o indivíduo mais instruído. Em outras palavras, a formação adicional não reduz o desemprego total, apenas o distribui dentro da força de trabalho. Evidentemente, ainda seria lógico para os indivíduos investir em educação se eles não considerassem essa externalidade.

Gráfico de pizza

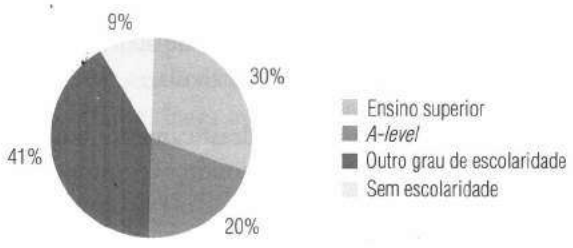
Outra maneira útil de apresentar as informações graficamente é por meio do **gráfico de pizza**, bastante adequado para descrever como uma variável se distribui por diversas categorias. Por exemplo, na tabela 1.1 temos a distribuição de pessoas em atividade por grau de escolaridade (primeira linha da tabela). Isso pode ser mostrado num gráfico de pizza como o que se vê na figura 1.5.

O gráfico de pizza, com a área de cada fatia proporcional à freqüência correspondente, é uma alternativa de apresentação ao gráfico de barras mostrado na figura 1.1. As porcentagens de cada nível de ensino foram colocadas em volta do gráfico, mas isso não é essencial. Para fins de apresentação, é melhor não usar muitas fatias: com mais de seis pedaços o gráfico tende a parecer excessivamente quebrado.

O gráfico revela que aproximadamente 40% das pessoas que estão trabalhando situam-se na categoria "outro grau de escolaridade" e que apenas 9% não têm escolari-

FIGURA 1.5

Grau de escolaridade das pessoas em atividade



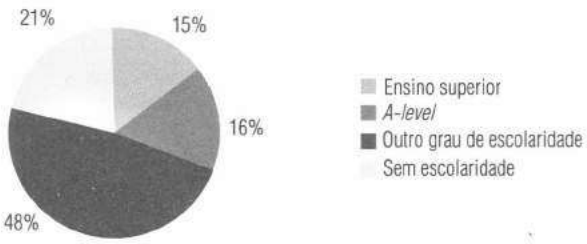
Nota: Se você precisar desenhar um gráfico de pizza à mão, o ângulo de cada fatia pode ser calculado da seguinte maneira:

$$\text{ângulo} = \frac{\text{freqüência}}{\text{freqüência total}} \times 360$$

O ângulo da primeira fatia, por exemplo, é igual a $\frac{8.224}{27.628} \times 360 = 107,2^{\circ}$

FIGURA 1.6

Grau de escolaridade das pessoas desempregadas



dade. Isso pode ser contraposto com o que se vê na figura 1.6, que apresenta um gráfico semelhante para os desempregados (segunda linha da tabela 1.1).

A categoria "outro grau de escolaridade" tem tamanho semelhante, mas o grupo "sem escolaridade" é maior, representando 21% dos indivíduos desempregados. Além disso, a proporção de quem tem diploma universitário foi reduzida à metade, de 30% para 15%.

Com o uso de tais gráficos, somos ainda capazes de apresentar os principais aspectos revelados pelos dados de um jeito interessante. Feito corretamente, esse modo de transmitir uma mensagem é extremamente eficaz.

Confecção de gráficos no Microsoft Excel

A maior parte dos gráficos deste livro foi produzida com os recursos do Excel. Sem pretender ditar um estilo específico, você deve tentar obter uma representação semelhante, que não pareça confusa. Veja algumas sugestões úteis:

- Desenhe as linhas de grade em um tom claro de cinza (como elas não fazem realmente parte do gráfico, é melhor que fiquem discretas).
- Livre-se do preenchimento do fundo (como o padrão é cinza, altere a escolha para "sem preenchimento"), pois ele não vai sair muito definido se você tiver de imprimir o gráfico.
- No eixo horizontal, utilize rótulos horizontais ou verticais, mas não-inclinados – isso dificulta perceber a qual ponto eles se referem. Se forem inclinados, clique duas vezes sobre o eixo horizontal e depois escolha o botão de alinhamento.
- Gráficos coloridos ficam muito bem na tela do computador, mas pouco claros se tiverem de ser impressos em preto-e-branco. Altere o tipo de estilo das linhas ou dos marcadores (por exemplo, faça com que alguns sejam tracejados) para poder distingui-los na cópia em preto-e-branco.
- Pelo padrão, os dois eixos partem do zero. Se todas as suas observações forem números grandes, entretanto, os pontos podem ficar concentrados num dos cantos do gráfico. Altere a escala dos eixos para resolver isso – escolha como valor mínimo do eixo um número ligeiramente inferior ao da observação de menor valor.

Com exceção dessas observações, as opções de padrão do Excel geralmente produzem bons resultados.

Exercício 1.1

A tabela fornecida a seguir apresenta o número total de turistas (em milhões) e o de turistas ingleses que visitam cada país:

	França	Alemanha	Itália	Espanha
Todos os turistas	12,4	3,2	7,5	9,8
Turistas ingleses	2,7	0,2	1,0	3,6

a) Desenhe um gráfico de barras mostrando o número total de turistas que visitam cada país.
b) Desenhe um gráfico de barras empilhadas mostrando o número de turistas ingleses e não-ingleses proporcionalmente ao total de visitantes de cada país.
c) Desenhe um gráfico de pizza mostrando a distribuição do total de turistas entre os quatro países. Faça o mesmo para os turistas ingleses e compare os resultados.

Exame de dados em *cross section*: distribuição de riqueza no Reino Unido em 2001

Tabelas de freqüência e histogramas

Vamos agora examinar os dados de forma diferente. Os dados de escolaridade e emprego não passavam de simples freqüências, e uma característica (por exemplo, ensino superior) estava presente ou não num indivíduo em particular. Examinemos agora a distribuição de riqueza, uma variável que pode ser medida numa **escala razão**, de tal modo que um valor diferente esteja associado a cada indivíduo. Por exemplo, uma pessoa pode ter um patrimônio de £1.000, e uma outra ter £1.000.000. Serão usadas técnicas distintas de apresentação para analisar dados desse tipo. Utilizamos essas técnicas para discutir questões tais como quanta riqueza uma pessoa média tem e se a riqueza é ou não uniformemente distribuída.

Os dados são apresentados na tabela 1.3, que indica a distribuição de riqueza no Reino Unido em 2001 (os dados mais recentes na época em que este livro foi escrito), extraídos de *Inland revenue statistics 2003*. Esse é um exemplo de **tabela de freqüência**. É difícil definir e medir riqueza; os dados aqui apresentados referem-se a bens *negociáveis* (ou seja, itens como aposentadoria, que não pode ser comprada, são excluídos) e correspondem a estimativas para a população como um todo, com base em dados da arrecadação de impostos.

Dividida em 14 **intervalos de classe** – de £0 a £10.000 (exclusive); de £10.000 a £24.999, e assim por diante –, a variável riqueza apresenta o número de indivíduos (isto é, a **freqüência**) em cada intervalo de classe. Note que a **amplitude da classe** varia com a escala da riqueza: a amplitude da primeira classe é igual a £10.000, a da segunda

TABELA 1.3 Distribuição de riqueza, Reino Unido, 2001

Intervalos de classe	Números (em milhares)
0-9.999	3.417
10.000-24.999	1.303
25.000-39.999	1.240
40.000-49.999	714
50.000-59.999	642
60.000-79.999	1.361
80.000-99.999	1.270
100.000-149.999	2.708
150.000-199.999	1.633
200.000-299.999	1.242
300.000-499.999	870
500.000-999.999	367
1.000.000-1.999.999	125
2.000.000 ou mais	41
Total	16.933

é igual a £15.000, a da terceira também £15.000, e assim por diante. Esse será um fator importante quando abordarmos a apresentação gráfica dos dados.

Essa tabela foi construída com base nas 16.933.000 observações existentes sobre a riqueza de indivíduos, já constituindo, assim, uma síntese dos dados originais (observe que todas as freqüências na tabela foram expressas em milhares) com grande parte da informação original. A primeira decisão a tomar quando se monta uma tabela de freqüência com base nos dados brutos é determinar quantos intervalos de classe serão usados, bem como que amplitude eles deverão ter. As coisas serão mais simples se a amplitude for a mesma para todos os intervalos, mas, nesse caso, isso não é factível: se escolhêssemos a **amplitude-padrão** de 10.000, haveria muitos intervalos entre 500.000 e 1.000.000 (50 deles, para ser preciso), a maioria dos quais com freqüência baixa ou nula. Se a amplitude-padrão fosse igual a 100.000, haveria um número pequeno de intervalos, e o primeiro (0 a 100.000) teria 9.947 observações (59% do total), de maneira que quase todos os detalhes interessantes seriam perdidos. É preciso encontrar um meio-termo entre esses extremos.

Uma regra útil consiste em escolher um número de intervalos igual à raiz quadrada da freqüência total, desde que o máximo de intervalos seja igual a 12. Assim, por exemplo, 25 observações no total seriam alocadas a cinco intervalos; 100 observações deveriam ser agrupadas em 10 intervalos; e 16.933, em 12 (aqui são usados 14). Os intervalos

de classes precisam ser iguais, na medida do possível, mas devem aumentar quando as freqüências se tornarem pequenas.

Para que esses dados sejam apresentados graficamente, pode-se desenhar um gráfico de barras como no caso da escolaridade, conforme ilustra a figura 1.7. Antes de continuar a leitura deste texto, passe algum tempo examinando essa figura e perguntando-se o que há de errado com ela.

A resposta é: a figura apresenta uma visão inteiramente enganosa dos dados! (Por sinal, essa é a visão que você obterá usando um programa de planilha eletrônica, como foi feito aqui. Todos os aplicativos comuns parecem fazer isso; portanto, tome cuidado. É o caso de se perguntar quantas decisões já foram influenciadas pela apresentação incorreta dos dados.)

Por que a figura está errada? Considere o seguinte raciocínio. O gráfico parece mostrar que existe uma concentração de indivíduos com riqueza acima de £60.000 (a freqüência salta de 642 para 1.361) e de £100.000 (um salto de 1.270 para 2.708). Mas isso resulta apenas da mudança da amplitude da classe nesses pontos (para 20.000 em £60.000 e para 50.000 em £100.000). Imagine dividir a classe de £100.000 a £150.000 em duas: de £100.000 a £125.000 e de £125.000 a £150.000. Nesse caso, repartimos a freqüência de 2.708 igualmente entre as duas classes (essa é uma decisão arbitrária, usada apenas para ilustrar o problema). O gráfico agora passa a ser o da figura 1.8.

Comparando-se as figuras 1.7 e 1.8, nota-se uma diferença: o pico em £100.000 desaparece. Isso é perturbador, pois significa que é possível alterar a forma da distribui-

FIGURA 1.7

Gráfico de barras da distribuição de riqueza no Reino Unido, 2001

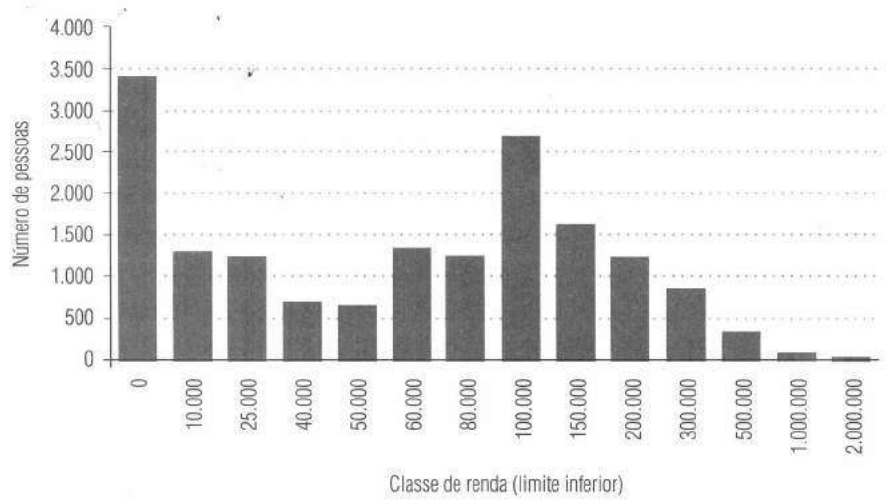


FIGURA 1.8

Distribuição de riqueza com diferentes intervalos de classe

ção simplesmente mudando a amplitude das classes. Nesse caso, como podemos confiar num exame visual da distribuição? Um método melhor consiste em fazer com que a forma da distribuição não dependa de como são escolhidos os intervalos das classes. Isso pode ser feito traçando-se um **histograma**.

Histograma

Um histograma assemelha-se a um gráfico de barras, exceto pelo fato de que corrige as diferenças de amplitude das classes. Se todas as classes tiverem a mesma amplitude, não haverá diferença entre um gráfico de barras e um histograma. Os cálculos necessários para gerar um histograma são fornecidos na tabela 1.4.

A nova coluna da tabela apresenta a **densidade de freqüência**, assim definida:

$$(1.1)\quad \text{densidade de freqüência} = \frac{\text{freqüência}}{\text{amplitude da classe}}$$

O emprego dessa fórmula corrige as figuras pelo uso de amplitudes de classes diferentes. O princípio por trás dessa correção é que, se a amplitude da classe dobrar, para compensar, a freqüência deverá ser reduzida à metade. Se a amplitude quadruplicar, então a dividiremos por quatro, e assim por diante. A maneira mais simples de fazer essa correção é dividir cada freqüência pela amplitude da classe. Portanto, 0,3417 = 3.417/10.000 é a primeira densidade de freqüência; 0,0869 = 1.303/15.000 é a segunda, etc. Acima de £200.000, as amplitudes das classes são muito grandes, e as freqüências, pequenas (pequenas demais para aparecer no histograma). Por isso, essas classes foram combinadas.

TABELA 1.4 Cálculo da densidade de freqüência

Faixa	Número	Amplitude da classe	Densidade de freqüência
0-	3.417	10.000	0,3417
10.000-	1.303	15.000	0,0869
25.000-	1.240	15.000	0,0827
40.000-	714	10.000	0,0714
50.000-	642	10.000	0,0642
60.000-	1.361	20.000	0,0681
80.000-	1.270	20.000	0,0635
100.000-	2.708	50.000	0,0542
150.000-	1.633	50.000	0,0327
200.000-	2.645	3.800.000	0,0007

Nota: Alternativamente à densidade de freqüência, pode-se calcular a freqüência por amplitude da classe-"padrão" escolhendo-se a amplitude-padrão de 10.000 (a classe mais estreita). Os valores na coluna 4 seriam, então, 3.417; 868,7 (= 1.303 ÷ 1,5); 826,7; etc. Isso produziria um histograma com a mesma forma obtida com o uso da densidade de freqüência.

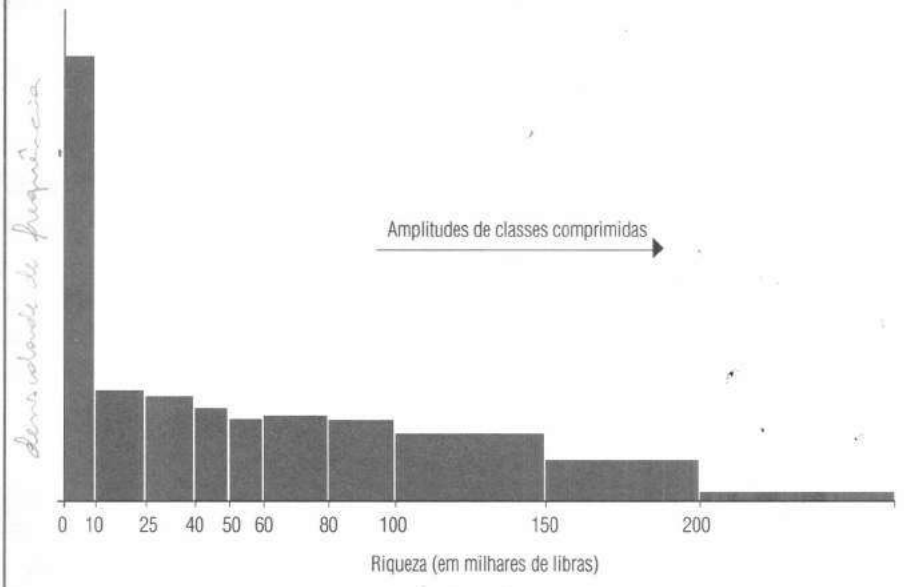
Como a amplitude do intervalo final é desconhecida, deve-se fazer uma estimativa para calcular a densidade de freqüência. Ela tende a ser bastante grande, pois uma pessoa rica pode ter ativos avaliados em vários milhões (ou até mesmo bilhões) de libras; o valor que adotarmos influenciará o cálculo da densidade de freqüência e, portanto, a forma do histograma. Felizmente, por se encontrar na cauda da distribuição, a amplitude atinge um número pequeno de observações. Nesse caso, estamos supondo (arbitrariamente) que uma amplitude de £3,8 milhões é "razoável", gerando um limite superior de £4 milhões para essa classe.

A seguir, colocam-se a densidade de freqüência e o nível de riqueza nos eixos vertical e horizontal, respectivamente, para produzir o histograma. Mas um aspecto precisa ser salientado: a escala no eixo da riqueza deve ser tão linear quanto possível, ou seja, £50.000 devem estar a uma distância da origem duas vezes maior do que £25.000. Entretanto, é difícil encaixar todos os valores no eixo horizontal sem comprimir o gráfico excessivamente nos níveis baixos de riqueza, nos quais se situa a maioria das observações. Portanto, as classes acima de £100.000 foram comprimidas, e chamamos a atenção do leitor para isso. O resultado é apresentado na figura 1.9.

O efeito do cálculo de densidade de freqüência é fazer com que a *área*, e não a altura, de cada bloco do histograma represente a freqüência. É a altura que agora mostra a densidade. Em conseqüência, esse procedimento fornece uma imagem precisa da forma da distribuição.

FIGURA 1.9

Histograma da distribuição de riqueza no Reino Unido, 2001



Nota: Teríamos um **polígono de freqüência** se, em vez de desenhar blocos no histograma, contruíssemos linhas ligando o centro do topo de cada bloco.

Tendo feito tudo isso, o que mostra o histograma? Vejamos os pontos que mais se destacam:

- O histograma tem uma forte **assimetria** à direita (ou seja, a cauda mais longa ocorre à direita). A maioria das pessoas tem níveis modestos de riqueza; poucas pessoas são muito ricas.
- O intervalo **modal** é de £0 a £10.000 (isto é, tem a maior densidade; nenhum outro intervalo de £10.000 é mais numeroso).
- A maioria das pessoas (na verdade, 51,2%) tem menos de £80.000 de riqueza em termos de valor de mercado.
- Aproximadamente 16% das pessoas têm riqueza superior a £200.000.[1]

A figura indica a existência de um grau elevado de desigualdade na distribuição de riqueza. Se isso é aceitável, ou até mesmo desejável, é uma questão de juízo de valor. Observe que parte dessa desigualdade deve-se à diferença de idade: os indivíduos mais jovens ainda não tiveram tempo suficiente para acumular muita riqueza e, portanto, parecem estar em posição de inferioridade; mas, se considerarmos sua estimativa de vida, isso pode mudar. Para ter uma visão melhor da distribuição de riqueza, talvez seja

1. Devido à compressão das amplitudes de algumas classes de riqueza, é difícil perceber esse fato com clareza no histograma. A apresentação gráfica tem limitações.

necessário analisar o processo de aquisição dessa riqueza ao longo da vida. Na verdade, uma correção por diferença de idade não produz grande efeito sobre o padrão de distribuição de riqueza. A respeito desse aspecto e da desigualdade da distribuição de riqueza em geral, veja Atkinson (1983), capítulos 7 e 8.

Distribuição de freqüência relativa e de freqüência relativa acumulada

A distribuição de riqueza também pode ser ilustrada com o uso da distribuição de **freqüências relativa** e **acumulada** dos dados, cujos valores são calculados na tabela 1.5.

As freqüências relativas indicam a *proporção* das observações que se situam dentro de cada intervalo de classe. Por exemplo, 4,2% dos indivíduos têm nível de riqueza entre £40.000 e £50.000. As freqüências relativas são apresentadas na terceira coluna, provenientes da seguinte fórmula:[2]

TABELA 1.5 Cálculo de freqüências relativa e acumulada

Faixa	Freqüência	Freqüência relativa (%)	Freqüência acumulada
0-	3.417	20,2	3.417
10.000-	1.303	7,7	4.720
25.000-	1.240	7,3	5.960
40.000-	714	4,2	6.674
50.000-	642	3,8	7.316
60.000-	1.361	8,0	8.677
80.000-	1.270	7,5	9.947
100.000-	2.708	16,0	12.655
150.000-	1.633	9,6	14.288
200.000-	1.242	7,3	15.530
300.000-	870	5,1	16.400
500.000-	367	2,2	16.767
1.000.000-	125	0,7	16.892
2.000.000-	41	0,2	16.933
Total	16.933	100,0	

Nota: As freqüências relativas são calculadas exatamente como as porcentagens de colunas na tabela 1.2. Assim, por exemplo, obtém-se 20,2% dividindo 3.417 por 16.933. As freqüências acumuladas são obtidas juntando-se freqüências, ou seja, somando-se as freqüências sucessivamente. Por exemplo, 4.720 é igual a 3.417 + 1.303; 5.960 é igual a 4.720 + 1.240, etc.

2. Se você não estiver familiarizado com a notação Σ, leia o Apêndice 1A deste capítulo antes de prosseguir.

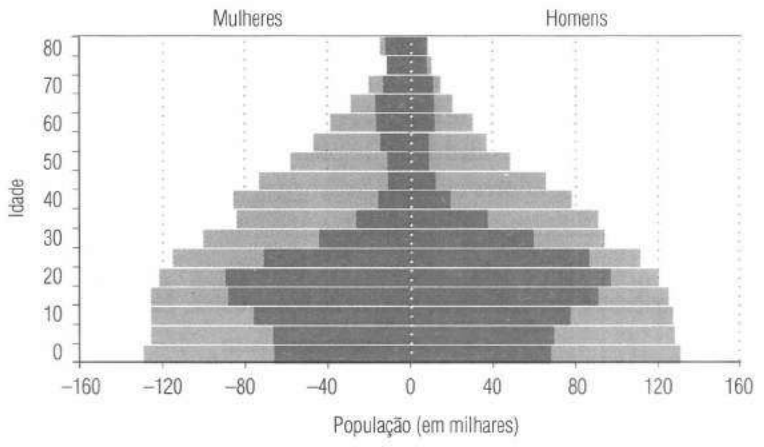
$$(1.2)\ \text{freqüência relativa} = \frac{\text{freqüência}}{\text{soma de freqüências}} = \frac{f}{\sum f}$$

A soma das freqüências relativas precisa ser igual a 100%, o que funciona como verificação dos cálculos realizados.

As freqüências acumuladas, mostradas na quarta coluna, são obtidas somando-se as freqüências sucessivamente. Elas indicam o número total de indivíduos com níveis de riqueza até um dado valor. Por exemplo, cerca de dez milhões de pessoas têm patrimônio inferior a £100.000.

A epidemia de aids

Para mostrar como a estatística descritiva pode ser útil na apresentação de informações, expomos abaixo a "pirâmide populacional" de Botsuana (um dos países mais seriamente afetados pela aids), projetada para 2020. A representação corresponde, em sua essência, a dois gráficos de barras (um para homens, outro para mulheres), dispostos lado a lado, apresentando as freqüências em cada uma das categorias de idade (em lugar de categorias de riqueza). A pirâmide interna (em cor mais escura) indica a população projetada, dada a existência da aids; a pirâmide externa supõe que não ocorram mortes por causa da aids.

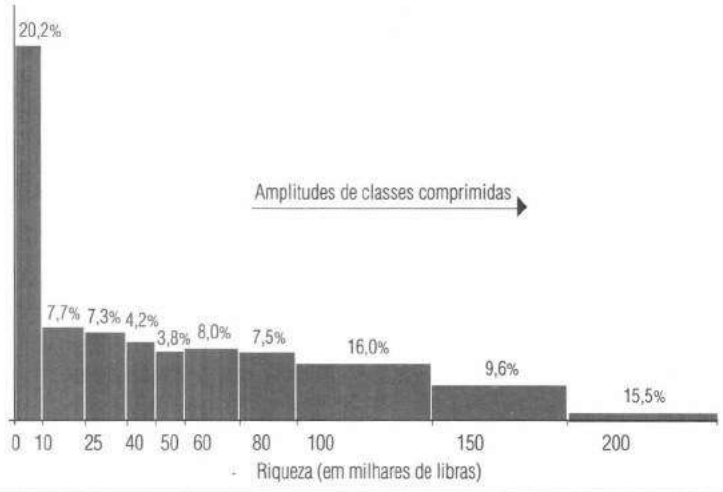


Fonte dos dados originais: US Census Bureau, *World population profile 2000*. Gráfico adaptado do site da Unaids: http://www.unaids.org/epidemic_update/report/Epi_report.htm#the population.

Pode-se perceber imediatamente o enorme impacto da aids, em especial sobre a faixa de 40 a 60 anos (atualmente com 20 a 40 anos), tanto para homens quanto para mulheres. Essas pessoas estariam normalmente na fase mais produtiva de sua vida, mas, com a aids, o país sofrerá bastante, pois terá muitas pessoas velhas e jovens dependendo de uma pequena população ativa. A gravidade dos problemas futuros é mostrada de maneira clara nesse diagrama simples, baseado no gráfico de barras.

FIGURA 1.10

Distribuição de freqüências relativas de riqueza no Reino Unido, 2001

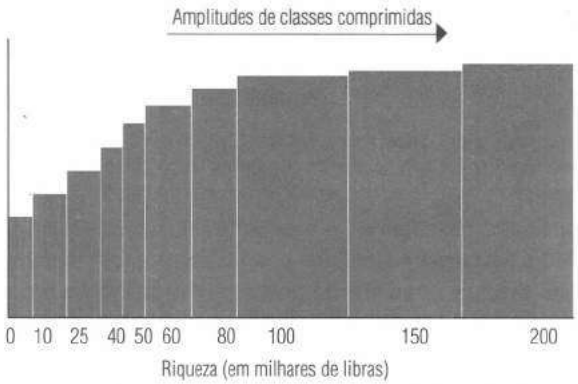


A distribuição de freqüências relativa e acumulada pode ser desenhada de maneira semelhante à do histograma. Na verdade, a distribuição de freqüência relativa tem exatamente a mesma forma da distribuição de freqüência. Isso é mostrado na figura 1.10. Dessa vez, escrevemos as freqüências relativas acima da coluna correspondente, embora isso não seja essencial.

A distribuição de freqüências acumuladas é fornecida na figura 1.11, na qual a altura dos blocos aumenta à medida que se eleva o nível de riqueza. A maneira mais simples de fazer isso é acumulando as densidades de freqüência (apresentadas na última coluna da tabela 1.4) e usando esses valores como coordenadas no eixo vertical.

FIGURA 1.11

Distribuição de freqüências acumuladas de riqueza no Reino Unido, 2001



Nota: As coordenadas do eixo vertical são obtidas acumulando-se as densidades de freqüência da tabela 1.4. Por exemplo, as duas primeiras coordenadas no eixo vertical são 0,3417 e 0,4286.

Exemplo resolvido 1.1

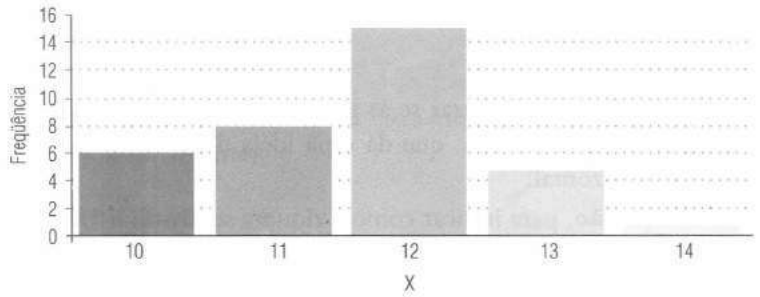
Como há grande volume de detalhes nas seções anteriores, o exemplo resolvido apresentado a seguir visa enfatizar os cálculos básicos necessários para produzir os gráficos resumidos. São utilizados dados artificiais deliberadamente para evitar uma interpretação longa dos resultados e de seu significado. Os dados da variável X e suas freqüências f são fornecidos na tabela a seguir, junto com os cálculos exigidos:

X	Freqüência, f	Freqüência relativa	Freqüência acumulada, F
10	6	0,17	6
11	8	0,23	14
12	15	0,43	29
13	5	0,14	34
14	1	0,03	35
Total	35	1,00	

Notas: Os valores de X são pontuais, mas poderiam representar o ponto médio de uma faixa, como anteriormente considerado.
As freqüências relativas são assim calculadas: 0,17 = 6/35; 0,23 = 8/35, etc.
As freqüências acumuladas são assim calculadas: 14 = 6 + 8; 29 = 6 + 8 + 15, etc.
O símbolo F designa, em geral, a freqüência acumulada nos trabalhos da área de estatística.

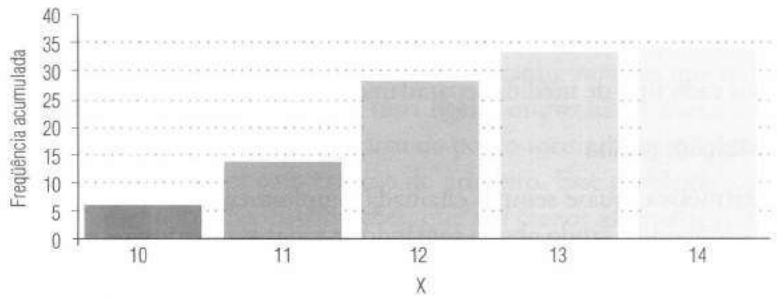
O gráfico de barras e a distribuição de freqüências acumuladas resultantes são:

Gráfico de barras da variável X



e

Distribuição de freqüências acumuladas de X



Exercício 1.2

Considerando os dados a seguir:

Faixa	Freqüência
0-10	20
11-30	40
31-60	30
61-100	20

a) Desenhe um gráfico de barras e um histograma dos dados e compare-os.
b) Calcule as freqüências acumuladas e desenhe um gráfico com essas freqüências.

Síntese de dados com o uso de técnicas numéricas

Os métodos gráficos representam um jeito excelente de oferecer uma visão geral rápida dos dados, mas não são particularmente precisos e não se prestam a análises adicionais. Para esse fim, precisamos recorrer a medidas numéricas, como a média.

Há várias maneiras distintas pelas quais podemos descrever uma distribuição como a de riqueza. Se pensássemos na possibilidade de tentar descrever o histograma, seria útil contar com:

- Uma **medida de posição**, para mostrar se as pessoas possuem muita ou pouca riqueza. Um exemplo disso é a média, que dá uma idéia de onde está localizada a distribuição no eixo horizontal.
- Uma **medida de dispersão**, para indicar como a riqueza se distribui em torno (geralmente) da média; se é concentrada perto da média ou se está muito afastada dela. Um exemplo disso é o desvio-padrão.
- Uma **medida de assimetria**, para mostrar quão simétrica ou não é a distribuição, ou seja, se a metade esquerda da distribuição é uma imagem da metade direita refletida em um espelho ou não. Obviamente, isso não ocorre na distribuição de riqueza.

Examinemos cada tipo de medida separadamente.

Medidas de posição: média

A **média aritmética**, quase sempre chamada simplesmente de média, é a medida de posição mais conhecida, sendo obtida somando-se todas as observações e dividindo-se essa soma pelo número de observações. Representamos a riqueza da i-ésima unidade familiar por x_i (de modo que o índice i vai de 1 a N, sendo N o número de observa-

ções). Por exemplo, x_3 é a riqueza da terceira unidade familiar. Assim sendo, a média é dada pela seguinte fórmula:

$$(1.3)\quad \mu = \frac{\sum_{i=1}^{i=N} x_i}{N}$$

em que a letra grega μ (pronuncia-se "mi") indica a média e $\sum_{i=1}^{i=N} x_i$ (lido assim: "sigma x i, para i = 1 a i = N", sendo Σ a letra maiúscula grega sigma) representa a soma dos valores de x. Isso pode ser simplificado para:

$$(1.4)\quad \mu = \frac{\Sigma x}{N}$$

o que deixa claro que valores de x estão sendo somados (geralmente, todas as observações estão disponíveis). Essa última forma, de leitura mais fácil, será em geral a utilizada neste livro.

A fórmula 1.3 pode ser usada somente quando todos os valores individuais de x são conhecidos. A tabela de freqüências não mostra as 17 milhões de observações, mas apenas a faixa de valores de cada intervalo de classe e a freqüência correspondente. Nesse caso de dados agrupados, pode-se utilizar a seguinte fórmula:

$$(1.5)\quad \mu = \frac{\sum_{i=1}^{i=C} f_i x_i}{\sum_{i=1}^{i=C} f_i}$$

ou, mais simplificadamente,

$$(1.6)\quad \mu = \frac{\Sigma fx}{\Sigma f}$$

Nessa fórmula:

- x representa o **ponto médio** de cada intervalo de classe, pois os valores individuais de x são desconhecidos. O ponto médio é utilizado como valor representativo de x para cada classe. No primeiro intervalo, por exemplo, não sabemos exatamente onde está cada uma das 3.417 observações. Portanto, *supomos* que todas estejam no ponto médio, £5.000. Isso causará uma ligeira imprecisão – como a distribuição é assimétrica, haverá mais famílias abaixo do ponto médio do que acima dele em cada intervalo de classe, talvez com exceção do primeiro. Esse problema é aqui ignorado, e é menor na maioria das distribuições, que são menos assimétricas que essa.
- A soma vai de 1 a C, o número de intervalos de classe, ou de valores distintos de x. O produto entre f e x nos dá a riqueza total em cada intervalo de classe. Se so-

marmos o conjunto de 14 intervalos de classe, obteremos a riqueza total de todos os indivíduos.

- $\sum f_i = N$ fornece o número total de observações, ou seja, a soma das freqüências individuais. O cálculo da média, μ, para os dados de riqueza é fornecido na tabela 1.6.

Nesse caso, obtemos:

$$\mu = \frac{2.225.722,5}{16.933} = 131,443$$

Note que os valores de x são expressos em milhares de libras, o que nos lembra que a média também será medida em milhares de libras; portanto, a riqueza média é igual a £131.443. Observe que as freqüências também foram divididas por 1.000, mas isso não tem efeito algum sobre o cálculo da média, pois f aparece tanto no numerador quanto no denominador da fórmula da média.

A média nos diz que se a riqueza total fosse dividida igualmente entre todos os indivíduos, cada um teria £131.443. Esse valor pode parecer surpreendente, pois o his-

TABELA 1.6 Cálculo da riqueza média

Faixa	x	f	fx
0-	5,0	3.417	17.085,0
10.000-	17,5	1.303	22.802,5
25.000-	32,5	1.240	40.300,0
40.000-	45,0	714	32.130,0
50.000-	55,0	642	35.310,0
60.000-	70,0	1.361	95.270,0
80.000-	90,0	1.270	114.300,0
100.000-	125,0	2.708	338.500,0
150.000-	175,0	1.633	285.775,0
200.000-	250,0	1.242	310.500,0
300.000-	400,0	870	348.000,0
500.000-	750,0	367	275.250,0
1.000.000-	1.500,0	125	187.500,0
2.000.000-	3.000,0	41	123.000,0
Total		16.933	2.225.722,5

Nota: A coluna fx fornece o produto dos valores das colunas f e x (assim, por exemplo, 5,0 × 3.417 = 17.085,0, que é a riqueza total que os indivíduos pertencentes ao primeiro intervalo possuem). A soma dos valores de fx fornece a riqueza total.

tograma mostra que a maioria das pessoas tem patrimônio menor que esse (aproximadamente 70% dos indivíduos estão de fato abaixo da média). A média não parece ser típica da riqueza da maioria das pessoas. O motivo para a média apresentar um valor tão elevado é a existência de alguns indivíduos cuja riqueza está muito acima de £131.443 – chegando aos milhões de libras, na verdade. A média é o "ponto de equilíbrio" da distribuição – se o histograma fosse um modelo físico, ele se equilibraria sobre um fulcro situado em 131.443. Os poucos níveis altos de riqueza exercem muita força e contrabalançam os indivíduos mais numerosos abaixo da média.

Exemplo resolvido 1.2

Imagine que haja dez famílias, cada uma com um único televisor em sua residência, 12 famílias com dois televisores cada uma, e três famílias com três televisores cada uma. Você pode calcular mentalmente que há um total de 43 televisores (10 + 24 + 9) pertencentes a 25 famílias (10 + 12 + 3). O número médio de televisores por família, portanto, seria 43/25 = 1,72. Expressando isso mais formalmente, temos (tal como na distribuição de riqueza, mas numa situação mais simples):

X	f	fx
1	10	10
2	12	24
3	3	9
Total	25	43

Isso gera a média de 1,72. Note que nossos dados são valores discretos nesse caso e que temos os valores exatos, e não um intervalo amplo de classe.

A média como valor esperado

Também nos referimos à média como o **valor esperado** de x e escrevemos:

(1.7) $E(x) = \mu = 131.443$

E(x) é lido "E de x" ou "valor esperado de x". A média é o valor esperado no sentido de que, se selecionarmos uma família ao acaso da população, "esperamos" que seu nível de riqueza seja igual a £131.443. É importante observar que se trata de expectativa *estatística*, e não do sentido mais usado do termo. A maior parte dos indivíduos selecionados ao acaso tem riqueza substancialmente inferior a esse valor. Portanto, a maioria das pessoas talvez "esperasse" um valor mais baixo, porque essa é sua experiência no dia-a-dia; mas os estatísticos são diferentes e sempre esperam o valor médio.

A notação de valor esperado é particularmente útil para acompanhar os efeitos de certas transformações de dados sobre a média (por exemplo, dividir a riqueza por

1.000 também divide a média por 1.000); o Apêndice 1B traz uma explicação detalhada sobre isso. Também se usa o operador E em inferência estatística para descrever as propriedades de estimadores (veja o capítulo 4).

Média amostral e média da população

Freqüentemente, dispomos somente de uma amostra dos dados (como no exemplo resolvido antes), e é importante distinguir esse caso daquele em que temos todas as observações possíveis. Por causa disso, a média amostral é dada por:

$$(1.8)\quad \bar{x} = \frac{\Sigma x}{n} \text{ ou } \bar{x} = \frac{\Sigma fx}{\Sigma f} \text{ para dados agrupados}$$

Note as diferenças entre μ (a média da população) e $\bar{x}$ (a média da amostra), e entre N (o tamanho da população) e n (o tamanho da amostra). Com exceção desses aspectos, os cálculos são idênticos. Por convenção, usam-se letras gregas, como μ, quando nos referimos à população, e letras romanas, como $\bar{x}$, em referência a uma amostra.

Média ponderada

Às vezes, as observações precisam receber pesos diferentes no cálculo da média, como indica o exemplo a seguir. Considere o problema de cálculo do gasto médio por aluno por parte de uma escola. Na tabela 1.7 são fornecidos os dados de gasto referentes a alunos de nível primário (idade entre cinco e 11 anos), secundário (de 11 a 16 anos) e com idade superior a 16 anos.

Fica bastante claro que se gasta muito mais com alunos de nível secundário e com os de idade acima de 16 anos (um padrão generalizado na Inglaterra e na maioria dos outros países), e que a média deve estar em algum ponto entre 890 e 1.910. Entretanto, se tirássemos uma média simples desses três valores, obteríamos a resposta errada, pois pode haver quantidades muito diferentes de crianças nas três faixas etárias. Os números e as proporções de crianças em cada grupo etário são apresentados na tabela 1.8.

Como há relativamente mais crianças no nível primário do que no secundário, e menos alunos com mais de 16 anos, o custo unitário do ensino primário deve receber o maior peso no cálculo da média, e o custo unitário do ensino de alunos com mais de 16 anos, o menor peso. A **média ponderada** é obtida multiplicando-se cada custo unitário pela proporção de crianças em cada categoria e depois somando os resultados. Portanto, a média ponderada é:

$$(1.9)\quad 0,444 \times 890 + 0,389 \times 1.450 + 0,167 \times 1.910 = 1.277,8$$

A média ponderada fornece uma resposta mais próxima do custo unitário do ensino primário do que a média simples (1.416,7, nesse caso) dos três valores, o que seria enganoso. A fórmula da média ponderada é:

$$(1.10)\quad \bar{x}_w = \Sigma_i w_i x_i$$

em que w representa os pesos, *cuja soma deve ser igual a 1*, ou seja:

Cálculo de sua nota final

Se você é estudante universitário, sua nota final provavelmente será calculada pela média ponderada de suas notas nas várias disciplinas. Os pesos podem estar baseados nos créditos associados a cada disciplina, ou em algum outro fator. Por exemplo, na minha universidade os estudantes que cursavam a faculdade de direito, passando um ano no exterior, tinham sua nota final (a média geral, G) calculada da seguinte maneira:

$$G = \frac{0,75L + 0,25S + 0,25Y}{1,25}$$

em que L corresponde à sua nota na área de direito, S é a nota no curso fora da área de direito e Y é a nota da dissertação referente ao trabalho feito no exterior.

Note que o peso do curso principal é, de fato, 0,75/1,25 = 0,60, e não 0,75, mas é mais fácil expressar a fórmula no formato acima. Pela minha experiência, muitas pessoas (mesmo alguns professores e membros da administração da escola) têm dificuldade para calcular a média ponderada, e por esse motivo é bom você conferir se a sua foi calculada corretamente. Esse é, sem dúvida, um bom motivo para aprender a calcular a média ponderada.

$$(1.11)\quad \Sigma_i w_i = 1$$

e x representa os valores do custo unitário.

Mediana

Retornando ao estudo da riqueza, o resultado pouco representativo para a média sugere que podemos dar preferência a uma medida de posição que não seja fortemente afetada por observações extremas e pela presença de assimetria.

A **mediana**, definida pelo procedimento a seguir, é uma medida de posição mais robusta a tais valores extremos, isto é, menos propensa a ser afetada por eles. Imagine que todas as pessoas sejam dispostas numa linha que vai do mais pobre ao mais rico.

TABELA 1.7 Custo por aluno por grau de escolaridade (£ por ano)

	Primário	Secundário	Mais de 16 anos
Custo unitário	890	1.450	1.910

TABELA 1.8 Número e proporção de alunos em cada faixa etária

	Primário	Secundário	Mais de 16 anos	Total
Número	8.000	7.000	3.000	18.000
Proporção	44,4%	38,9%	16,7%	

Procure o indivíduo localizado na metade da linha e pergunte-lhe qual é o seu nível de riqueza. A resposta é a mediana. Claramente, a mediana, ao contrário da média, não é afetada por valores extremos: dobrar o nível de riqueza da pessoa mais rica (sem redução da riqueza de qualquer outra) não produz efeito algum sobre a mediana. O cálculo da mediana não é tão simples quanto o da média, em particular no caso de dados agrupados. O seguinte exemplo resolvido mostra como a mediana é calculada para dados não-agrupados.

Exemplo resolvido 1.3

Mediana

Calcule a mediana dos seguintes valores: 45, 12, 33, 80, 77.

Inicialmente, colocamos os valores em ordem crescente: 12, 33, 45, 77, 80.

Isso facilita ver que o valor intermediário é 45. Esta é a mediana. Note que, se o valor da maior observação subir, digamos, para 150, o valor da mediana não mudará. Já o valor da média, com esse aumento, mudará de 49,4 para 63,4.

Se o número de observações for par, não haverá observação intermediária. A solução será tirar a média das duas observações intermediárias. Por exemplo:

Calcule a mediana de 12, 33, 45, 63, 77, 80.

Note a presença de uma nova observação, 63, o que faz com que o total seja de seis observações. O valor da mediana está na metade do caminho entre a terceira e a quarta observações, isto é, (45 + 63)/2 = 54.

No caso de dados agrupados, o cálculo é feito em duas etapas: primeiro, devemos identificar o intervalo de classe que contém o indivíduo mediano; depois, devemos calcular a posição dessa pessoa no intervalo. Para ilustrar, calcularemos a mediana dos dados de riqueza:

1. Identificação do intervalo de classe apropriado: como há 16.933.000 observações, precisamos descobrir qual é o patrimônio da pessoa que está na posição 8.466.500. A tabela de freqüências acumuladas (veja a tabela 1.5) é a mais adequada para essa finalidade. Há 7.316.000 indivíduos com patrimônio inferior a £60.000 e 8.677.000 com menos de £80.000. A pessoa intermediária, portanto, está na classe de riqueza de £60.000 a £80.000. Além disso, como 8.466.500 é um número muito mais próximo de 8.677.000, percebe-se que a mediana está perto do limite superior desse intervalo. Em seguida, procuramos tornar essa afirmação mais precisa.
2. Para determinar a posição dentro do intervalo, agora podemos usar a fórmula 1.12:

$$(1.12)\quad \text{mediana} = x_L + (x_U - x_L)\left\{\frac{\frac{N+1}{2} - F}{f}\right\}$$

em que:

x_L = limite inferior do intervalo contendo a mediana;
x_U = limite superior desse intervalo de classe;
N = número de observações (o uso de N + 1 em lugar de N na fórmula torna-se importante quando o valor de N é relativamente pequeno);
F = freqüência acumulada dos intervalos de classe até o intervalo que contém a mediana, mas sem incluir esse intervalo;
f = freqüência do intervalo de classe contendo a mediana.

A expressão entre chaves nos diz quanto é preciso avançar no intervalo até chegar à mediana.

No caso da distribuição de riqueza, temos:

$$\text{mediana} = 60.000 + (80.000 - 60.000)\left\{\frac{\frac{16.933.000}{2} - 7.316.000}{1.361.000}\right\} = £76.907$$

Essa medida alternativa de posição passa uma impressão muito diferente: é pouco superior à metade da média. Apesar disso, é igualmente válida, a despeito de ter um significado distinto. Demonstra que a pessoa que está "no meio" possui patrimônio de £76.907 e, nesse sentido, é típico da população do Reino Unido. Note que a mediana está realmente perto do limite superior desse intervalo. A expressão entre chaves vale 0,84, indicando que a mediana se situa 84% acima do limite inferior do intervalo. Antes de comparar essas medidas, devemos examinar a moda, uma terceira medida.

Generalização da mediana – quantis

A idéia da mediana como meio de distribuição pode ser ampliada: os **quartis** dividem a distribuição em quatro partes iguais, os **quintis**, em cinco partes, os **decis**, em dez partes, e, finalmente, os **percentis** dividem a distribuição em cem partes iguais. Genericamente, são conhecidos como **quantis**. Ilustraremos essa idéia examinando os decis (os quartis serão discutidos mais adiante).

O primeiro decil ocorre a um décimo da linha que representa as pessoas ordenadas da mais pobre à mais rica. Isso significa que precisamos saber qual é o patrimônio da pessoa colocada na posição 1.693.300 (= N/10) da distribuição. Com base na tabela de freqüências acumuladas, essa pessoa situa-se no primeiro intervalo de classe. Adaptando a fórmula 1.12, obtemos:

$$\text{primeiro decil} = 0 + (10.000 - 0) \times \left\{\frac{1.693.300 - 0}{3.417.000}\right\} = £\,4.956$$

Assim, estimamos que qualquer família com patrimônio inferior a £4.956 situe-se nos 10% inferiores da distribuição de riqueza. De maneira análoga, o nono decil pode ser encontrado calculando-se o patrimônio da família na posição 15.239.700 (= N × 9/10) da distribuição.

Moda

A **moda** é definida pelo nível de riqueza que ocorre com maior freqüência; em outras palavras, o valor que ocorre mais vezes. É mais útil e de cálculo mais fácil quando se tem acesso a todos os dados e há relativamente poucas observações diferentes. Isso é o que acontece no exemplo simples apresentado a seguir.

Suponhamos os seguintes dados de vendas de vestidos numa loja, de acordo com o tamanho dos vestidos:

Tamanho	Vendas
8	7
10	25
12	36
14	11
16	3
18	1

O tamanho modal é 12. Há mais mulheres comprando vestidos desse tamanho do que de qualquer outro. Essa pode ser a medida de posição mais útil para a loja. Embora ela precise estocar tamanhos diferentes, sabe que deve encomendar um número maior de vestidos tamanho 12. A média não seria tão útil nesse caso (é igual a $\bar{x} = 11,7$), pois realmente não existe um vestido desse tamanho.

Para dados agrupados, a questão é mais complicada. Nesse caso, exige-se o intervalo de classe modal, uma vez que os intervalos tenham sido ajustados por suas amplitudes (caso contrário, um intervalo mais amplo será indevidamente comparado a um intervalo mais estreito). Para esse fim, podemos usar novamente a densidade de freqüência. Com base na tabela 1.4, vemos que o intervalo que apresenta a maior densidade de freqüência é o primeiro, de £0 a £10.000. Ele é "típico" da distribuição por ser o intervalo que ocorre mais freqüentemente (usando as densidades de freqüência, e *não* as freqüências). A distribuição de riqueza é mais concentrada nesse nível, e mais pessoas têm a mesma riqueza. Mais uma vez, é digna de nota a diferença em relação tanto à mediana quanto à média.

As três medidas de posição transmitem mensagens distintas por causa da assimetria da distribuição: se ela fosse simétrica, todas dariam aproximadamente a mesma resposta. Temos aqui um caso extremo de assimetria, mas ele serve para ilustrar a comparação das diferentes medidas de posição. Quando a distribuição é assimétrica à direita, como ocorre aqui, a ordem das três medidas é moda, mediana, média; se é assimétrica à esquerda, a ordem é invertida. E se a distribuição tiver mais de um pico, essa regra de ordenamento das medidas de posição poderá não funcionar.

Qual das medidas é "correta" ou mais útil? Nesse caso específico, a média não é muito útil: ela é fortemente influenciada por valores extremos. Portanto, a mediana é mais utilizada quando se discute distribuição de riqueza e renda. Quando a desigualdade é muito pronunciada, como ocorre em alguns países menos desenvolvidos, a média é ainda menos informativa. A moda também é bastante útil como informação a respeito de uma faixa importante da população, embora possa ser sensível ao modo pelo qual os intervalos de classe são montados. Se houvesse um intervalo de £5.000 a £15.000, essa poderia muito bem ser a classe modal, dando uma impressão ligeiramente diferente.

As três medidas de posição são assinaladas no histograma da figura 1.12. Ele mostra as diferenças substanciais entre as medidas de posição quando temos uma distribuição assimétrica, como a de riqueza.

Exercício 1.3

a) Considerando os dados fornecidos no exercício 1.2, calcule a média, a mediana e a moda.

b) Assinale esses valores no histograma que você desenhou no exercício 1.2.

FIGURA 1.12

Histograma com média, mediana e moda assinaladas

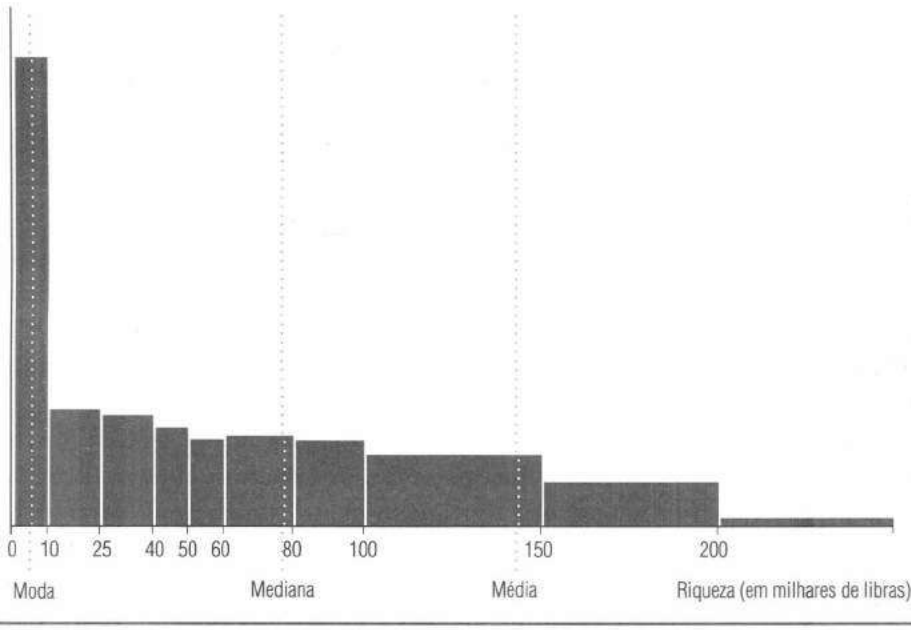
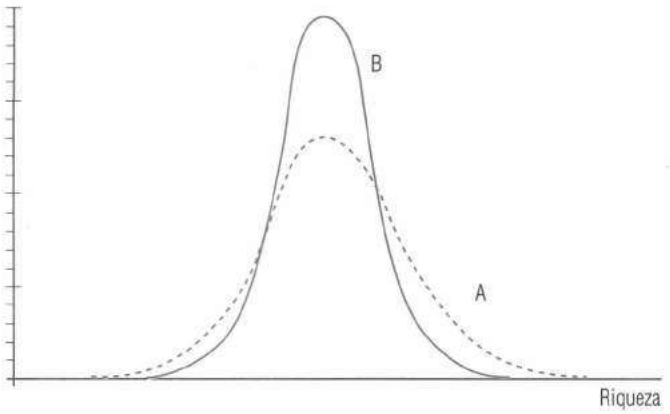


FIGURA 1.13

Duas distribuições com graus diferentes de dispersão



Nota: A distribuição A tem um grau maior de dispersão do que B, na qual todos têm níveis semelhantes de riqueza.

Medidas de dispersão

Duas distribuições distintas (por exemplo, distribuição de riqueza em dois países) poderiam ter a mesma média e, no entanto, parecer muito diferentes, como vemos na figura 1.13 (as distribuições foram traçadas com curvas ligando os pontos, em lugar de barras, para aumentar a clareza de sua apresentação). Em um dos países, seria possível que todos tivessem níveis semelhantes de riqueza (curva B). No outro, embora a média seja a mesma, poderiam existir extremos de riqueza e pobreza (curva A). Uma medida de dispersão é um número que nos permite diferenciar essas duas situações.

A medida mais simples de dispersão é a **amplitude**, ou seja, a diferença entre as observações mínima e máxima. É impossível calcular esse valor na tabela de níveis de riqueza, pois a observação máxima não está disponível. De qualquer forma, não é um indicador muito útil, pois depende de dois valores extremos e ignora o restante da distribuição. Em casos mais simples, poderia ser mais informativo. Por exemplo, num exame as notas podem variar de um mínimo de 28% a um máximo de 74%. Nesse caso, a amplitude é 74 – 28 = 46, o que representa uma informação útil.

Um aperfeiçoamento disso é o **intervalo entre quartis**, ou seja, a diferença entre o primeiro e o terceiro quartil. Define, portanto, os limites de riqueza da metade interna da distribuição. Para calcular o primeiro quartil (que representaremos por Q_1), precisamos avançar até um quarto da linha de indivíduos possuidores de riqueza (ordenados dos mais pobres aos mais ricos) e perguntar à pessoa nessa posição qual é o seu patrimônio. A resposta é o primeiro quartil. O cálculo é o seguinte:

- um quarto de 16.933 é 4.233,25;
- a pessoa situada na posição 4.233,25 pertence ao intervalo entre £10.000 e £25.000;
- adaptando a fórmula 1.12:

$$(1.13)\quad Q_1 = 10.000 + (25.000 - 10.000)\left\{\frac{4.233,25 - 3.417}{1.303}\right\} = 19.396,58$$

O terceiro quartil é calculado seguindo-se procedimento semelhante:

- três quartos de 16.933 é 12.699,75;
- a pessoa situada na posição 12.699,75 está no intervalo entre £150.000 e £200.000;
- novamente, usando a fórmula 1.12:

$$Q_3 = 150.000 + (200.000 - 150.000)\left\{\frac{12.699,75 - 12.655}{1.633}\right\} = 151.370,18$$

Portanto, o intervalo entre quartis é $Q_3 - Q_1 = 151.370 - 19.396 = 131.974$.

O cálculo fornece uma medida sintética da dispersão da distribuição: ela é maior quanto maior o seu valor. Duas distribuições diferentes de riqueza poderiam ser comparadas por meio de seus intervalos entre quartis, e o país que apresentasse o valor mais elevado seria o país com maior desigualdade de riqueza. Observe que os valores precisariam ser medidos numa mesma moeda para que essa comparação fosse válida.

Exemplo resolvido 1.4

Amplitude e intervalo entre quartis (2) intervalo entre quartis

Suponhamos que 110 crianças se submetessem a um teste, cujos resultados estão registrados na tabela abaixo:

Nota, X	Freqüência, f	Freqüência acumulada, F
13	5	5
14	13	18
15	29	47
16	33	80
17	17	97
18	8	105
19	4	109
20	1	110
Total	110	

A amplitude é simplesmente 20 – 13 = 7. O intervalo entre quartis exige o cálculo dos quartis. Q_1 é dado pelo valor da 27,5ª observação (= 110/4), que é igual a 15. Q_3 é

o valor da 82,5ª observação (= 110 x 0,75), que é igual a 17. O intervalo entre quartis (IQ) é, portanto, igual a 17 – 15 = 2 pontos. Metade dos alunos obteve notas nessa faixa.

Note que uma pequena alteração nos dados (três alunos obtendo 16 pontos, em lugar de 17) mudaria o IQ em 1 ponto (16 – 15). O resultado deve ser tratado com certo cuidado, portanto. Esse é um problema freqüente quando há poucos valores distintos da variável (oito, neste exemplo).

Variância

Uma medida mais útil de dispersão é a **variância**, que utiliza toda a informação disponível, em lugar de desprezar os extremos da distribuição. A variância é representada pelo símbolo σ^2 (σ é a letra grega sigma minúscula, de modo que σ^2 se lê "sigma ao quadrado"). O σ tem um significado completamente diferente do Σ (sigma maiúsculo), utilizado anteriormente. Sua fórmula é:[3]

$$(1.14)\quad \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

A variância é a média dos quadrados de todos os desvios em relação à média. Uma distribuição com maior dispersão (tal como A na figura 1.13) tenderá a ter desvios maiores em relação à média e, portanto, uma variância mais elevada. Ao comparar duas distribuições com médias semelhantes, portanto, devemos examinar suas variâncias para ver qual das duas tem o maior grau de dispersão. No caso de dados agrupados, a fórmula passa a ser:

$$(1.15)\quad \sigma^2 = \frac{\Sigma f(x_i - \mu)^2}{\Sigma f}$$

O cálculo da variância é apresentado na tabela 1.9. Obtemos, conseqüentemente:

$$\sigma^2 = \frac{895.418.240,28}{16.933} = 52.880,07$$

Esse valor é calculado antes da reconversão às unidades originais de medida, como foi feito no caso da média, ao multiplicá-la por 1.000. No caso da variância, porém, devemos multiplicar por 1.000.000, ou seja, o *quadrado* de 1.000. A variância, portanto, é igual a 52.880.070.000. A multiplicação pelo quadrado de 1.000 é conseqüência do uso de desvios ao quadrado na fórmula da variância (veja o Apêndice 1B a respeito dos operadores E e V para mais detalhes).

Assim, é preciso tomar certo cuidado com as unidades de medida. Se for dito que a média é igual a 131,443, então o apropriado é dizer que a variância é igual a

3. $x_i - \mu$ representa o desvio de x_i em relação à média. Esses desvios são primeiro elevados ao quadrado e depois somados. Dividindo-se a soma por N, obtém-se a variância.

TABELA 1.9 Cálculo da variância da riqueza

Intervalo	Ponto médio (x mil libras)	Freqüência, f	Desvio $(x - \mu)$	$(x - \mu)^2$	$f(x - \mu)^2$
0-	5,0	3.417	-126,4	15.987,81	54.630.329,97
10.000-	17,5	1.303	-113,9	12.982,98	16.916.826,55
25.000-	32,5	1.240	-98,9	9.789,70	12.139.223,03
40.000-	45,0	714	-86,4	7.472,37	5.335.274,81
50.000-	55,0	642	-76,4	5.843,52	3.751.537,16
60.000-	70,0	1.361	-61,4	3.775,23	5.138.086,73
80.000-	90,0	1.270	-41,4	1.717,51	2.181.241,95
100.000-	125,0	2.708	-6,4	41,51	112.411,42
150.000-	175,0	1.633	43,6	1.897,22	3.098.162,88
200.000-	250,0	1.242	118,6	14.055,79	17.457.288,35
300.000-	400,0	870	268,6	72.122,92	62.746.940,35
500.000-	750,0	367	618,6	382.612,90	140.418.932,52
1.000.000-	1.500,0	125	1.368,6	1.872.948,56	234.118.569,53
2.000.000-	3.000,0	41	2.868,6	8.228.619,88	337.373.415,02
Total		16.933			895.418.240,28

52.880,07. Mas se for dito que a média é 131.443, então deve-se dizer que a variância é 52.880.070.000. Note que só a forma de apresentação muda; os fatos subjacentes continuam os mesmos.

Desvio-padrão

Em que unidade a variância é medida? Como ela envolve um procedimento de cálculo de quadrados, acabamos com libras esterlinas ao quadrado, o que não é muito conveniente. Por causa disso, definimos o **desvio-padrão** como a raiz quadrada da variância, o que nos coloca novamente em libras. O desvio-padrão, portanto, é dado por:

$$(1.16)\quad \sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

ou, para dados agrupados:

$$(1.17)\quad \sigma = \sqrt{\frac{\Sigma f(x - \mu)^2}{\Sigma f}}$$

Essas expressões são simplesmente a raiz quadrada das expressões 1.14 e 1.15. O desvio-padrão da riqueza, portanto, é igual a $\sqrt{52.880,07} = 229,957$. Isso é medido em milhares de libras, de modo que o desvio-padrão realmente é igual a £229.957 (note que esse valor é a raiz quadrada de 52.880.070.000, como era de esperar). Por si mesmo, o desvio-padrão e a variância não têm fácil interpretação, pois não representam algo que tenha um significado intuitivo, ao contrário da média. São mais úteis quando utilizados num contexto comparativo. Isso será mostrado adiante.

Variância e desvio-padrão de uma amostra

Tal como ocorre com a média, um símbolo diferente é utilizado para distinguir entre uma variância calculada na população e uma calculada numa amostra. Além disso, para se chegar à variância da amostra, utiliza-se uma fórmula ligeiramente diferente daquela usada para calcular a variância da população. A fórmula da variância da amostra, simbolizada por s^2, é dada pelas equações 1.18 e 1.19:

$$(1.18)\quad s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

$$(1.19)\quad s^2 = \frac{\sum f(x - \bar{x})^2}{n - 1}$$

em que n é o tamanho da amostra. O motivo pelo qual se usa n – 1 no denominador, em lugar de n (como seria de esperar), é o seguinte: estamos realmente interessados na variância da população, e a variância da amostra é uma estimativa dela. A primeira é medida pela dispersão em relação a μ; e a segunda, em termos ideais, também deveria ser medida relativamente a μ. Entretanto, o valor de μ é desconhecido, e usa-se o valor de $\bar{x}$ em seu lugar. Mas a variação das observações da amostra em torno de $\bar{x}$ tende a ser menor que as variações em torno de μ. O uso de n – 1 em vez de n contrabalança esse efeito, e o resultado é uma estimativa **não-viesada**[4] (isto é, correta, em média) da variância da população.

O uso da fórmula correta é tanto mais importante quanto menor for o tamanho da amostra, pois aumenta a diferença relativa entre n – 1 e n. Por exemplo, se n = 10, o ajuste corresponde a 10% da variância; quando n = 100, o ajuste é apenas 1%.

O desvio-padrão da amostra é dado pela raiz quadrada da equação 1.18 ou da equação 1.19.

Exemplo resolvido 1.5

Variância e desvio-padrão

Continuamos com o exemplo resolvido antes, relacionado às notas obtidas por alunos num teste. *A variância e o desvio-padrão podem ser calculados da seguinte maneira:*

4. O conceito de *viés* será tratado mais detalhadamente no capítulo 4.

X	f	fx	$x - \mu$	$(x - \mu)^2$	$f(x - \mu)^2$
13	5	65	–2,81	7,89	39,45
14	13	182	–1,81	3,27	42,55
15	29	435	–0,81	0,65	18,98
16	33	528	0,19	0,04	1,20
17	17	289	1,19	1,42	24,11
18	8	144	2,19	4,80	38,40
19	4	76	3,19	10,18	40,73
20	1	20	4,19	17,56	17,56
Total	110	1.739			222,99

A média é 1.739/110 = 15,81, e, com base nesse resultado, calculam-se os valores na coluna de desvios $(x - \mu)$ (assim sendo, –2,81 = 13 – 15,81, etc.).

A variância é calculada pela fórmula $\sum f(x - \mu)^2/(n - 1) = 222,99/109 = 2,05$. Conseqüentemente, o desvio-padrão é igual a 1,43, ou seja, a raiz quadrada de 2,05. (Os cálculos são apresentados com duas casas decimais, mas foram efetuados com valores exatos.)

No caso de distribuições que são aproximadamente simétricas e têm a forma de um sino (ou seja, as observações estão concentradas perto da média), há uma relação aproximada entre o desvio-padrão e o intervalo entre quartis. Essa regra diz que o IQ é 1,3 vezes o desvio-padrão. Nesse caso, 1,3 × 1,43 = 1,86, próximo do valor 2 obtido anteriormente.

Fórmulas alternativas de cálculo da variância e do desvio-padrão

As fórmulas fornecidas a seguir geram as mesmas respostas que as equações 1.14 a 1.17, mas seu cálculo é mais simples, seja feito manualmente, seja usando uma planilha. Para a variância da população, pode-se usar:

$$(1.20)\quad \sigma^2 = \frac{\sum x^2}{N} - \mu^2$$

ou, para dados agrupados:

$$(1.21)\quad \sigma^2 = \frac{\sum fx^2}{\sum f} - \mu^2$$

O cálculo da variância com o emprego da equação 1.21 é demonstrado na figura 1.14. A variância da amostra pode ser calculada usando-se:

$$(1.22)\quad s^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1}$$

ou, no caso de dados agrupados:

$$(1.23)\quad s^2 = \frac{\sum fx^2 - n\bar{x}^2}{n-1}$$

FIGURA 1.14

Estatísticas descritivas calculadas com o uso de Excel

<>	A	B	C	D	E	F	G	H
1			DADOS DE RIQUEZA 2001					
2								
3	Faixa de	Ponto Médio	Freqüência					
4	Riqueza	x	f	fx	f*x ao quadrado		Estatísticas sintéticas	
5	0	5,0	3.417	17.085,0	85.425,0			
6	10.000	17,5	1.303	22.802,5	399.043,8		Média	131,443
7	25.000	32,5	1.240	40.300,0	1.309.750,0		Variância	52880,071
8	40.000	45,0	714	32.130,0	1.445.850,0		Desvio-padrão	229,957
9	50.000	55,0	642	35.310,0	1.942.050,0		Coeficiente de variação	1,749
10	60.000	70,0	1.361	95.270,0	6.668.900,0			
11	80.000	90,0	1.270	114.300,0	10.287.000,0			
12	100.000	125,0	2.708	338.500,0	42.312.500,0			
13	150.000	175,0	1.633	285.775,0	50.010.625,0			
14	200.000	250,0	1.242	310.500,0	77.625.000,0			
15	300.000	400,0	870	348.000,0	139.200.000,0			
16	500.000	750,0	367	275.250,0	206.437.500,0			
17	1.000.000	1500,0	125	187.500,0	281.250.000,0			
18	2.000.000	3000,0	41	123.000,0	369.000.000,0			
19								
20	Totais		16.933	2.225.722,5	1.187.973.643,8			
21								
22								
23								
24								
25								
26								

O desvio-padrão, evidentemente, pode ser obtido com o cálculo da raiz quadrada dessas fórmulas.

Uso de calculadora ou computador

As calculadoras eletrônicas e particularmente os computadores têm simplificado cálculos como o da média. A figura 1.14 mostra como montar os cálculos anteriores numa planilha (Microsoft Excel, nesse caso), incluindo algumas das fórmulas das células apropriadas.

Nesse caso, a variância é calculada com a fórmula:

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \mu^2$$

ou seja, a fórmula fornecida na equação 1.21, mostrada anteriormente. Note que ela gera o mesmo resultado.

As seguintes fórmulas estão contidas nas células:

D5:	= C5*B5	para calcular f vezes x
E5:	= D5*B5	para calcular f vezes x^2
C20:	= SOMA(C5:C18)	para somar as freqüências
H6:	= D20/C20	calcula $\sum fx/\sum f$
H7:	= E20/C20–H6^2	calcula $\sum fx^2/\sum f - \mu^2$
H8:	= RAIZ(H7)	calcula σ
H9:	= H8/H6	calcula σ/μ

Coeficiente de variação

As medidas de dispersão examinadas até agora são medidas de **dispersão absoluta** e, em particular, seus valores dependem das unidades com as quais a variável é medida. É difícil, portanto, comparar os graus de dispersão de duas variáveis medidas em unidades distintas. Por exemplo, não é possível comparar níveis de riqueza no Reino Unido com níveis de riqueza na Alemanha se no primeiro país os valores são medidos em libras e no segundo, em euros. Também não se poderiam comparar as distribuições de riqueza num país em dois momentos distintos, porque a inflação altera o valor da moeda com o passar do tempo. A solução é usar uma medida de **dispersão relativa** que seja independente das unidades de medida. Uma dessas medidas é o **coeficiente de variação**, definido da seguinte maneira:

$$(1.24)\quad \text{coeficiente de variação} = \frac{\sigma}{\mu}$$

ou seja, o desvio-padrão dividido pela média. Sempre que as unidades de medida são alteradas, o efeito sobre a média e sobre o desvio-padrão é idêntico, de modo que o coeficiente de variação não é afetado. No caso da distribuição de riqueza, seu valor é 229,957/131,443 = 1,749, ou seja, o desvio-padrão é 175% da média. Isso pode ser comparado diretamente com o coeficiente de variação de uma distribuição diferente de riqueza para determinar qual delas apresenta um grau relativo maior de dispersão.

Desvio-padrão do logaritmo

Outra solução para o problema de unidades diferentes de medida consiste em empregar o logaritmo[5] da riqueza em lugar de seu valor efetivo. O motivo pelo qual isso funciona pode ser adequadamente ilustrado com um exemplo. Suponhamos que entre 1994 e 2001 a riqueza de cada indivíduo tenha dobrado, de modo que $X_i^{2001} = 2X_i^{1994}$, em que X_i^t indica a riqueza do indivíduo i no ano t. O desvio-padrão de X^{2001} é, portanto, o dobro do desvio-padrão de X^{1994}. Calculando os logaritmos naturais, temos $\ln X_i^{2001} = \ln 2 + \ln X_i^{1994}$, de modo que a distribuição de $\ln X^{2001}$ é igual à de $\ln X^{1994}$, exceto pelo fato de ser deslocada para a direita por duas unidades. A variância (e, portanto, o desvio-padrão) das duas distribuições de logaritmos é a mesma, indicando não ter havido alteração da dispersão relativa das duas distribuições de riqueza.

O desvio-padrão dos logaritmos da riqueza é calculado na tabela 1.10. A variância, portanto, é:

$$\sigma^2 = \frac{306.673,6}{16.933} - \left(\frac{67.584,6}{16.933}\right)^2 = 2,181$$

e o desvio-padrão é $\sigma = 1,477$. Veja mais adiante uma comparação da distribuição de 2001 com a distribuição de 1979, pouco antes da eleição de um governo conservador que dirigiu a Grã-Bretanha durante a década de 1980 e boa parte dos anos 90.

5. Veja o Apêndice 1C caso não esteja familiarizado com logaritmos.

TABELA 1.10 Cálculo do desvio-padrão do logaritmo da riqueza

Faixa	Ponto médio × (£ 000)	ln (x)	Freqüência, f	fx	fx^2
0-	5,0	1,609	3.417	5.499,4	8.851,0
10.000-	17,5	2,862	1.303	3.729,4	10.674,4
25.000-	32,5	3,481	1.240	4.316,7	15.027,6
40.000-	45,0	3,807	714	2.718,0	10.346,3
50.000-	55,0	4,007	642	2.572,7	10.309,7
60.000-	70,0	4,248	1.361	5.782,2	24.565,7
80.000-	90,0	4,500	1.270	5.714,8	25.715,3
100.000-	125,0	4,828	2.708	13.075,1	63.130,6
150.000-	175,0	5,165	1.633	8.434,1	43.560,3
200.000-	250,0	5,521	1.242	6.857,7	37.864,3
300.000-	400,0	5,991	870	5.212,6	31.231,0
500.000-	750,0	6,620	367	2.429,6	16.083,9
1.000.000-	1.500,0	7,313	125	914,2	6.685,4
2.000.000-	3.000,0	8,006	41	328,3	2.628,2
Total			16.933	67.584,6	306.673,6

Nota: Use a tecla "ln" em sua calculadora ou a função =LN() numa planilha para obter os logaritmos naturais dos dados. Você deve obter ln 5 = 1,609, ln 17,5 = 2,862, etc.

Mensuração de desvios em relação à média: escores z

Imagine o seguinte problema: um homem e uma mulher estão discutindo sobre desempenho em suas carreiras e o homem diz que, por ganhar mais do que ela, é mais bem-sucedido. A mulher responde que as mulheres são vítimas de discriminação e que, em relação às mulheres, ela está tendo mais êxito do que o homem em comparação a outros homens. Essa discussão pode ser dirimida?

Suponhamos que os dados sejam os seguintes: o salário médio dos homens é £19.500, e o das mulheres, é £16.800. Já o desvio-padrão dos salários masculinos é £4.750, enquanto, para as mulheres, é £3.800. O salário desse homem equivale a £31.375, e o dessa mulher é £26.800. O homem, portanto, está £11.875 acima da média, e a mulher, £10.000 acima da média correspondente. Entretanto, a dispersão dos salários das mulheres é menor do que a dos salários dos homens; logo, a mulher tem bom desempenho ao obter um salário de £26.800.

Uma das maneiras de resolver a questão é calcular o **escore z**, que nos dá o salário em termos de número de desvio-padrão em relação à média. Portanto, para o homem, o escore z é:

$$(1.25)\quad z = \frac{x - \mu}{\sigma} = \frac{31.375 - 19.500}{4.750} = 2,50$$

Assim sendo, o homem se encontra 2,5 desvios-padrão acima do salário médio masculino. Para a mulher, o cálculo é:

$$(1.26)\quad z = \frac{26.800 - 16.800}{3.800} = 2,632$$

A mulher está 2,632 desvios-padrão acima de sua média e, portanto, ganha a discussão – ela está mais perto do alto de sua distribuição do que o homem e, dessa maneira, corresponde mais a um valor extremo.

Desigualdade de Chebyshev

O uso do escore z leva naturalmente à **desigualdade de Chebyshev**, que nos diz qual é a proporção de observações situadas nas caudas de qualquer distribuição, independentemente de sua forma. O enunciado do teorema é o seguinte:

(1.27) Pelo menos $(1 - 1/k^2)$ das observações em qualquer distribuição situa-se a menos de k desvios-padrão da média.

Se tomarmos a distribuição de salários de mulheres mencionada acima, podemos nos perguntar que proporção de mulheres estará a mais de 2,632 desvios-padrão da média (nas duas caudas da distribuição). Sendo k = 2,632, então $(1 - 1/k^2) = (1 - 1/2,632^2) = 0,8556$. Portanto, pelo menos 85% das mulheres têm salário de no máximo ± 2,632 desvios-padrão da média, ou seja, entre £6.800 (= 16.800 – 2,632 × 3.800) e £26.800 (= 16.800 + 2,632 × 3.800). No máximo 15% das mulheres situam-se fora dessa faixa, e se poderia supor que cerca de 7,5% das mulheres, no máximo, teriam salários acima de £26.800. Isso sugere que a distribuição de salários seja aproximadamente simétrica, o que pode não ser verdade.

A desigualdade de Chebyshev é uma regra muito conservadora, pois se aplica a qualquer distribuição; se soubermos mais a respeito da forma de uma distribuição específica (por exemplo, que a estatura de homens tem distribuição normal – veja o capítulo 3), então poderemos fazer uma afirmação mais precisa. No caso da distribuição normal, mais de 99% dos homens estão a menos de 2,632 desvios-padrão da estatura média, porque há uma concentração de observações perto do centro da distribuição.

Também podemos utilizar a desigualdade de Chebyshev para investigar o intervalo entre quartis. A fórmula 1.27 indica que pelo menos 50% das observações estão a menos de $\sqrt{2} = 1,41$ desvios-padrão da média, um valor mais conservador do que o 1,3 utilizado anteriormente.

Exercício 1.4

a) Com os dados do exercício 1.2, calcule o intervalo entre quartis, a variância e o desvio-padrão.
b) Calcule o coeficiente de variação.

c) Verifique se a relação entre o IQ e o desvio-padrão, enunciada no texto, é aproximadamente correta para essa distribuição.
d) Aproximadamente quanto da distribuição está situado dentro de um desvio-padrão em qualquer lado da média? Como isso se compara com a predição proveniente da desigualdade de Chebyshev?

Mensuração de assimetria

A **assimetria** de uma distribuição é a terceira característica mencionada anteriormente, além da posição e da dispersão. A distribuição de riqueza apresenta forte assimetria à direita, ou seja, ela tem assimetria **positiva**; sua cauda longa está no lado direito da distribuição. Uma medida de assimetria fornece uma indicação numérica de quão assimétrica é a distribuição.

Uma medida de assimetria, chamada **coeficiente de assimetria**, é:

$$(1.28)\quad \frac{\sum f(x-\mu)^3}{N\sigma^3}$$

que se baseia em desvios em relação à média *elevados ao cubo*. O resultado da aplicação da fórmula 1.28 é positivo quando a distribuição tem assimetria à direita (como no caso da distribuição de riqueza), zero no caso de uma distribuição simétrica, e negativo quando a distribuição tem assimetria à esquerda. A tabela 1.11 mostra a realização do cálculo para os dados de riqueza (algumas linhas foram omitidas para economizar espaço). Portanto:

$$\frac{\sum f(x-\mu)^3}{N} = \frac{1.382.901.383.966}{16.933} = 81.669.012,2$$

Dividindo-se por σ^3, tem-se:

$$\frac{81.669.012,2}{12.160.125} = 6,716$$

que é positivo, como se esperava.

TABELA 1.11 Cálculo da assimetria dos dados de riqueza

Faixa	Ponto médio (x mil libras)	Freqüência, f	Desvio, $x-\mu$	$(x-\mu)^3$	$f(x-\mu)^3$
0	5,0	3.417	−126,4	−2.021.544	−6.907.616.944
10.000	17,5	1.303	−113,9	−1.479.319	−1.927.552.150
⋮	⋮	⋮	⋮	⋮	⋮
1.000.000	1.500,0	125	1.368,6	2.563.237.059	320.404.632.317
2.000.000	3.000,0	41	2.868,6	23.604.266.037	967.774.907.513
Total		16.933	4.674,8	26.419.423.675	1.382.901.383.966

A medida de assimetria é muito menos útil no trabalho prático do que as medidas de posição e dispersão, e o conhecimento do valor do coeficiente nem sempre dá uma idéia detalhada da forma da distribuição: duas distribuições muito diferentes podem ter o mesmo coeficiente. Na realização de trabalhos descritivos, talvez seja melhor desenhar o histograma da distribuição.

Comparação da distribuição de riqueza em 2001 e em 1979

Podemos tirar algumas lições úteis da comparação entre a distribuição de riqueza em 2001 e a distribuição correspondente em 1979, o que engloba o período do governo conservador (iniciado com Margareth Thatcher em 1979) até os quatro primeiros anos da administração trabalhista. Essa análise mostra quão útil podem ser as várias estatísticas sintéticas ao se comparar duas distribuições diferentes. Os dados de riqueza para 1979 são fornecidos no problema 1.5, no qual se pede que o leitor confirme os seguintes cálculos.

A riqueza média em 1979 era igual a £16.399, cerca de 1/8 de seu valor em 2001. A média aumentou substancialmente (cerca de 10% ao ano, em média), mas parte desse crescimento foi inflação, e não aumento real de ativos possuídos. Na verdade, entre 1979 e 2001 o índice de preços no varejo subiu de 59,9 para 183,2, ou seja, cresceu aproximadamente três vezes. Assim, o aumento nominal[6] (isto é, em termos monetários, antes de qualquer ajuste em função da elevação dos preços) da riqueza é composto por duas partes: (1) um componente devido à inflação, que aproximadamente "triplicou" a riqueza; e (2) um componente real, formado por um aumento de 2,6 vezes (portanto, $3 \times 2{,}6 = 8$). Os índices de preços são discutidos no capítulo 10, quando mostra-se como dividir um aumento nominal em componentes de variação de preço e real (variação de quantidade). É provável que a dimensão do aumento real de riqueza esteja sendo superestimada nesse caso, em virtude do uso do índice de preços no varejo em lugar de um índice de preços de ativos. Uma parte substancial do crescimento dos valores de ativos no período talvez se deva ao aumento muito rápido nos preços de residências (as quais constituem uma parcela significativa do patrimônio de muitas famílias).

O desvio-padrão é afetado de maneira semelhante pela inflação. O valor de 1979 é 25.552, contra 229.957 em 2001, ou seja, cerca de nove vezes maior. Portanto, a dispersão da distribuição parece ter aumentado (mesmo que levemos em conta o efeito dos preços em geral). Examinando o coeficiente de variação, porém, vê-se que ele subiu de 1,56 a 1,75, ou seja, um crescimento pouco acentuado. A dispersão da distribuição, *relativamente à sua média*, não se alterou muito. Isso é confirmado pelo cálculo do desvio-padrão do logaritmo: em 1979, o valor era 1,31, ligeiramente inferior ao de 2001 (1,48).

A medida de assimetria para os dados de 1979 é 5,723, inferior ao valor de 2001 (6,716). Isso indica que a distribuição de 1979 é menos assimétrica que a de 2001. No-

6. Esse é um sentido diferente do termo "nominal" utilizado antes para indicar dados medidos numa escala nominal, ou seja, dados agrupados em categorias sem a existência de uma ordem clara. Infelizmente, os dois sentidos da palavra são empregados em estatística, muito embora normalmente, pelo contexto, fique evidente em que sentido ela está sendo empregada.

vamente, esses dois valores podem ser diretamente comparados, pois não dependem das unidades com as quais a riqueza era medida. Entretanto, a diferença relativamente pequena é de difícil interpretação em termos de alteração da forma da distribuição.

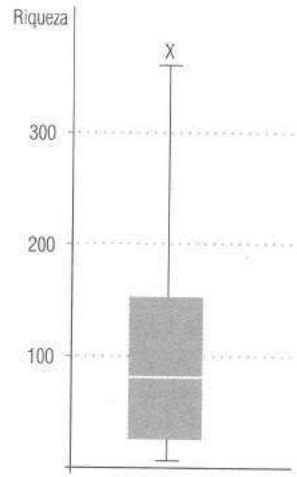
Box plot

Após o cálculo dessas diversas estatísticas de síntese, podemos retornar a um método gráfico útil de apresentação. É o ***box plot***, que mostra a mediana, os quartis e outros aspectos de uma distribuição num único diagrama. A figura 1.15 apresenta o *box plot* dos dados de riqueza.

A riqueza é medida no eixo vertical. A caixa retangular se estende verticalmente do primeiro ao terceiro quartil e, portanto, abrange a metade interna da distribuição. A linha horizontal que atravessa a caixa, um pouco abaixo da metade da caixa, representa a mediana. Isso nos diz que há certo grau de assimetria, mesmo dentro da metade central da distribuição, embora não pareça ser muito acentuada. Os dois "bigodes" se estendem acima e abaixo da caixa, correspondendo às observações máxima e mínima, *excluindo os valores extremos*. Um valor extremo é definido como qualquer observação cujo valor é mais de 1,5 vez o intervalo entre quartis (igual à altura da caixa), acima ou abaixo da caixa. Anteriormente, descobrimos que o valor de IQ era igual a 131.974, e o do quartil superior, 151.370, de modo que um valor extremo superior ficará acima de 151.370 + 1,5 × 131.974 = 349.331. Não há valores extremos abaixo da caixa, pois a riqueza não pode ser negativa. O "bigode" superior, portanto, é substancialmente mais longo do que o inferior e indica o grau de dispersão na direção das caudas da distribuição. A cruz indica os valores extremos e, na verdade, estende-se bem além do que é mostrado no diagrama.

FIGURA 1.15

***Box plot* da distribuição de riqueza**



Um diagrama simples, portanto, revela muita informação a respeito da distribuição. Outros *box plots* poderiam ser colocados lado a lado no mesmo diagrama (talvez representando outros países), o que tornaria as comparações bastante simples. Alguns pacotes estatísticos, como SPSS e Stata, podem gerar *box plots* a partir dos dados originais, sem que seja necessário o usuário calcular a mediana, etc. Entretanto, as planilhas eletrônicas ainda não dispõem desse recurso tão útil.

Dados em série temporal: gastos de investimento, 1970-2002

Os dados de distribuição de riqueza nos oferecem um retrato da situação em momentos específicos, e é possível comparar os retratos de 1979 e 2001. Com freqüência, porém, desejamos concentrar nossa atenção na trajetória de uma variável e, portanto, usamos **dados em série temporal**. As técnicas de apresentação e síntese são ligeiramente diferentes das empregadas para dados em *cross section*. Como exemplo, usamos dados de investimento no Reino Unido para o período de 1970-2002. Esses dados foram extraídos da Statbase (http://www.statistics.gov.uk/statbase/), embora você possa encontrá-los em *Economic trends annual supplement*. Para a economia, os gastos de investimento são importantes, pois constituem um dos principais fatores de seu crescimento. Até pouco tempo, o crescimento da economia do Reino Unido, segundo os padrões internacionais, foi limitado, e uma das causas pode ter sido a falta de investimento. A variável estudada é a formação doméstica bruta (isto é, antes de subtrair a depreciação) total de capital fixo, medida em milhões de libras. Os dados são apresentados na tabela 1.12.

TABELA 1.12 Investimento no Reino Unido, 1970-2002

Ano	Investimento	Ano	Investimento	Ano	Investimento
1970	10.036	1981	43.331	1992	100.583
1971	11.243	1982	47.394	1993	101.027
1972	12.347	1983	51.490	1994	108.314
1973	15.227	1984	58.589	1995	117.448
1974	18.134	1985	64.400	1996	126.291
1975	21.856	1986	68.546	1997	133.776
1976	25.516	1987	78.996	1998	150.540
1977	28.201	1988	96.243	1999	154.647
1978	32.208	1989	111.324	2000	161.210
1979	38.211	1990	114.300	2001	166.691
1980	43.238	1991	105.179	2002	169.972

Nota: Dados em série temporal consistem em observações de uma ou mais variáveis em diversos períodos. As observações podem ser diárias, semanais, mensais, trimestrais ou anuais, como neste caso.

É preciso lembrar que os dados são medidos em preços correntes, de modo que os valores refletem aumentos de preço, bem como variações do volume de investimento físico. A série na tabela 1.12, portanto, indica o volume efetivo de dinheiro investido em cada ano. As técnicas empregadas a seguir para sintetizar os dados de investimento poderiam ser igualmente aplicadas a uma série que apresentasse o volume de investimento.

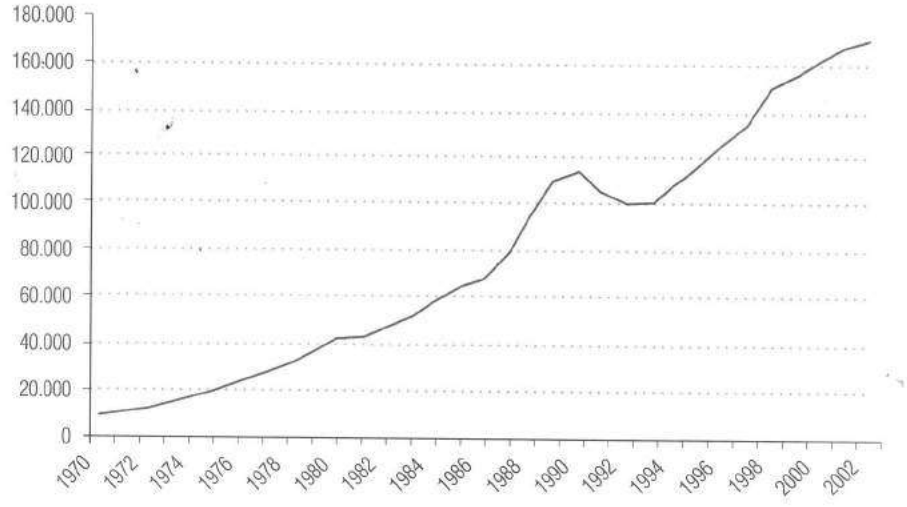
Inicialmente, podemos usar técnicas gráficas para ter uma noção das características do investimento. A figura 1.16 apresenta um gráfico dessa série de investimento – ele indica os períodos no eixo horizontal e a variável investimento no eixo vertical.

Uma representação gráfica de dados desse tipo destaca algumas características fundamentais da série:

- A **tendência** do investimento é crescente, com apenas alguns anos em que não houve aumento, ou houve redução.
- Existe uma "corcova" nos dados no final da década de 1980 e no início da década de 90, antes de a série voltar a sua tendência. Algo de incomum deve ter ocorrido nessa época. Se desejarmos saber que fatores determinam o investimento (ou o efeito do investimento sobre outras variáveis econômicas), será preciso obter alguma informação relevante desse período.
- A tendência é ligeiramente **não-linear** – comporta-se como uma curva cada vez mais inclinada com o tempo. Isso ocorre essencialmente porque o investimento cresce de acordo com uma *porcentagem* ou um *valor proporcional* a cada ano. Co-

FIGURA 1.16

Gráfico da série de investimento no Reino Unido, 1970-2002



Nota: As coordenadas são os valores {ano, investimento}; o primeiro ponto tem as coordenadas {1970, 10.036}, por exemplo.

mo veremos logo a seguir, ele cresce cerca de 9% ao ano. Portanto, à medida que o nível de investimento cresce anualmente, o mesmo se dá com o aumento do nível, o que gera um gráfico não-linear.

- Valores sucessivos da variável investimento são de magnitude semelhante, isto é, o valor no ano t é semelhante ao valor em t – 1. O investimento, por exemplo, não varia de £40 bilhões num ano para £10 bilhões no ano seguinte, e depois sobe para £50 bilhões. Na verdade, o valor num ano parece basear-se no valor do ano anterior, mais (em geral) cerca de 9%. Esse fenômeno, chamado de **correlação serial**, é um dos aspectos dos dados que poderíamos nos interessar em investigar. A *ordem* dos dados é relevante, ao contrário do que ocorria com os dados em *cross section*. Quando se decide modelar o comportamento do investimento, é preciso concentrar a atenção nas *variações* do investimento de um ano para outro.
- A série parece ter um comportamento "mais suave" nos primeiros anos (até 1986, talvez), apresentando maior volatilidade mais tarde. Em outras palavras, há maiores flutuações em torno da tendência nos anos finais. Poderíamos expressar esse fato mais formalmente dizendo que a variância do investimento em torno de sua tendência parece crescer (aumentar) com o tempo. Isso é chamado de **heteroscedasticidade**; uma variância constante é chamada de **homoscedasticidade**.

Podemos obter uma visão adicional sobre como o investimento evoluiu com o passar do tempo concentrando nossa atenção na variação do investimento de um ano para outro. Sendo o investimento no ano t representado por I_t, então a variação do investimento, ΔI_t, é dada por $I_t - I_{t-1}$.

A tabela 1.13 apresenta a variação do investimento ano a ano, e a figura 1.17 mostra um gráfico da série.

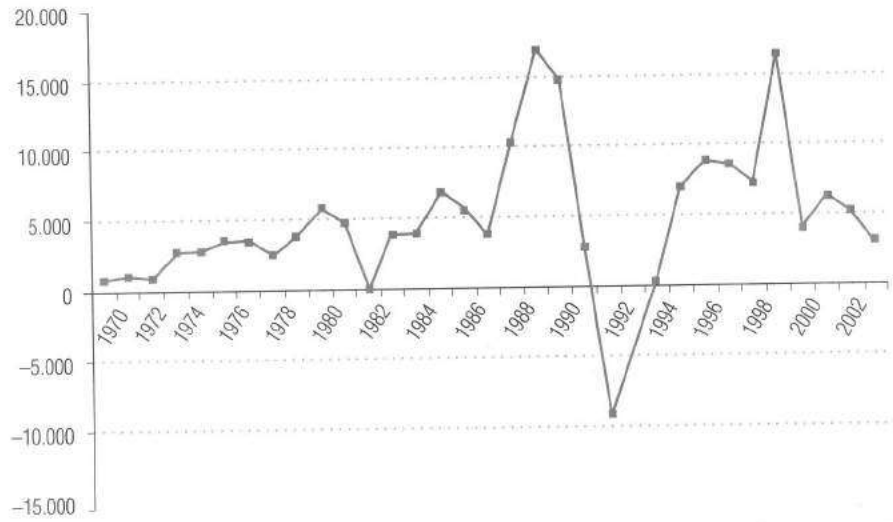
TABELA 1.13 Variação do investimento

Ano	Δ Investimento	Ano	Δ Investimento	Ano	Δ Investimento
1970	970	1981	93	1992	–4.596
1971	1.207	1982	4.063	1993	444
1972	1.104	1983	4.096	1994	7.287
1973	2.880	1984	7.099	1995	9.134
1974	2.907	1985	5.811	1996	8.843
1975	3.722	1986	4.146	1997	7.485
1976	3.660	1987	10.450	1998	16.764
1977	2.685	1988	17.247	1999	4.107
1978	4.007	1989	15.081	2000	6.563
1979	6.003	1990	2.976	2001	5.481
1980	5.027	1991	–9.121	2002	3.281

Nota: A variação do investimento é obtida calculando-se a diferença entre observações sucessivas. Por exemplo, 1.207 é a diferença entre 10.036 e 11.243.

FIGURA 1.17

Gráfico da série da variação do investimento



A série é composta principalmente por valores positivos, indicando que o investimento cresce com o passar do tempo. Também mostra que o crescimento aumenta a cada ano, talvez com maior volatilidade (do aumento) no final do período. O gráfico ainda indica de modo evidente a variação que ocorreu em torno de 1990.

Outra maneira útil de examinar os dados é utilizar o **logaritmo** do investimento. Essa transformação resulta na linearização da série não-linear de investimento. A tabela 1.14 apresenta os valores transformados e a figura 1.18 mostra um gráfico da série resultante. Nesse caso, foi usado o logaritmo natural (base e).

Essa nova série é muito mais suave do que a original (como geralmente acontece quando se calculam logaritmos) e nos ajuda a mostrar a tendência a longo prazo, em-

Valores extremos

A elaboração de gráficos de dados também nos permite detectar a presença de **valores extremos** (observações incomuns). Estes podem resultar de um erro de inserção de dados (por exemplo, digitar 97 em lugar de 970) ou porque algo de especial aconteceu (por exemplo, o investimento em 1991). Qualquer que seja o motivo, sua presença deve ser evidenciada por um gráfico apropriado. Por exemplo, o gráfico da variação do investimento realça o dado de 1991. No caso de um simples erro, você deve evidentemente corrigi-lo. Se achar que o valor extremo não é um simples erro de digitação, pense nos possíveis motivos da existência desse valor e considere se ele distorce a descrição que está tentando fazer.

TABELA 1.14 Logaritmo do investimento e variação do logaritmo

Ano	ln Investimento	Δ ln Investimento	Ano	ln Investimento	Δ ln Investimento	Ano	ln Investimento	Δ ln Investimento
1970	9,214	0,102	1981	10,677	0,002	1992	11,519	–0,045
1971	9,328	0,114	1982	10,766	0,090	1993	11,523	0,004
1972	9,421	0,094	1983	10,849	0,083	1994	11,593	0,070
1973	9,631	0,210	1984	10,978	0,129	1995	11,674	0,081
1974	9,806	0,175	1985	11,073	0,095	1996	11,746	0,073
1975	9,992	0,187	1986	11,135	0,062	1997	11,804	0,058
1976	10,147	0,155	1987	11,277	0,142	1998	11,922	0,118
1977	10,247	0,100	1988	11,475	0,197	1999	11,949	0,027
1978	10,380	0,133	1989	11,620	0,146	2000	11,990	0,042
1979	10,551	0,171	1990	11,647	0,026	2001	12,024	0,033
1980	10,674	0,124	1991	11,563	–0,083	2002	12,043	0,019

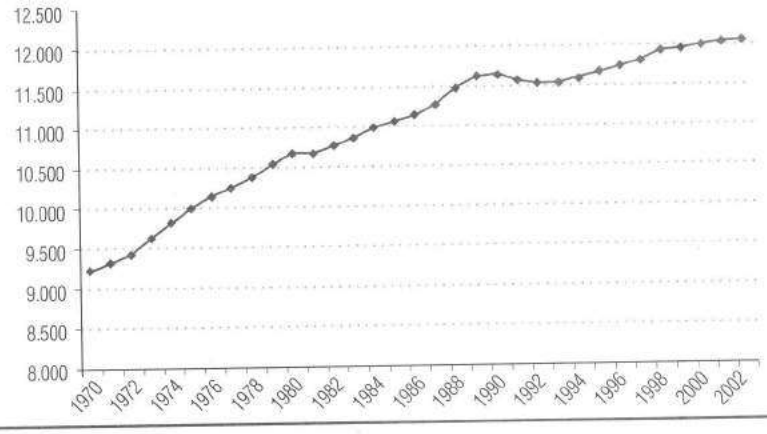
Nota: Para o ano de 1970, 9,214 é o logaritmo natural de 10.036, isto é, ln 10.036 = 9,214.

bora oculte parte da volatilidade do investimento. A curva do gráfico é uma boa aproximação da taxa média de crescimento do investimento no período (em representação decimal). Isso é calculado da seguinte maneira:

$$(1.29)\quad \text{inclinação} = \frac{\text{variação de (ln) investimento}}{\text{número de anos}} = \frac{12,043 - 9,214}{32} = 0,088$$

FIGURA 1.18

Gráfico da série temporal do logaritmo das despesas de investimento



ou seja, 8,8% ao ano. Note que, embora existam 33 observações, só há 32 anos de crescimento. Uma advertência: você deve utilizar logaritmos naturais (base e), e não logaritmos em base 10, para que esse cálculo funcione. Lembre-se também de que o crescimento do *volume* de investimento será inferior a 8,8% ao ano porque parte do crescimento deve-se a aumento de preços.

A apresentação logarítmica é útil quando se deseja comparar duas séries diferentes: em sua representação gráfica, com seus logaritmos, é fácil perceber qual delas está crescendo mais depressa – basta ver a que tem maior inclinação.

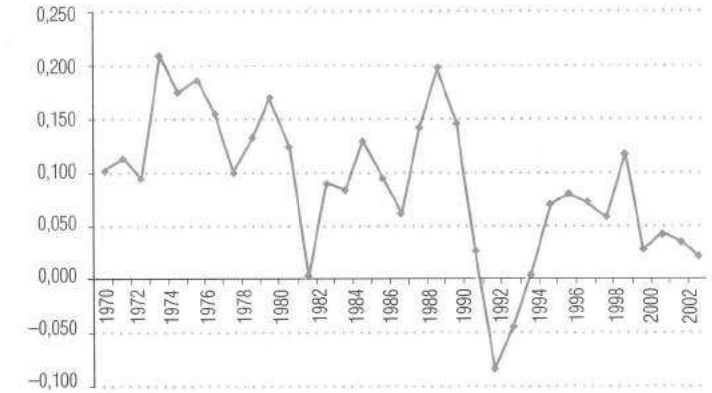
Um corolário da equação 1.29 é o fato de a variação do logaritmo natural do investimento de um ano para outro representar a variação *percentual* dos dados. Por exemplo, o logaritmo natural do investimento em 1970 é 9,214, ao passo que o de 1971 é 9,328. A diferença é igual a 0,114, o que significa que a taxa de crescimento foi 11,4%. Você pode verificar esse resultado usando os dados originais. Essa aproximação é satisfatória desde que a taxa de crescimento não tenha sido superior a 20%.

Finalmente, podemos fazer o gráfico da diferença dos logaritmos, assim como fizemos para a diferença do nível da série. Isso é mostrado na figura 1.19 (os cálculos estão na tabela 1.14).

Esse gráfico é bastante revelador. Ele mostra que a série flutua em torno de aproximadamente 0,10 (a média calculada na equação 1.29), com uma ligeira tendência declinante. Além disso, a série não parece apresentar volatilidade crescente com o tempo, como ocorreu com as outras. O gráfico, portanto, demonstra que, em termos *proporcionais*, não há aumento de volatilidade; a variância da série em torno de 0,10 não muda muito com o tempo (embora a observação de 1991 ainda pareça "incomum"). Poderíamos nos referir a essa série, de maneira não muito rigorosa, como **estacionária** – tem média e variância constantes. Essa não é uma definição formal do termo estatístico, que impõe condições mais rigorosas à série, mas é útil como referência.

FIGURA 1.19

Gráfico da série temporal da diferença da série logarítmica



Elaboração de gráficos de séries múltiplas

O investimento é composto por várias categorias, e a tabela fornecida no problema 1.14 apresenta os dados de investimento em diversas rubricas: residências; outros prédios e obras; transporte; máquinas; e ativos fixos intangíveis. Em conjunto, elas formam o investimento total. Geralmente é útil mostrar todas as séries juntas num único gráfico. A figura 1.20 apresenta um **gráfico de séries múltiplas** com os dados de investimento.

A elaboração desse tipo de gráfico é simples; não passa de uma extensão da técnica de apresentação de uma única série. Pode ser elaborado à mão com facilidade, embora trabalhosamente, mas a maioria dos aplicativos de computador é capaz de produzir esse tipo de gráfico com rapidez. A única complicação surge quando as séries têm ordens de magnitude diferentes e se torna difícil visualizar todas elas no gráfico. Nesse caso, você pode medir algumas séries numa segunda escala vertical, no eixo à direita. Um exemplo é fornecido na figura 1.21, que mostra a série de investimento em relação à taxa de juros, que tem valores numéricos muito menores.

Voltando às categorias de investimento, a figura 1.20 indica que "máquinas" e "outros prédios" são as duas categorias principais e de tamanhos aproximadamente iguais. Isso não tem mudado muito com o passar do tempo. Todas as séries exibem padrões semelhantes de comportamento no período. (Seria possível produzir também gráficos múltiplos do logaritmo do investimento e de sua variação.)

FIGURA 1.20

Gráfico de séries temporais múltiplas de investimento

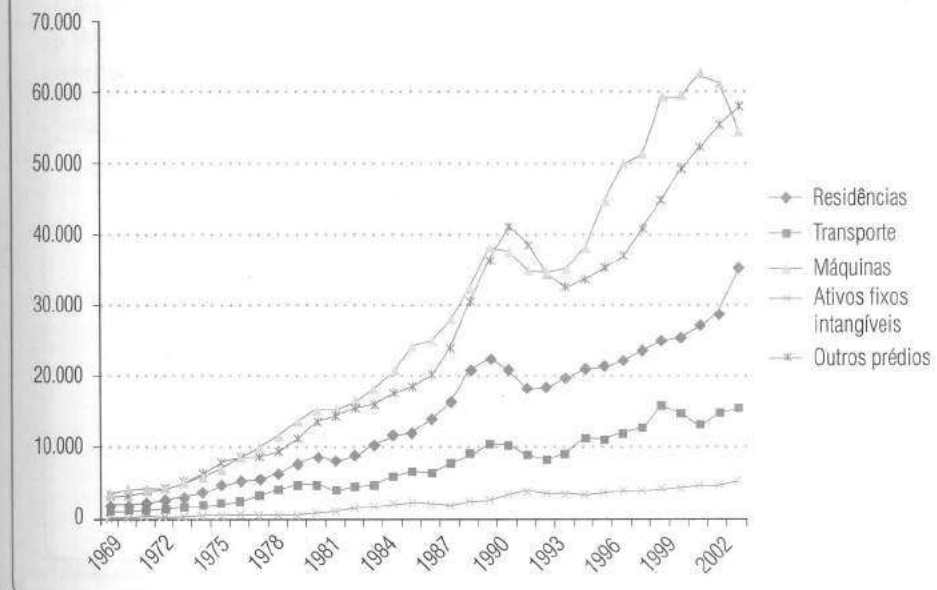
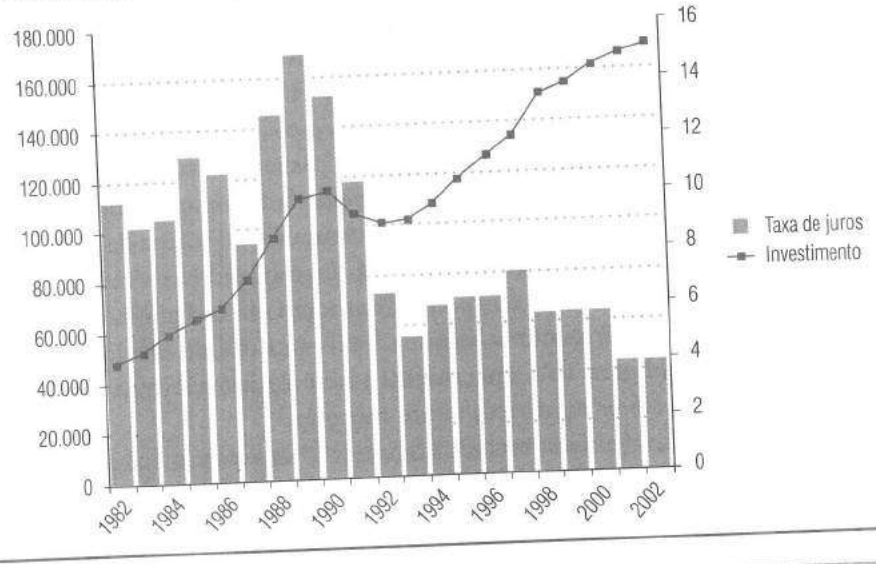


FIGURA 1.21

Gráfico de séries temporais com o uso de duas escalas verticais: investimento (escala à esquerda) e taxa de juros (escala à direita), 1989-2002



Sobreposição de faixas das séries de dados

O gráfico abaixo, extraído de *Treasury briefing*, de fevereiro de 1994, apresenta um belo exemplo de como se deve fazer um gráfico de séries múltiplas e compará-las. O objetivo é comparar as recessões e recuperações de 1974-78, 1979-83 e 1990-93. Em vez de pôr o tempo no eixo horizontal, colocou-se o número de trimestres desde o início de cada recessão, fazendo com que as séries se sobrepusessem. Isso facilita a visualização da profundidade da última recessão e do longo período até o início da recuperação. Em contraposição, a recessão de 1974-78 encerrou-se muito rapidamente e a recuperação foi bastante veloz.

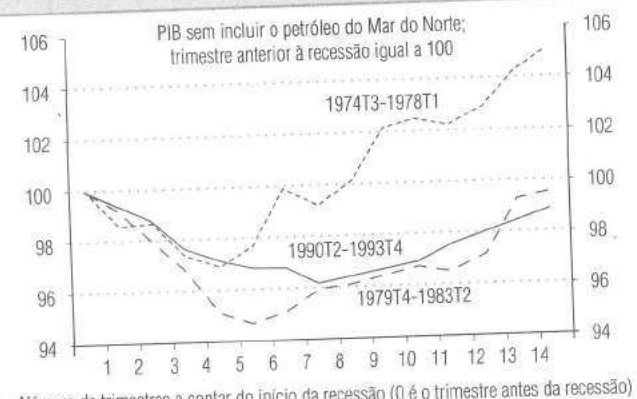
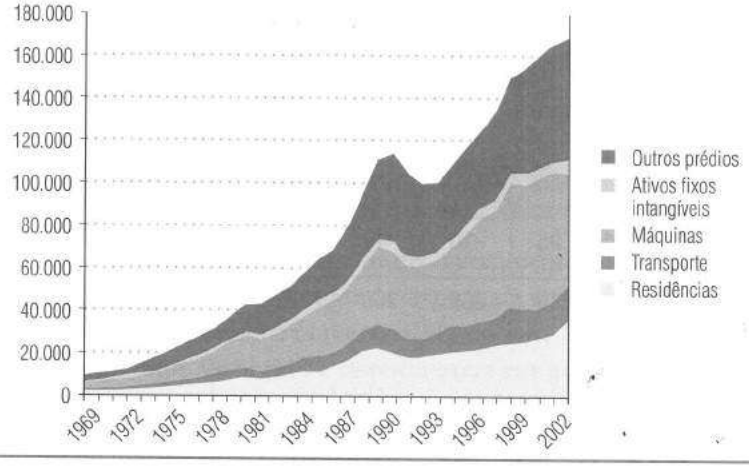


FIGURA 1.22

Gráfico de área de categorias de investimento, 1969-2002



Essas séries também podem ser ilustradas por meio de um **gráfico de áreas**, que apresenta as séries empilhadas, como na figura 1.22.

"Lixo gráfico"

Com a disponibilidade de aplicativos modernos, é fácil ir longe demais e produzir um gráfico que mais esconde do que revela. Há forte tentação para adicionar efeitos em três dimensões, realçá-lo um pouco com o uso de cores, girar e inclinar a perspectiva, etc. Chama-se esse tipo de produção de "lixo gráfico". Como exemplo, veja a figura 1.23, uma representação alternativa do gráfico de áreas da figura 1.22. Foi divertido criá-la, mas não transmite o que precisa de maneira alguma! Gosto é uma questão pessoal, evidentemente, mas um pouco de moderação é essencial.

FIGURA 1.23

Gráfico exagerado do investimento

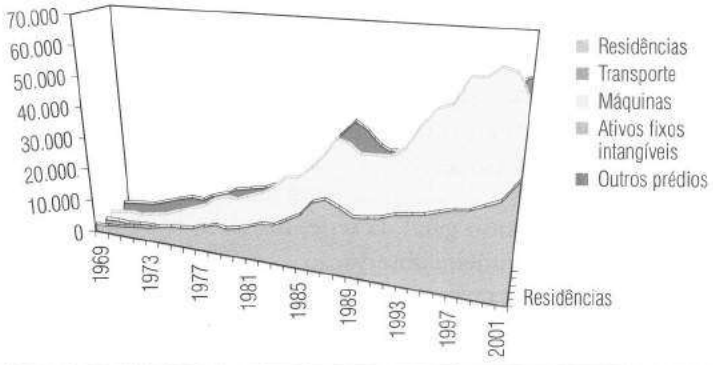


TABELA 1.15 Cálculo da média geométrica – fatores de crescimento anual

Ano	Investimento	Fatores de crescimento
1970	10.036	
1971	11.243	1,1203 (=11.243/10.036)
1972	12.347	1,0982 (=12.347/11.243)
1973	15.227	1,2333 etc.
⋮	⋮	⋮
1997	133.776	1,0593
1998	150.540	1,1253
1999	154.647	1,0273
2000	161.210	1,0424
2001	166.691	1,0340
2002	169.972	1,0197

Nota: Cada fator de crescimento indica simplesmente o quociente entre o investimento daquele ano em relação ao investimento do ano anterior.

Média geométrica

Ao determinar a taxa média de crescimento do investimento, calculamos implicitamente a **média geométrica** de uma série. Numa série de n valores, sua média geométrica é calculada pela raiz n do produto dos valores, ou seja:[7]

$$(1.30)\quad \text{média geométrica} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Os valores de x, nesse caso, são os fatores de crescimento de cada ano, como vemos na tabela 1.15 (os valores dos anos intermediários foram omitidos).

O produto dos 32 fatores de crescimento é igual a 16,936 (o mesmo obtido dividindo a observação final pela observação inicial – por quê?) e a 32ª raiz desse número é 1,092. Esse último valor, 1,092, é a média geométrica dos fatores de crescimento, e dela podemos obter a taxa de crescimento de 9,2% a.a. subtraindo 1.

Sempre que alguém está lidando com dados de crescimento (ou qualquer série baseada num processo multiplicativo), deve usar a média geométrica, e não a média aritmética, para obter a resposta. Entretanto, o uso da média aritmética nesse caso geralmente produz apenas um erro pequeno, como é mostrado a seguir.

7. O símbolo $\prod$ (letra grega maiúscula "pi") significa o "produto de", assim como Σ significa a sua soma.

Uma maneira aproximada de calcular a taxa média de crescimento

Vimos que, ao calcular taxas de crescimento, deve-se usar a média geométrica, mas, se a taxa de crescimento for pequena, a média aritmética fornecerá aproximadamente a resposta correta. A média aritmética dos fatores de crescimento é:

$$\frac{1{,}1203 + 1{,}0982 + \ldots + 1{,}0340 + 1{,}0197}{32} = 1{,}095$$

o que nos dá uma estimativa de 9,5% a.a. para a taxa de crescimento – próxima do valor correto. Observe também que se poderia, de maneira equivalente, tirar a média das taxas anuais de crescimento (0,1203; 0,0982, etc.), obtendo 0,095, ou seja, o mesmo resultado. O uso da média aritmética se justifica nesse contexto, pois se necessita apenas de uma aproximação da resposta correta, e as taxas anuais de crescimento são razoavelmente pequenas.

Juros compostos

Os cálculos que fizemos em relação a taxas de crescimento são análogos ao cálculo de **juros compostos**. Se investirmos £100 a uma taxa anual de juros de 10%, então o investimento crescerá a 10% a.a. (supondo que todos os juros sejam reaplicados). Portanto, após um ano, o total terá aumentado para £100 × 1,1 (£110); após dois anos, para £100 × $1{,}1^2$ (£121); e após t anos, para £100 × $1{,}1^t$. A fórmula geral do valor final S_t de uma quantia S_0 aplicada por t anos a uma taxa de juros r é:

$$(1.31)\quad S_t = S_0(1+r)^t$$

em que r é representado por um número decimal. Reorganizando 1.31 para isolar o valor de r, temos:

$$(1.32)\quad r = \sqrt[t]{S_t / S_0} - 1$$

que é exatamente a fórmula da taxa média de crescimento. Como exemplo adicional: suponhamos que um fundo de investimento transforme um depósito inicial de £8.000 em £13.500 no prazo de 12 anos. Qual é a taxa média de retorno do investimento? Sendo $S_0 = 8$, $S_t = 13{,}5$, $t = 12$ e usando 1.32, obtemos:

$$r = \sqrt[12]{13{,}5/8} - 1 = 0{,}045$$

ou seja, 4,5% ao ano.

A fórmula 1.32 também pode ser empregada para calcular a **taxa de depreciação** e o valor da depreciação anual dos ativos de uma empresa. Nesse caso, S_0 é o valor inicial do ativo, S_t representa o valor final ou valor de liquidação, e a taxa anual de depreciação é dada por r na equação 1.32.

Variância de uma série temporal

Como devemos descrever a variância de uma série temporal? A variância dos dados de investimento pode ser calculada, mas seria tão pouco informativa quanto a média. Como a série tem tendência, e isso deve continuar num prazo mais longo, a variância, em princípio, é infinita. A variância calculada estaria intimamente relacionada ao tamanho da amostra: quanto maior for a amostra, maior será a variância. Mais uma vez, faz mais sentido calcular a variância da taxa de crescimento, cuja tendência de longo prazo é muito mais fraca.

Essa variância pode ser calculada com a seguinte fórmula:

$$(1.33)\quad s^2 = \frac{\sum(x-\overline{x})^2}{n-1} = \frac{\sum x^2 - n\overline{x}^2}{n-1}$$

em que $\overline{x}$ é a taxa média (aritmética) de crescimento. O cálculo é executado na tabela 1.16, usando a fórmula da direita na equação 1.33.

A variância, portanto, é igual a:

$$s^2 = \frac{0,4635 - 32 \times 0,092^2}{31} = 0,0056$$

e o desvio-padrão é 0,075, a raiz quadrada da variância. O coeficiente de variação é igual a:

$$CV = \frac{0,075}{0,092} = 0,815$$

ou seja, o desvio-padrão da taxa de crescimento é aproximadamente 80% da média.

TABELA 1.16 Cálculo da variância da taxa de crescimento

Ano	Investimento	Taxa de crescimento	
		x	x^2
1971	11.243	0,1203	0,0145
1972	12.347	0,0982	0,0096
1973	15.227	0,2333	0,0544
1997	⋮	⋮	⋮
1998	150.540	0,1253	0,0157
1999	154.647	0,0273	0,0007
2000	161.210	0,0424	0,0018
2001	166.691	0,0340	0,0012
2002	169.972	0,0197	0,0004
Total		3,0392	0,4635

Três coisas devem ser notadas nesse cálculo: em primeiro lugar, usamos a média aritmética (optar pela média geométrica faz muito pouca diferença); em segundo lugar, aplicamos a fórmula da variância da amostra, pois o período de 1970-2002 é uma amostra de todos os dados que poderíamos coletar; em terceiro lugar, também poderíamos ter utilizado os fatores de crescimento no cálculo da variância (por quê?).

Exemplo resolvido 1.6

A partir dos seguintes dados:

Ano	1999	2000	2001	2002	2003
Preço de um *laptop*	1.100	900	800	750	700

podemos calcular a taxa média de crescimento anual do preço da seguinte maneira: o fator geral de crescimento é 700/1.100 = 0,6363. O fato de que esse número é inferior a 1 simplesmente reflete o fato de que o preço caiu com o passar do tempo. Caiu a 64% de seu valor original. Para encontrar a taxa anual, tiramos a raiz quarta de 0,6363 (quatro anos de crescimento). Portanto, obtemos $\sqrt[4]{0,6363} = 0,893$, isto é, a cada ano o preço cai a 89% de seu valor no ano anterior. Isso significa que o preço está caindo a 0,893 – 1 = –0,107, ou seja, cai aproximadamente 11% por ano.

Podemos ver se a queda é mais ou menos a mesma a cada ano calculando o fator de crescimento de cada ano. Os fatores de crescimento são:

Ano	1999	2000	2001	2002	2003
Preço de um *laptop*	1.100	900	800	750	700
Fator de crescimento		0,818	0,889	0,9375	0,933
Queda de preço		–19%	–11%	–6%	–7%

A queda de preço foi maior nos primeiros anos, tanto em termos percentuais quanto em termos absolutos. O cálculo do desvio-padrão dos valores na última linha fornece uma medida da variabilidade ano a ano. A variância é dada por:

$$s^2 = \frac{(19-11)^2 + (11-11)^2 + (6-11)^2 + (7-11)^2}{3} = 30,7$$

e o desvio-padrão, portanto, é igual a 5,54%. (Os cálculos foram arredondados, mas a resposta é precisa.)

Exercício 1.6

a) Usando os dados do exercício 1.5, calcule o nível médio de lucro no período e a taxa média de crescimento do lucro durante o período. Que informação parece ser mais útil?
b) Calcule a variância do lucro e compare-a à variância das vendas.

Gráficos de dados bivariados: diagrama de dispersão

A análise da série de investimento constitui um exemplo do uso de **métodos univariados**: há apenas uma variável envolvida. Entretanto, é comum examinarmos a relação entre duas (às vezes mais) variáveis, em que precisamos usar **métodos bivariados** (ou **multivariados**). Para ilustrar esses métodos, examinaremos a relação entre gastos de investimento e produto interno bruto (PIB). A teoria econômica nos diz que deve haver uma relação positiva entre essas variáveis, com PIB mais elevado associado a maior investimento. A tabela 1.17 expõe dados do PIB do Reino Unido.

Um **diagrama de dispersão** (também chamado de **gráfico XY**) apresenta uma das variáveis (investimento, nesse caso) no eixo vertical, a outra (PIB), no eixo horizontal e, portanto, mostra a relação entre elas. Por exemplo, isso permite ver se valores elevados de uma variável tendem a estar associados a valores altos da outra. A figura 1.24 apresenta a relação entre investimento e PIB.

O gráfico indica a existência de uma forte relação linear entre as duas variáveis, com exceção de uma queda curiosa no meio. Ela reflete a forte queda do investimento após 1990, que *não* é acompanhada por uma queda do PIB (se fosse, o gráfico XY indicaria uma relação linear sem a queda). É importante reconhecer a diferença entre o gráfico de séries temporais e o gráfico XY. Por causa da inflação, as observações mais recentes tendem a aparecer no canto superior direito do gráfico XY (pois tanto investimento quanto PIB estão crescendo no tempo), mas isso *não precisa* acontecer; se ambas as variáveis oscilassem para cima e para baixo, as observações mais recentes poderiam estar no canto inferior esquerdo (ou no centro, ou em qualquer outro lugar). Em contraposição, num gráfico de séries temporais, as observações recentes estão sempre mais à direita.

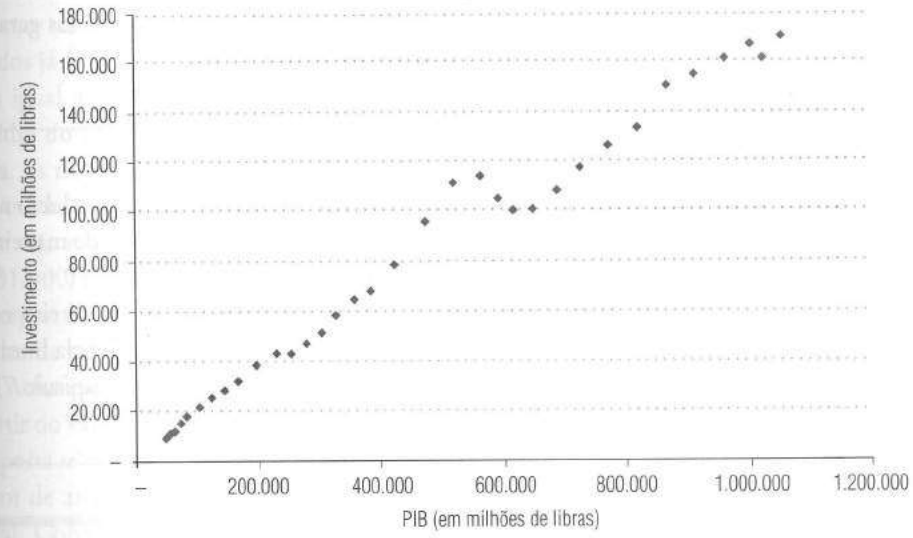
Note que as duas variáveis são medidas em termos nominais, ou seja, não são corrigidas pela inflação ocorrida no período. Isso pode ser percebido algebricamente: o gasto de investimento é composto pelo *volume* de investimento (I) vezes seu *preço* (P_I). De maneira análoga, o PIB nominal é igual ao PIB real (Y) vezes seu preço (P_Y). Portanto, o gráfico de fato representa a relação entre $P_I \times I$ e $P_Y \times Y$. É bem provável que os dois preços tenham tendência parecida e que isso domine os movimentos do investimento real e do PIB real. O gráfico, portanto, mostra uma mistura de preços e quantidades quando a relação mais interessante é entre as *quantidades* de investimento e produto.

TABELA 1.17 Dados do PIB do Reino Unido

Ano	PIB	Ano	PIB	Ano	PIB
1970	51.515	1981	253.000	1992	610.854
1971	57.449	1982	277.090	1993	642.327
1972	64.317	1983	302.774	1994	681.327
1973	73.979	1984	324.407	1995	719.176
1974	83.742	1985	354.952	1996	763.290
1975	105.773	1986	381.317	1997	810.944
1976	125.098	1987	419.631	1998	859.436
1977	145.528	1988	468.386	1999	903.865
1978	167.806	1989	514.168	2000	951.265
1979	197.355	1990	557.300	2001	994.037
1980	230.695	1991	586.149	2002	1.043.306

FIGURA 1.24

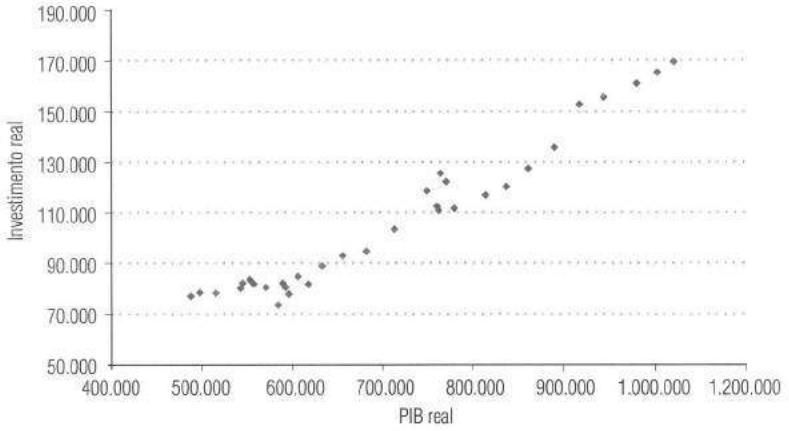
Diagrama de dispersão do investimento (eixo vertical) e do PIB (eixo horizontal) (valores nominais)



Nota: As coordenadas (x, y) de cada ponto são dadas pelos valores de investimento e PIB, respectivamente. Portanto, o primeiro ponto (1970) é marcado 10.036 unidades acima do eixo horizontal e a 51.515 unidades do eixo vertical.

FIGURA 1.25

Relação entre investimento real e PIB real



A figura 1.25 apresenta a relação entre as quantidades de investimento e PIB reais, isto é, após a remoção dos efeitos de preço. Esta figura não é tão clara quanto a do gráfico de valores nominais; agora há um "nó" de pontos no centro, onde talvez tanto o investimento real quanto o PIB real tenham oscilado. Está claro que algo "interessante" ocorreu por volta de 1990 e merece investigação adicional.

O capítulo 10, que trata de números índices, explica detalhadamente como extrair variáveis reais de variáveis nominais, como foi feito aqui, e descreve em termos gerais como corrigir os efeitos da inflação sobre magnitudes econômicas.

Exercício 1.7

a) Usando novamente os dados do exercício 1.5, faça um gráfico XY com o lucro no eixo vertical e as vendas no eixo horizontal. Escolha a escala dos eixos de maneira apropriada.
b) Se estiver usando Excel para produzir gráficos, clique com o botão direito no gráfico, selecione "adicionar linha de tendência" e escolha uma tendência linear. Isso gera a "linha de melhor ajuste" (abordada detalhadamente no capítulo 7). O que parece ser indicado pelo resultado?

Transformações de dados

Ao analisar os dados de emprego e investimento nos exemplos anteriores, com freqüência alteramos as variáveis de alguma forma para realçar características impor-

tantes. Em estatística, normalmente trabalha-se com dados transformados de alguma maneira, e não com os números originais. Vale a pena, portanto, resumir as principais transformações de dados disponíveis, apresentando justificativas para o seu emprego e examinando as implicações de tais ajustes dos dados originais. Discutiremos rapidamente as seguintes transformações:

- arredondamento;
- agrupamento;
- divisão ou multiplicação por uma constante;
- diferenças;
- cálculo de logaritmos;
- cálculo de recíprocos;
- deflacionamento.

Arredondamento

O arredondamento de dados aumenta a facilidade de leitura. Um detalhamento excessivo pode ocultar a mensagem a ser transmitida, de modo que o arredondamento da resposta faz com que seja mais fácil lembrá-la. Para dar um exemplo, vimos anteriormente neste capítulo que a riqueza média na Inglaterra é igual a £131.442,893 (usando três casas decimais). Porém, seria absurdo apresentá-la dessa forma, pois não sabemos com certeza se esse número é correto (na verdade, é quase certo que não é). Existe um grau espúrio de precisão que, ao contrário, pode iludir o leitor. Portanto, quanto devemos arredondar esse número para fins de apresentação? Lembre que os dados já foram de fato arredondados ao ser colocados em intervalos cuja amplitude era igual a 10.000 ou mais (todas as observações foram arredondadas pelo ponto médio do intervalo). Entretanto, a maior parte desse arredondamento se anula, ou seja, os números arredondados para cima compensam os arredondados para baixo, de modo que a média é razoavelmente precisa. Arredondar para £131.000 faz com que seja mais fácil memorizar o valor, e trata-se de uma alteração de apenas 0,3% (131.000/131.443 = 0,997), ou seja, é uma conciliação razoável. No texto anterior, a resposta não foi arredondada a esse ponto porque a finalidade era mostrar os métodos de cálculo.

A função de arredondamento é irreversível: não se pode obter o valor original a partir do valor transformado (arredondado). Portanto, você não deverá arredondar sua resposta se necessitar do valor original em cálculos subseqüentes. Além disso, pequenos erros de arredondamento podem se acumular, causando um erro grande na resposta final. Conseqüentemente, você *nunca* deve arredondar uma resposta intermediária. Mesmo que arredonde uma resposta intermediária por um valor pequeno, a resposta final pode resultar muito inexata. Tente o seguinte: calcule 60,29 × 30,37 – 1.831, antes e depois de arredondar os dois primeiros números para valores inteiros. No primeiro caso, você obtém 0,0073, e, no segundo, –31.

Agrupamento

Quando há excesso de dados para uma apresentação simples, o problema pode ser resolvido por agrupamento, embora o custo seja ocultar parte da informação. Usamos dados agrupados nos exemplos que relacionaram educação e emprego e os dados de riqueza. O uso de dados brutos nos teria dado informação excessiva, de modo que o agrupamento é uma primeira etapa da análise de dados. Do mesmo modo que o arredondamento, o agrupamento é outra transformação irreversível: uma vez realizada, não é possível recuperar a informação original.

Divisão/multiplicação por uma constante

Essa transformação é realizada para tornar a leitura dos números mais fácil ou para facilitar os cálculos com remoção de zeros à direita. Os dados de riqueza foram divididos por 1.000 para ajudar nos cálculos; se isso não tivesse sido feito, a coluna fx^2 conteria valores extremamente grandes. Algumas estatísticas sintéticas (por exemplo, a média) serão afetadas pela transformação, mas nem todas (como o coeficiente de variação). Procure recordar quais são afetadas! Os operadores E e V podem ajudar (veja o Apêndice 1B). Essa transformação é de fácil reversão.

Diferenças

Pode haver uma tendência no caso de dados de séries temporais, e por isso é melhor descrever as características dos dados em relação a ela. O resultado pode ter significado econômico mais útil; por exemplo, os governos geralmente se preocupam mais com o crescimento do produto do que com seu nível. O cálculo de diferenças, uma das maneiras de eliminar a tendência, foi utilizado com os dados de investimento por esses dois motivos. Uma das implicações desse cálculo é que a informação sobre o *nível* da variável é perdida e não pode ser recuperada.

Cálculo de logaritmos

O cálculo de logaritmos é empregado para linearizar uma série não-linear, em particular uma série que cresce a uma taxa constante. Geralmente, é mais fácil perceber as características fundamentais de uma série como essa com a elaboração de um gráfico do logaritmo em vez dos dados brutos. A transformação logarítmica também é útil em análise de regressão (veja o capítulo 9) porque gera estimativas de **elasticidade** (por exemplo, da demanda). O cálculo do logaritmo dos dados de investimento linearizou a série e tendeu a suavizá-la. Os inversos das transformações logarítmicas são 10^x (no caso de logaritmos comuns) e e^x (no caso de logaritmos naturais), o que permite recuperar os dados originais.

Cálculo de recíprocos

O recíproco de uma variável pode ter uma interpretação útil e proporcionar uma explicação mais intuitiva de um fenômeno. Essa transformação também converterá uma série linear numa série não-linear. O recíproco da rotatividade no mercado de

trabalho (ou seja, o número de pessoas que deixam de estar desempregadas dividido pelo número de pessoas desempregadas) nos dá uma idéia da duração do desemprego. Se metade dos desempregados encontra emprego a cada ano (rotatividade = 0,5), então a duração média do desemprego é igual a dois anos (= 1/0,5). Se um gráfico de rotatividade apresentar uma queda linear com o tempo, então a duração média do desemprego estará aumentando a uma taxa cada vez mais alta. A repetição da transformação em recíproco recupera os dados originais.

Deflacionamento

O deflacionamento converte uma série nominal numa série real, ou seja, numa série que reflete as variações de quantidades sem a contaminação causada por variações de preços. Isso é discutido em detalhes no capítulo 10. Geralmente, tem mais significado, em termos econômicos, tratar de uma variável real do que de uma variável nominal. Por exemplo, os consumidores se preocupam mais com sua renda real do que com sua renda monetária.

É perigoso confundir variáveis reais com variáveis nominais! Por exemplo, a renda nominal (monetária) de uma pessoa pode estar subindo e, no entanto, sua renda real estar caindo (caso os preços estejam crescendo mais depressa do que a renda monetária). É importante saber com que série você está lidando (esse é um erro comum entre os estudantes principiantes). Uma série de renda que está crescendo a 2-3% ao ano é provavelmente uma série real; uma série que está crescendo a 10% ao ano provavelmente é nominal.

Orientação para o estudante: como medir seu progresso

Você chegou ao final do capítulo, mas o trabalho ainda não terminou! É pouco provável que você tenha entendido tudo após uma leitura de ponta a ponta. Por isso, agora você deve fazer o seguinte:

- Examinar novamente os objetivos de estudo enunciados no início do capítulo. Você acha que os atingiu? Por exemplo, é capaz de reconhecer e enumerar os vários tipos distintos de dados (primeiro objetivo de estudo)?
- Ler o resumo do capítulo, a seguir, que vai ajudá-lo a encaixar tudo em um contexto. Você precisa reconhecer cada tópico e estar ciente das principais questões, técnicas, etc. dentro do tópico. Não deve haver surpresas ou lacunas!
- Ler a lista de termos-chave. Você deve ser capaz de dar uma definição ou descrição sucinta e precisa de cada um. Não se preocupe se não conseguir lembrar de todas as fórmulas (embora deva tentar memorizar as mais simples, como a da média).
- Tentar resolver os problemas (isso é o mais importante!). As respostas de problemas ímpares são fornecidas no final do livro para que você possa conferir o resultado que obteve.

Com tudo isso, você deve ser capaz de descobrir se absorveu o conteúdo do capítulo ou não. Não fique surpreso se não tiver conseguido – mais de uma leitura será necessária. Volte às partes em que você se sente inseguro e use essas mesmas técnicas de aprendizado para cada capítulo do livro.

Resumo

- As estatísticas descritivas são úteis como síntese de grandes volumes de informação, destacando as características principais, mas omitindo os detalhes.
- Distintas técnicas são apropriadas à representação de diferentes tipos de dados, a saber: gráficos de barras para dados em *cross section* e taxas de crescimento para séries temporais.
- Métodos gráficos, como o gráfico de barras, dão uma visão dos dados. Fornecem uma síntese informal, mas são inadequados como base de análise adicional.
- Técnicas gráficas importantes incluem o gráfico de barras, a distribuição de freqüências, a distribuição de freqüência relativa e a de freqüência acumulada, o histograma e o gráfico de pizza. Para dados em série temporal, um gráfico da série de dados é informativo.
- Técnicas numéricas são mais precisas como síntese. A base dessas técnicas é formada por medidas de posição (como a média), dispersão (a variância) e assimetria.
- Estatísticas numéricas de síntese importantes são a média, a mediana e a moda; a variância, o desvio-padrão e o coeficiente de variação; o coeficiente de assimetria.
- No caso de dados bivariados, o diagrama de dispersão (ou gráfico XY) é uma maneira útil de ilustrar os dados.
- Dados freqüentemente são transformados de alguma maneira antes da análise, por exemplo, calculando-se logaritmos. Em geral as transformações ampliam a visualização de características essenciais das informações em gráficos e às vezes facilitam a interpretação de estatísticas de síntese. Por exemplo, no caso de séries temporais, a taxa média de crescimento pode ser mais apropriada do que a média da série.

Termos e conceitos fundamentais

- dados em *cross section*
- gráfico de barras
- histograma
- mediana
- variância
- escore z
- transformação de dados
- *box plot*
- dados em série temporal
- gráfico de pizza
- freqüências relativa e acumulada
- moda
- desvio-padrão
- assimetria
- crescimento composto
- tabulação cruzada
- tabela de freqüências
- média
- quantis
- coeficiente de variação
- valores extremos
- diagrama de dispersão (gráfico XY)

Problemas

Os problemas mais difíceis têm o **número** em cor.

Problema 1.1

Os dados a seguir mostram a situação de mulheres com idade entre 20 e 29 anos, em termos de educação e emprego (extraídos de *General household survey*, 1991):

	Ensino superior	*A-level*	Outro grau de escolaridade	Sem escolaridade	Total
Em atividade	209	182	577	92	1.060
Desempregadas	12	9	68	32	121
Inativas	17	34	235	136	422
Tamanho da amostra	238	225	880	260	1.603

a) Desenhe um gráfico de barras do número de mulheres em atividade em cada grau de escolaridade. Ele pode ser comparado facilmente com o diagrama semelhante de 2003 (figura 1.1)?

b) Desenhe um gráfico de barras empilhadas usando todas as situações em termos de emprego, semelhante ao da figura 1.3. Comente quaisquer semelhanças e diferenças em relação ao diagrama apresentado no texto.

c) Converta os valores da tabela em porcentagens (por coluna) e produza um gráfico de barras empilhadas semelhante ao da figura 1.4. Comente quaisquer semelhanças e diferenças.

d) Faça um gráfico de pizza mostrando a distribuição das mulheres em atividade, conforme cada grau de escolaridade, e o compare ao da figura 1.5 no texto.

Problema 1.2

Os dados a seguir indicam o rendimento mediano semanal (em libras) das pessoas empregadas em tempo integral na Grã-Bretanha em 1992, por grau de escolaridade.

	Diploma superior	Outra formação superior	*A-level*	GCSE A-C	GCSE D-G	Nenhum
Homens	433	310	277	242	226	220
Mulheres	346	278	201	183	173	146

Nota: A sigla GCSE (*General Certificate of Secondary Education*) refere-se a exames de várias disciplinas que os estudantes da Inglaterra, País de Gales e Irlanda do Norte fazem para comprovar seu aproveitamento no ensino secundário obrigatório (dos 11 aos 16 anos). Os conceitos vão de A (mais alto) a G. Os alunos classificados de A a C são os "aprovados". (N. do e.)

a) De que modo básico os dados dessa tabela diferem dos dados do problema 1.1?
b) Construa um gráfico de barras que indique os rendimentos de homens e mulheres por grau de escolaridade. O que é mostrado pelo gráfico?
c) Por que seria inadequado fazer um gráfico de barras empilhadas com esses dados? Como apresentar graficamente os dados combinados de homens e mulheres? Que informação adicional é necessária para fazer isso?

Problema 1.3

Usando os dados do problema 1.1:

a) Que grau de escolaridade apresenta maior proporção de pessoas em atividade? Qual é essa proporção?
b) Em que situação de emprego há maior proporção de pessoas com diploma superior? Qual é essa proporção?

Problema 1.4

Usando os dados do problema 1.2:

a) Qual é o diferencial, em termos de rendimento mediano, entre um diplomado no ensino superior e uma pessoa com níveis avançados? Isso varia de homem para mulher?
b) Você esperava que os rendimentos *médios* apresentassem situação semelhante? Que diferenças você acha que deveria haver, se é que deveria haver alguma?

Problema 1.5

A distribuição de ativos negociáveis em 1979 no Reino Unido é indicada na tabela abaixo (extraída de *Inland revenue statistics 1981*, p. 105):

Faixa	Número (milhares)	Valor (milhões de libras)
0-	1.606	148
1.000-	2.927	5.985
3.000-	2.562	10.090
5.000-	3.483	25.464
10.000-	2.876	35.656
15.000-	1.916	33.134
20.000-	3.425	104.829
50.000-	621	46.483
100.000-	170	25.763
200.000-	59	30.581

Faça um gráfico de barras e um histograma dos dados (suponha que o intervalo final tenha amplitude igual a 200.000). Fale sobre as diferenças entre os dois gráficos. Comente a respeito de qualquer diferença entre o histograma que obteve e o do ano de 2001 apresentado no texto.

Problema 1.6

Os dados fornecidos abaixo mostram o número de indústrias no Reino Unido em 1991/1992, conforme o número de empregados nelas existente:

Número de empregados	Número de empresas
1-	95.409
10-	15.961
20-	16.688
50-	7.229
100-	4.504
200-	2.949
500-	790
1.000-	332

Desenhe um gráfico de barras e um histograma com esses dados (suponha que o ponto médio do último intervalo seja igual a 2.000). Quais são as principais características em cada representação? E quais são as diferenças?

Problema 1.7

Usando os dados do problema 1.5:

a) Calcule a média, a mediana e a moda da distribuição. Por que são diferentes?
b) Calcule o intervalo entre quartis, a variância, o desvio-padrão e o coeficiente de variação dos dados.
c) Calcule a assimetria da distribuição.
d) Com base no que você calculou, bem como nos dados do capítulo, é possível chegar a alguma conclusão a respeito do grau de desigualdade de riqueza e sobre como isso se modificou?
e) Qual seria o efeito sobre a média se a amplitude da classe final fosse igual a £10 milhões? Quais seriam os efeitos sobre a mediana e a moda?

Problema 1.8

Usando os dados do problema 1.6:

a) Calcule a média, a mediana e a moda da distribuição. Por que são diferentes?
b) Calcule o intervalo entre quartis, a variância, o desvio-padrão e o coeficiente de variação dos dados.
c) Calcule o coeficiente de assimetria da distribuição.

Problema 1.9

Um motorista registrou suas compras de combustível numa viagem longa, a saber:

Posto de combustível	1	2	3
Litros adquiridos	33	40	25
Preço por litro	55,7	59,6	57,0

Calcule o preço médio do combustível nessa viagem.

Problema 1.10

Demonstre que o cálculo da média ponderada apresentado na equação 1.9 do texto é equivalente à determinação dos gastos totais com educação divididos pelo número de pessoas.

Problema 1.11

Num teste ministrado a cem estudantes, a nota média foi igual a 65, com variância de 144. O estudante A tirou 83 e o estudante B obteve 47.

a) Calcule os escores z para esses dois estudantes.
b) Qual é o número máximo de estudantes com um escore melhor do que o de A ou pior do que o de B?
c) Qual é o número máximo de estudantes com um escore melhor do que o de A?

Problema 1.12

A renda média de um grupo de pessoas é igual a £8.000. 80% dos membros do grupo têm renda na faixa entre £6.000 e £10.000. Qual é o valor mínimo do desvio-padrão da distribuição?

Problema 1.13

Os dados a seguir mostram os registros de automóveis no Reino Unido no período de 1970 a 1991:

Ano	Registros	Ano	Registros	Ano	Registros
1970	91,4	1978	131,6	1986	156,9
1971	108,5	1979	142,1	1987	168,0
1972	177,6	1980	126,6	1988	184,2
1973	137,3	1981	124,5	1989	192,1
1974	102,8	1982	132,1	1990	167,1
1975	98,6	1983	150,5	1991	133,3
1976	106,5	1984	146,6		
1977	109,4	1985	153,5		

Fonte: *ETAS 1993*, p. 57.

a) Faça um gráfico da série de registros de automóveis. Comente as principais características da série.

b) Faça gráficos da variação dos registros, do logaritmo natural dos registros e da variação do logaritmo. Comente os resultados.

Problema 1.14

A tabela abaixo apresenta as diferentes categorias de investimento no período de 1983-2002.

Ano	Residências	Transporte	Máquinas	Ativos fixos intangíveis	Outros prédios
1983	10.447	4.781	18.377	1.728	16.157
1984	11.932	5.938	20.782	2.229	17.708
1985	12.219	6.726	24.349	2.458	18.648
1986	14.140	6.527	25.218	2.184	20.477
1987	16.548	7.872	28.225	2.082	24.269
1988	21.097	9.227	32.614	2.592	30.713
1989	22.771	10.624	38.417	2.823	36.689
1990	21.048	10.571	37.776	3.571	41.334
1991	18.339	9.051	35.094	4.063	38.632
1992	18.825	8.420	35.071	3.782	34.485
1993	19.892	9.315	35.316	3.648	32.856
1994	21.233	11.395	38.226	3.613	33.847
1995	21.664	11.295	45.012	3.939	35.538
1996	22.516	12.222	50.197	4.136	37.220
1997	23.928	12.972	51.533	4.249	41.094
1998	25.222	16.143	59.512	4.547	45.116
1999	25.700	15.067	59.766	4.645	49.469
2000	27.394	13.444	62.698	4.966	52.708
2001	29.311	15.168	61.461	5.016	55.735
2002	35.597	15.825	54.624	5.542	58.384

Utilize técnicas gráficas apropriadas para analisar as propriedades de qualquer uma das séries de investimento. Comente os resultados.

Problema 1.15

Usando os dados do problema 1.13:

a) Calcule a taxa média de crescimento da série.

b) Calcule o desvio-padrão em torno da taxa média de crescimento.

c) A série parece ser mais ou menos volátil do que os dados de investimento utilizados no capítulo? Apresente os motivos.

Problema 1.16

Usando os dados do problema 1.14:

a) Calcule a taxa média de crescimento da série de residências.
b) Calcule o desvio-padrão em torno da taxa média de crescimento.
c) A série parece ser mais ou menos volátil do que os dados de investimento utilizados no capítulo? Apresente os motivos.

Problema 1.17

Como você esperaria que as seguintes séries se apresentassem se fossem representadas graficamente? Por exemplo: com tendência? Com tendência linear? Com tendência crescente ou decrescente? Estacionária? Autocorrelacionada? Cíclica? Algum outro tipo?

a) Renda nacional nominal.
b) Renda nacional real.
c) Taxa nominal de juros.

Problema 1.18

Como você esperaria que as seguintes séries se apresentassem se fossem representadas graficamente?

a) Nível de preços.
b) Taxa de inflação.
c) Taxa de câmbio de libras por dólares.

Problema 1.19

a) Um título público que promete pagar £1.000 ao portador daqui a cinco anos é emitido, com taxa corrente de juros de mercado de 7%. Que preço você esperaria que o título tivesse agora? Qual seria seu preço daqui a dois anos? E se depois de dois anos a taxa de juros de mercado saltasse para 10%, qual seria o preço do título?
b) Um título que promete pagar £200 por ano nos próximos cinco anos é emitido. Sendo a taxa de juros de mercado igual a 7%, quanto você estaria disposto a pagar pelo título? Por que a resposta difere da obtida na questão anterior? (Suponha que os juros sejam pagos no final de cada ano.)

Problema 1.20

Uma empresa compra por £30.000 uma máquina cuja durabilidade é de dez anos, após os quais será vendida por seu valor de liquidação, de £3.000. Calcule a taxa média de depreciação por ano e o valor líquido depreciado da máquina após um, dois e cinco anos.

Problema 1.21

A depreciação de automóveis BMW e Mercedes é fornecida pela tabela a seguir:

Idade	BMW 525i	Mercedes 200E
0 km	22.275	21.900
1 ano	18.600	19.700
2 anos	15.200	16.625
3 anos	12.600	13.950
4 anos	9.750	11.600
5 anos	8.300	10.300

a) Calcule a taxa média de depreciação de cada modelo de automóvel.
b) Use as taxas de depreciação calculadas para estimar o valor do automóvel após um, dois, etc. anos de uso. Como isso se compara aos valores efetivos?
c) Faça um gráfico dos valores calculados e dos valores estimados de cada automóvel.

Problema 1.22

Um título emitido promete pagar £400 por ano permanentemente. Quanto vale o título agora, caso a taxa de juros seja de 5%? (Dica: a soma de uma série infinita da forma

$$\frac{1}{1+r} + \frac{1}{(1+r)^2} + \frac{1}{(1+r)^3} + \ldots$$

é $1/r$, desde que $r > 0$.)

Problema 1.23

Demonstre, usando a notação Σ, que $E(x + k) = E(x) + k$.

Problema 1.24

Demonstre, usando a notação Σ, que $V(kx) = k^2V(x)$.

Problema 1.25

Critique o seguinte raciocínio estatístico. O preço médio de uma residência é £54.150. O empréstimo hipotecário médio, £2.760. Portanto, os compradores precisam pagar £21.390 de seu bolso, ou seja, aproximadamente 40% do preço de compra. Em quaisquer termos, essa é uma quantia enorme a dispender, que os casais jovens, em particular os que estão comprando uma casa pela primeira vez, teriam grande dificuldade para conseguir.

Problema 1.26

Critique o seguinte raciocínio estatístico. Entre os diplomados em artes, 10% são incapazes de encontrar emprego. Entre os formados em ciências, somente 8% ficam desempregados. Portanto, os diplomados em ciências são melhores do que os formados em artes. (Dica: imagine que existam dois tipos de emprego: popular e impopular. Os diplomados em artes tendem a se candidatar aos primeiros, e os cientistas, aos últimos.)

Problema 1.27 (Projeto 1)

Tem sido alegado que, apesar do desejo do governo do Reino Unido em promover uma tributação mais baixa, o nível de impostos em 2000 era mais alto do que o de 1979. Essa afirmação é correta? Você deve coletar dados que considere apropriados para essa tarefa, sintetizá-los como seja necessário e escrever um relatório sucinto com suas constatações. Você talvez queira analisar os seguintes aspectos:

- Deve-se considerar a receita tributária total, ou a receita como proporção do PNB?
- Deve-se distinguir entre alíquotas de imposto e a base do imposto (isto é, o que é tributado)?
- O equilíbrio entre tributação direta e indireta tem se alterado?
- Segmentos diferentes da população têm sido tributados de maneira distinta?

Você pode considerar outros pontos e resolver o problema para um outro país. Fontes adequadas de dados sobre o Reino Unido são: *Inland revenue statistics*, *UK national accounts*, *Annual abstract of statistics* ou *Financial statistics*.

Problema 1.28 (Projeto 2)

A experiência de emprego e desemprego na economia do Reino Unido é pior do que a de seus concorrentes? Escreva um relatório sobre esse tema em formato semelhante ao do projeto anterior. Você pode considerar taxas de desemprego no Reino Unido e em outros países; tendências de desemprego em cada um dos países; crescimento do emprego em cada país; sua estrutura (por exemplo, tempo integral/temporário) e também a do desemprego (por exemplo, longo prazo/curto prazo). Você pode usar dados sobre vários países, ou concentrar-se em dois países com maior profundidade. Algumas fontes de dados adequadas são: *OECD Main Economic Indicators*, *European Economy* (publicado pela Comissão Européia), *Employment Gazzette*.

APÊNDICE 1A: Notação Σ

O símbolo grego Σ (sigma maiúsculo) significa "soma" e é uma maneira simplificada de escrever o que, não fosse o seu uso, seria uma expressão algébrica muito longa. Por exemplo, dadas as seguintes observações de x:

x_1	x_2	x_3	x_4	x_5
3	5	6	4	8

então

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 3 + 5 + 6 + 4 + 8 = 26$$

Para expandir a expressão com sigma, o subscrito i é substituído por números inteiros sucessivos, começando com o número sob o sinal Σ e terminando com o número acima do sinal. De maneira similar:

$$\sum_{i=2}^{4} x_i = x_2 + x_3 + x_4 = 5 + 6 + 4 = 15$$

Quando está claro qual é a faixa de valores assumidos por i, a fórmula pode ser simplificada para $\sum_i x_i$, ou $\sum x_i$, ou mesmo $\sum x$.

Quando há freqüências associadas a cada uma das observações, como nos dados abaixo:

i	1	2	3	4	5
x_i	3	5	6	4	8
f_i	2	2	4	3	1

então

$$\sum_{i=1}^{i=5} f_i x_i = f_1 x_1 + \ldots + f_5 x_5 = 2 \times 3 + \ldots + 1 \times 8 = 60$$

e

$$\sum f_i = 2 + 2 + 4 + 3 + 1 = 12$$

Portanto, a soma das 12 observações é 60, e a média é

$$\frac{\sum fx}{\sum f} = \frac{60}{12} = 5$$

Outros exemplos são:

$$\sum x^2 = x_1^2 + x_2^2 + \ldots + x_5^2 = 150$$

$$(\sum x)^2 = (x_1 + x_2 + \ldots + x_5)^2 = 676$$

$$\sum fx^2 = f_1x_1^2 + f_2x_2^2 + \ldots + f_5x_5^2 = 2 \times 3^2 + 2 \times 5^2 + \ldots + 1 \times 8^2 = 324$$

Usando a notação Σ, podemos ver o efeito da transformação de x ao dividir por 1.000, como fizemos no cálculo do nível médio de riqueza. Em lugar de trabalhar com x, usamos kx, em que k = 1/1.000. Ao determinar a média, calculamos:

$$(1.34) \quad \frac{\sum kx}{N} = \frac{kx_1 + kx_2 + \ldots}{N} = \frac{k(x_1 + x_2 + \ldots)}{N} = k\frac{\sum x}{N}$$

Portanto, para encontrar a média da variável original x, fomos obrigados a dividir novamente por k, isto é, multiplicar por 1.000. Em geral, sempre que cada observação numa soma é multiplicada por uma constante, a constante pode ser colocada para fora do operador de soma, como em 1.34.

Problemas usando a notação Σ

Problema 1A.1

A partir dos seguintes dados de x_i: {4, 6, 3, 2, 5}, avalie:

$$\sum x_i, \sum x_i^2, (\sum x_i)^2, \sum (x_i - 3), \sum x_i - 3, \sum_{i=2}^{4} x_i$$

Problema 1A.2

A partir dos seguintes dados de x_i: {8, 12, 6, 4, 10}, avalie:

$$\sum x_i, \sum x_i^2, (\sum x_i)^2, \sum (x_i - 3), \sum x_i - 3, \sum_{i=2}^{4} x_i$$

Problema 1A.3

Dadas as seguintes freqüências, f_i, associadas aos valores de x no problema 1A.1: {5, 3, 3, 8, 5}, avalie:

$$\sum fx, \sum fx^2, \sum f(x - 3), \sum fx - 3$$

Problema 1A.4

Dadas as seguintes freqüências, f_i, associadas aos valores de x no problema 1A.2: {10, 6, 6, 16, 10}, avalie:

$\sum fx, \sum fx^2, \sum f(x-3), \sum fx-3$

Problema 1A.5

Dados os seguintes pares de observações de x e y:

x	4	3	6	8	12
y	3	9	1	4	3

avalie:

$\sum xy, \sum x(y-3), \sum (x+2)(y-1)$

Problema 1A.6

Dados os seguintes pares de observações de x e y:

x	3	7	4	1	9
y	1	2	5	1	2

avalie:

$\sum xy, \sum x(y-2), \sum (x-2)(y+1)$

Problema 1A.7

Demonstre que:

$$\frac{\sum f(x-k)}{\sum f} = \frac{\sum fx}{\sum f} - k$$

em que k é uma constante.

Problema 1A.8

Demonstre que:

$$\frac{\sum f(x-\mu)^2}{\sum f} = \frac{\sum fx^2}{\sum f} - \mu^2$$

APÊNDICE 1B: Operadores E e V

Esses operadores constituem uma notação extremamente útil da qual faremos uso mais adiante. É muito fácil controlar os efeitos de transformações de dados utilizando essa notação. Há umas poucas regras simples de manipulação desses operadores, as quais permitem que alguns problemas sejam resolvidos rápida e elegantemente.

E(x) é a média de uma distribuição e V(x) é a sua variância. Mostramos em 1.34 que a multiplicação de cada observação por uma constante k multiplica a média por k. Podemos expressar essa propriedade da seguinte maneira:

(1.35) $E(kx) - kE(x)$

Se uma constante for adicionada a cada observação, o efeito será somar essa constante à média (veja o problema 1.23):

(1.36) $E(x + a) = E(x) + a$

(Graficamente, a distribuição inteira se desloca a unidades para a direita e, portanto, o mesmo ocorre com a média.) Combinando 1.35 e 1.36:

(1.37) $E(kx + a) = kE(x) + a$

De maneira análoga, para o operador variância, pode ser mostrado que:

(1.38) $V(x + k) = V(x)$

Demonstração:

$$V(x+k) = \frac{\sum((x+k)-(\mu+k))^2}{N} = \frac{\sum((x-\mu)+(k-k))^2}{N} = \frac{\sum(x-\mu)^2}{N} = V(x)$$

(Um deslocamento de toda a distribuição não altera a variância.) Além disso:

(1.39) $V(kx) = k^2V(x)$

(Veja o problema 1.24.) Esse é o motivo pelo qual, ao dividir os dados de riqueza por 1.000, dividimos a variância por 1.000^2. Aplicando 1.38 e 1.39:

(1.40) $V(kx + a) = k^2V(x)$

Finalmente, devemos observar que V pode ser expresso em termos de E:

(1.41) $V(x) = E(x - E(x))^2$

APÊNDICE 1C: Utilização de logaritmos

Os logaritmos são utilizados com menos freqüência hoje em dia, dada a existência de calculadoras eletrônicas baratas. Entretanto, a transformação logarítmica é útil em outros contextos de estatística e economia, e seu uso é exposto sucintamente a seguir.

O logaritmo (à base 10) de um número x é definido como a potência à qual o número 10 precisa ser elevado para gerar o valor de x. Por exemplo, $10^2 = 100$, de modo que o log de 100 é 2; escrevemos $\log_{10}100 = 2$, ou simplesmente $\log 100 = 2$.

De maneira semelhante, o log de 1.000 é 3 ($1.000 = 10^3$), de 10.000 é 4, etc. Não estamos limitados ao uso de potências inteiras de 10, de modo que, por exemplo, $10^{2,5} = 316,227766$ (tente fazer esse cálculo com uma calculadora científica), e assim o log de 316,227766 é 2,5. Cada número x, portanto, pode ser representado por seu logaritmo.

Multiplicação de dois números

Podemos usar logaritmos para multiplicar dois números, x e y, com base na seguinte propriedade:

$$\log xy = \log x + \log y$$

Por exemplo, multiplicando 316,227766 por 10:

$$\begin{aligned}\log(316,227766 \times 10) &= \log 316,227766 + \log 10\\ &= 2,5 + 1\\ &= 3,5\end{aligned}$$

O *antilog* de 3,5 é dado por $10^{3,5} = 3.162,27766$, que é a resposta que buscamos.

Calcular o antilog (ou seja, o número 10 elevado a um expoente) é a operação inversa da transformação logarítmica. Esquematicamente, temos:

x → calcular logaritmo → a (= log x) → elevar 10 ao expoente a → x

Divisão

Para dividir um número por outro, subtraímos seus logaritmos. Por exemplo, dividindo 316,227766 por 100:

$$\begin{aligned}\log(316,227766/100) &= \log 316,227766 - \log 100\\ &= 2,5 - 2\\ &= 0,5\end{aligned}$$

e

$$10^{0,5} = 3,16227766$$

Potências e raízes

Os logaritmos simplificam o processo de elevação de um número a um expoente. Para encontrar o quadrado de um número, multiplica-se seu logaritmo por 2, ou seja, para encontrar $316{,}227766^2$:

$$\log(316{,}227766^2) = 2\log(316{,}227766) = 5$$

e

$$10^5 = 100.000$$

Para encontrar a raiz quadrada de um número (o que equivale a elevá-lo a ½), divide-se o log por 2. Para se calcular a raiz n, divide-se o log por n. Por exemplo, no texto foi necessário encontrar a 32ª raiz de 16,936:

$$\frac{\log(16{,}936)}{32} = 0{,}0384$$

e

$$10^{0{,}0384} = 1{,}092$$

Logaritmos comuns e naturais

Os logaritmos na base 10 são conhecidos como logaritmos comuns, mas pode-se usar qualquer número como base. Os logaritmos *naturais* têm como base o número e (= 2,71828...), e escrevemos ln x, em lugar de log x, para distingui-los dos logaritmos comuns. Assim, por exemplo:

$$\ln 316{,}227766 = 5{,}756462732$$

pois

$$e^{5{,}756462732} = 316{,}227766$$

Os logaritmos naturais podem ser usados da mesma forma que os logaritmos comuns e têm propriedades semelhantes. Utilize a tecla "ln" em sua calculadora, assim como usaria a tecla "log", mas lembre que a transformação inversa é e^x, e não 10^x.

Problemas sobre logaritmos

Problema 1C.1

Calcule os logaritmos comuns de: 0,15; 1,5; 15; 150; 1.500; 83,7225; 9,15; –12.

Problema 1C.2

Calcule o log dos seguintes valores: 0,8; 8; 80; 4; 16; –37.

Problema 1C.3

Encontre os logaritmos naturais de: 0,15; 1,5; 15; 225; –4.

Problema 1C.4

Calcule o ln dos seguintes valores: 0,3; e; 3; 33; –1.

Problema 1C.5

Encontre o antilog dos seguintes valores: –0,823909; 1,1; 2,1; 3,1; 12.

Problema 1C.6

Encontre o antilog dos seguintes valores: –0,09691; 2,3; 3,3; 6,3.

Problema 1C.7

Encontre o anti-ln dos seguintes valores: 2,70805; 3,70805; 1; 10.

Problema 1C.8

Encontre o anti-ln dos seguintes valores: 3,496508; 14; 15; –1.

Problema 1C.9

Avalie: $\sqrt[2]{10}, \sqrt[4]{3,7}, 4^{1/4}, 12^{-3}, 25^{-3/2}$.

Problema 1C.10

Avalie: $\sqrt[3]{30}, \sqrt[6]{17}, 8^{1/4}, 15^{0}, 12^{0}, 3^{-1/3}$.

Perspectiva**brasileira**

Crescimento no Brasil: investimento e PIB

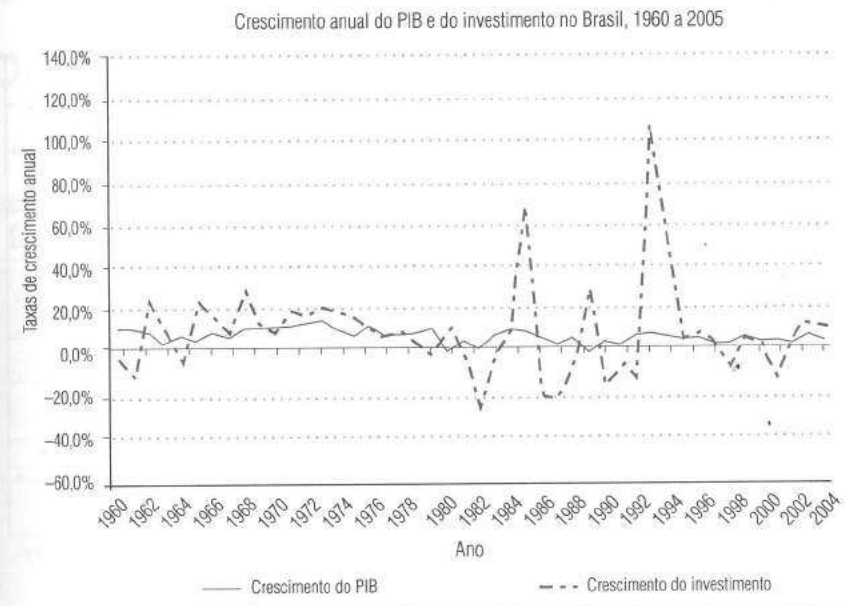
A tabela 1 contém dados anuais referentes ao crescimento do produto interno bruto (PIB) e do investimento no Brasil (ou seja, a chamada "formação bruta de capital fixo") – ambos em termos reais, isto é, descontada a inflação ocorrida no período de 46 anos. Os dados foram publicados pela revista *Conjuntura Econômica*, da Fundação Getulio Vargas, na edição de março de 2007, p. XIX.

TABELA 1 Taxas de crescimento do investimento real e do PIB em termos reais no Brasil, 1960-2005

Ano	Crescimento do PIB	Crescimento do investimento	Ano	Crescimento do PIB	Crescimento do investimento
1960	9,6%	–8,0%	1983	–2,9%	–36,1%
1961	8,8%	–17,7%	1984	5,4%	–5,5%
1962	6,5%	25,3%	1985	7,9%	7,1%
1963	0,4%	9,4%	1986	7,5%	78,9%
1964	3,6%	–10,4%	1987	3,6%	–29,5%
1965	2,4%	19,4%	1988	–0,1%	–30,9%
1966	6,8%	14,5%	1989	3,2%	–14,0%
1967	4,4%	7,6%	1990	–5,1%	26,8%
1968	9,7%	27,9%	1991	1,0%	–24,2%
1969	9,4%	12,2%	1992	–0,5%	–13,9%
1970	10,4%	6,4%	1993	4,9%	–17,9%
1971	11,4%	19,2%	1994	5,9%	122,9%
1972	11,9%	17,7%	1995	4,2%	60,0%
1973	13,9%	22,1%	1996	2,7%	3,2%
1974	8,1%	16,9%	1997	3,3%	7,3%
1975	5,2%	16,0%	1998	0,1%	2,3%
1976	10,3%	6,7%	1999	0,8%	–14,8%
1977	4,9%	4,5%	2000	4,4%	4,9%
1978	5,0%	8,2%	2001	1,3%	0,0%
1979	6,8%	–1,4%	2002	1,9%	–16,5%
1980	9,2%	–3,7%	2003	0,5%	4,1%
1981	–4,3%	6,4%	2004	4,9%	11,7%
1982	0,8%	–3,6%	2005	2,3%	10,2%

FIGURA 1

Taxas de crescimento do investimento real e do PIB em termos reais no Brasil, 1960-2005



A figura 1 mostra graficamente os dados da tabela 1, na forma de séries temporais, tal como explicadas no capítulo 1 deste livro. Já que as duas taxas de crescimento são medidas na mesma unidade (porcentagem de variação anual), usamos uma única escala para ambas as séries.

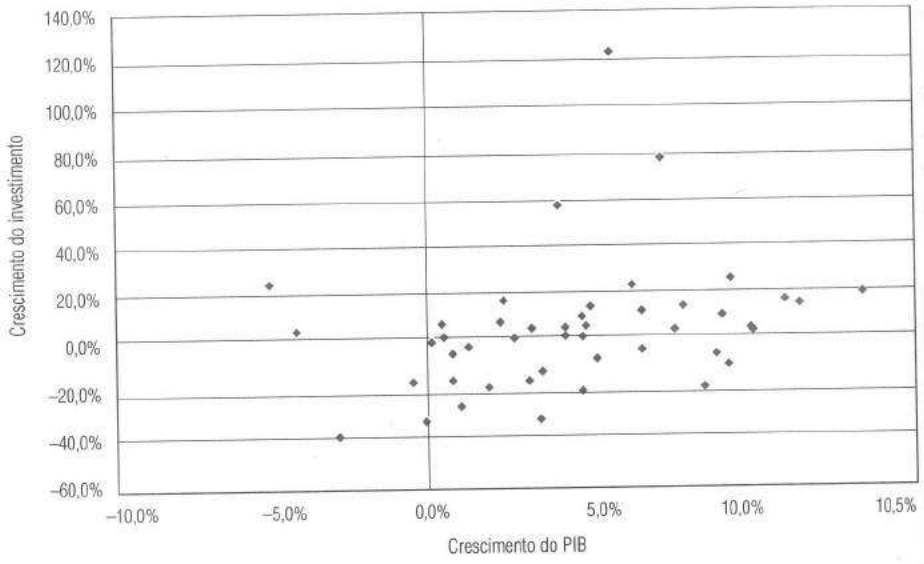
Em primeiro lugar, pode-se observar facilmente que o crescimento anual do investimento é muito mais volátil do que a série de crescimento do PIB. Em particular, destacam-se as variações ocorridas no período de 1986 a 1995, quando diversos planos de estabilização foram introduzidos. Na verdade, isso significa que boa parte da variação do investimento, em termos reais, pode ser atribuída às alterações radicais sofridas pelos índices de preços nesses anos.

Em segundo lugar, deve ser informado que a média aritmética das taxas de crescimento do PIB, no período, atingiu 4,6% ao ano. Isso mostra com clareza como as taxas mais recentes de crescimento da economia brasileira estão bastante abaixo dos níveis médios historicamente verificados, pois, com exceção da taxa de crescimento em 2004 (4,9%), precisaríamos voltar até 1994 para ver uma taxa anual de crescimento superior à média do período como um todo.

Finalmente, a figura 2 mostra o diagrama de dispersão das taxas de crescimento do PIB (no eixo horizontal) e do investimento (no eixo vertical). Como indica o texto do capítulo 1, essa é uma ferramenta descritiva útil para a verificação preliminar da existência de alguma relação entre as variáveis. Aparentemente, as duas variáveis têm uma fraca relação direta em vista da inclinação positiva que o conjunto de pontos apresenta. Po-

FIGURA 2

Diagrama de dispersão: taxas de crescimento do investimento e do produto interno bruto em termos reais no Brasil, 1960 a 2005



rém, isso era esperado, uma vez que, junto com as despesas de consumo final (privado e governamental) e o saldo da balança comercial, o investimento agregado é um dos componentes do produto interno bruto medido a preços de mercado.