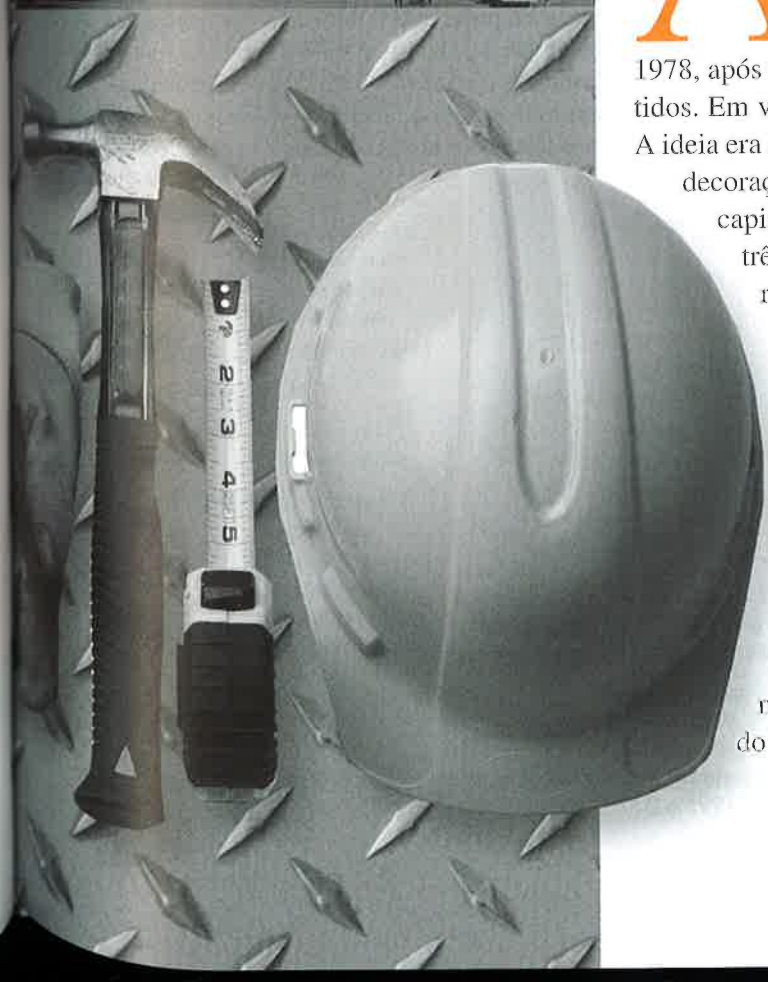
CAPÍTULO 7

Diagramas de Dispersão, Associação e Correlação



Home Depot

A Handy Dan era uma rede de lojas de materiais de construção e decoração de sucesso nos anos 1970, graças, em parte, aos esforços de dois dos seus executivos, Bernie Marcus e Arthur Blank. Entretanto, num dia de primavera, em 1978, após uma discussão com seu chefe, os dois foram demitidos. Em vez de desanimar, eles formaram a MB Associates. A ideia era abrir uma rede de lojas de materiais de construção e decoração no estilo de um armazém. Após levantar algum capital para investimento, a dupla conseguiu inaugurar três lojas no primeiro ano, contratando 200 vendedores e gerando $7 milhões em vendas. Eles trocaram o nome das lojas para Home Depot. Cinco anos depois, Bernie e Arthur tinham 10 vezes mais lojas, 20 vezes mais vendedores e 60 vezes mais vendas. O crescimento extraordinário continuou nos anos 1980 e 1990. A Home Depot atingiu o marco de $30, $40, $50 e $60 bilhões em vendas mais rapidamente do que qualquer outro varejista na história. Em 2005, ela era a segunda maior loja varejista nos Estados Unidos, atrás somente da Wal-Mart, e a terceira maior loja varejista do mundo, atrás da Wal-Mart e do gigante supermercado francês Carrefour.

Em 2000, Bernie e Arthur se aposentaram para desempenhar atividades filantrópicas. Robert (Bob) L. Nardelli, que trabalhou quase 30 anos na General Electric, onde se tornou presidente e CEO da GE Power Systems, se tornou diretor, presidente e CEO da Home Depot. Depois da posse de Nardelli, a empresa continuou a crescer, quase dobrando novamente nesses cinco anos, mas o preço da ação definhou. Em setembro de 2006, Nardelli foi reeleito como presidente na reunião anual dos acionistas, mas aproximadamente um terço dos acionistas retirou seu apoio, mencionando seu desapontamento com o preço da ação e questionando seus rendimentos, que, num período de mais de cinco anos, totalizou $123,7 milhões. Nardelli abruptamente anunciou sua retirada, em 3 de janeiro de 2007 – com uma indenização estimada em mais de $200 milhões –, afirmando ter sido uma "decisão mútua". Frank Blake, que se juntou à Home Depot em 2002 como vice-presidente executivo, tornou-se o CEO após a saída de Nardelli.

Quem	Trimestres econômicos
O quê	*Novas Residências* sazonalmente ajustadas por trimestres nos Estados Unidos
Unidades	Milhares de unidades
Quando	1995 – 2005
Onde	Estados Unidos
Por quê	Para examinar as tendências na construção de residências

Durante a década de 1995 a 2005, a Home Depot vivenciou um crescimento extraordinário. Naturalmente, a empresa está interessada nas tendências do mercado de construções de casas. Qual tem sido a tendência neste ramo? Eis um diagrama mostrando o número de *Novas Residências* (sazonalmente ajustadas por trimestres, em milhares de unidades) por *Trimestres* para aquela década. Se tivesse de resumir a tendência das novas residências nessa década, o que você diria?

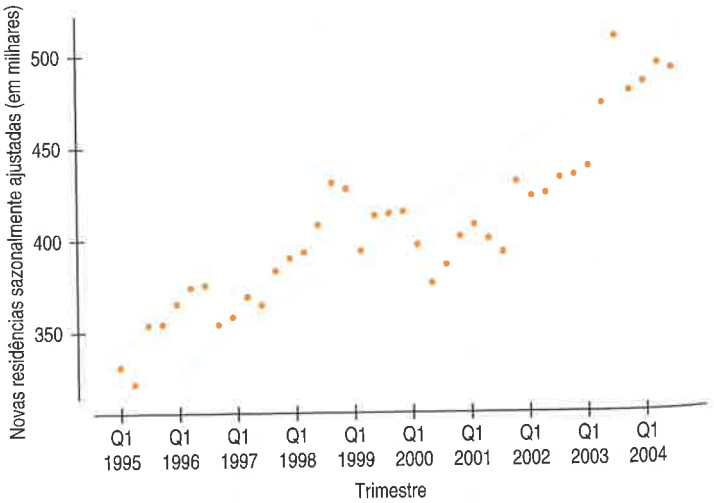


Figura 7.1 Novas residências sazonalmente ajustadas por trimestres (em milhares), de 1995 a 2005.

Evidentemente, as novas residências cresceram entre 1995 e 2005, começando abaixo de 350000 por trimestre e terminando próximo de 500000. O diagrama mostra um crescimento estável de 1995 até o final de 1999, época em que começa a decrescer, e novamente um crescimento a partir de 2001.

Esse diagrama de séries temporais é um exemplo de um tipo mais geral de representação gráfica, chamado de diagrama de dispersão. **Um diagrama de dispersão**, que representa uma variável quantitativa em relação a outra, pode ser uma apresentação eficiente para os dados. Sempre que você quiser entender o relacionamento entre duas variáveis quantitativas, deve fazer um diagrama de dispersão. Ao analisar o diagrama de dispersão, você consegue ver padrões, tendências, relacionamentos e até mesmo valores incomuns ocasionais, que diferem dos outros. Os diagramas de dispersão são a melhor maneira de começar a observação da relação entre duas variáveis *quantitativas*.

As relações entre variáveis estão geralmente no núcleo do que gostaríamos de aprender sobre os dados.

- A confiança do consumidor está relacionada aos preços do petróleo?
- O que acontece com a satisfação do consumidor à medida que as vendas aumentam?
- Um aumento de dinheiro gasto com propaganda está relacionado às vendas?
- Qual é a relação entre as vendas de ações e seus preços?

Perguntas como essas relacionam duas variáveis quantitativas e perguntam se existe uma **associação** entre elas. Os diagramas de dispersão são a maneira ideal de ilustrar tais associações.

Quem	Cidades nos Estados Unidos
O quê	*Custo do Congestionamento* por pessoa e período de pico na *Velocidade na Autoestrada*
Unidades	*Custo do Congestionamento* por pessoa ($ por pessoa por ano); período de pico *Velocidade na Autoestrada* (mph)
Quando	2000
Onde	Em todos os Estados Unidos
Por quê	Para examinar a relação entre o congestionamento em autoestradas e seu impacto na sociedade e nos negócios

7.1 Analisando os diagramas de dispersão

O Instituto de Transporte do Texas, que estuda a mobilidade fornecida pelo sistema nacional de transporte, publica um relatório anual do congestionamento de tráfego e seus custos para a sociedade e a economia. A Figura 7.2 mostra um diagrama de dispersão do *Custo do Congestionamento* anual por pessoa em atrasos no trânsito (em dólares), em 65 cidades nos Estados Unidos, *versus* Período de Pico na *Velocidade na Autoestrada* (mph).

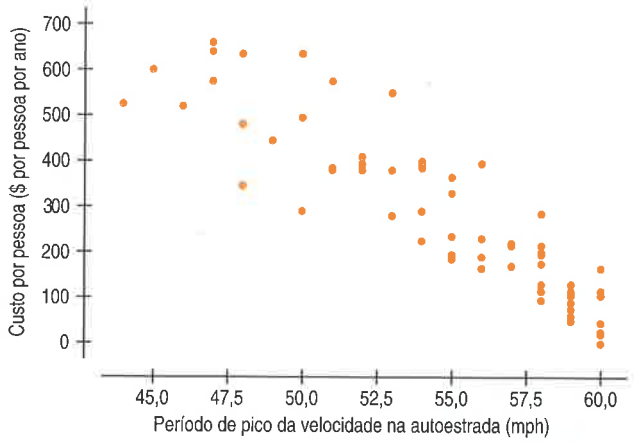


Figura 7.2 Custo do congestionamento por pessoa ($ por ano) de atrasos no trânsito *versus* o período de pico na velocidade na autoestrada (mph), para 65 cidades dos Estados Unidos.

Todos observam os diagramas de dispersão, mas, se perguntados, muitos julgariam difícil dizer o que observar num diagrama de dispersão. O que *você* vê? Tente descrever o diagrama de dispersão do *Custo do Congestionamento* versus *Velocidade na Autoestrada*.

Procure pela **Direção**: qual é o sinal – positivo, negativo ou nenhum deles?

Talvez você diga que a **direção** da associação é importante. À medida que o pico da velocidade na autoestrada sobe, o custo do congestionamento desce. Um padrão que vai do canto superior esquerdo ao canto inferior direito  é chamado de **negativo**. Um padrão que vai na outra direção  é chamado de **positivo**.

O segundo aspecto a ser procurado no diagrama de dispersão é sua **forma**. Se existe uma relação linear, ela irá aparecer como uma nuvem ou um agrupamento de pontos estendido geralmente numa forma consistente e reta. Por exemplo, o diagrama de dispersão do congestionamento do tráfego tem uma forma **linear** subjacente, embora alguns pontos se desviem dela.

Os diagramas de dispersão podem revelar diferentes tipos de padrões. Muitas vezes eles não serão lineares, mas padrões de linhas retas são mais comuns e mais úteis para a estatística.

Procure pela **Forma**: linear, curva, algo exótica ou mesmo sem padrão?

Se a relação não for linear, mas uma curva suave sempre crescente ou decrescente,  em geral, podemos encontrar maneiras de linearizá-la. Porém, se ela é uma curva aguda que sobe e desce, por exemplo,  então você precisará de métodos mais avançados.

A terceira característica a ser observada num diagrama de dispersão é a **força** da relação.

Procure por **Força**: quanta dispersão?

Num extremo, os pontos parecem estar bem aglomerados num único bloco  (retos, curvados ou se curvando por todos os lados)? Ou os pontos parecem ser tão variáveis e espalhados, que dificilmente podemos perceber qualquer tendência ou padrão?  O diagrama do congestionamento do tráfego mostra dispersão moderada ao redor de uma forma geralmente linear. Isso indica que existe uma relação moderadamente linear entre custo e velocidade.

Procure por **Características Incomuns**: existem observações incomuns ou subgrupos?

Finalmente, sempre procure pelo inesperado. Geralmente, a descoberta mais interessante num diagrama de dispersão é algo que você nunca pensou em procurar. Um exemplo de tal surpresa é uma observação incomum, ou **atípica**, fora do padrão geral do diagrama de dispersão. Tal ponto é quase sempre interessante e merece uma atenção especial. Você pode ver grupos inteiros ou subgrupos que se afastam ou mostram uma tendência numa direção diferente do resto do diagrama. Isso deve levantar questões sobre por que eles são diferentes. Talvez seja um sinal de que você deve separar os dados em subgrupos, em vez de observá-los todos juntos.

Descartes foi um filósofo, famoso por sua afirmação *cogito ergo sum*: penso, logo existo.

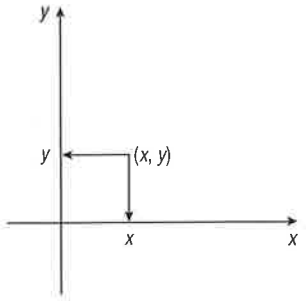
7.2 Atribuindo papéis a variáveis nos diagramas de dispersão

Os diagramas de dispersão foram umas das primeiras representações da matemática moderna. A ideia de utilizar dois eixos em ângulos retos para definir uma área onde exibir valores surgiu com René Descartes (1596 – 1650), e a área que ele definiu dessa maneira é chamada de *Plano Cartesiano*, em sua homenagem.

Os dois eixos que Descartes especificou caracterizam o diagrama de dispersão. O eixo vertical, por convenção, é chamado de eixo *y* e o eixo horizontal é chamado de

eixo *x*. Esses termos são padrão. Se alguém se referir ao eixo *y*, você pode ter certeza de que o eixo é vertical; o mesmo vale para o eixo *x*, horizontal.[1]

Para fazer um diagrama de dispersão de duas variáveis quantitativas, atribua uma ao eixo *y* e a outra ao eixo *x*. Como com qualquer gráfico, certifique-se de rotular os eixos claramente e indique as escalas dos eixos com números. Os diagramas de dispersão exibem variáveis *quantitativas*. Cada variável tem unidades e elas devem aparecer na representação – em geral, próximas de cada eixo. Cada ponto é colocado num diagrama de dispersão numa posição que corresponda aos valores dessas duas variáveis. Sua localização horizontal é especificada pelo seu valor *x* e sua localização vertical, pela variável de valor *y*. Juntas, elas são conhecidas como *coordenadas* e escritas como (*x*, *y*).



Os diagramas de dispersão feitos por programas de computador (como os dois que vimos neste capítulo), em geral, não mostram – e normalmente não deveriam mostrar – a *origem,* o ponto $x = 0$, $y = 0$, onde os eixos se encontram. Se as duas variáveis têm valores próximos ou em ambos os lados de zero, então a origem será parte da exibição. Se os valores estão longe de zero, não há motivos para incluir a origem. De fato, é melhor focar a parte do plano cartesiano que contém os dados. No nosso exemplo sobre autoestradas, nenhuma das velocidades estava próxima de 0 mph, portanto, o computador traçou o diagrama de dispersão na Figura 7.2 com eixos que não se cruzam.

Qual variável deve ir no eixo *x* e qual deve ir no eixo *y*? O que queremos saber a partir da relação pode nos indicar como fazer o gráfico. Geralmente, temos perguntas como estas:

- A satisfação dos empregados da Home Depot está relacionada com a produtividade?
- O aumento das vendas na Home Depot terá reflexos no preço das ações?
- Que outros fatores econômicos, além da construção de novas residências, estão relacionados às vendas da Home Depot?

Em todos esses exemplos, uma variável tem o papel de **explanatória** ou **variável previsora,** enquanto a outra tem o papel de **variável resposta**. Colocamos a variável explanatória no eixo *x* e a variável resposta no eixo *y*. Portanto, ao criar um diagrama de dispersão, escolha cuidadosamente quais variáveis atribuir a cada eixo.

ALERTA DE NOTAÇÃO:

As letras *x* e *y* não servem apenas para rotular os eixos de um diagrama de dispersão. Em estatística, a designação de variáveis para os eixos *x* e *y* (e a escolha da notação para elas em fórmulas) geralmente fornece informações sobre seus papéis como previsora ou resposta.

Os papéis que escolhemos para as variáveis estão mais relacionados a como *pensamos* sobre elas do que com as variáveis propriamente ditas. Só porque a variável está no eixo *x* não significa necessariamente que ela explique ou preveja *algo,* e a variável no eixo *y* pode não ser uma resposta. Traçamos o *Custo do Congestionamento por pessoa* versus o pico da *Velocidade na Autoestrada*, pensando que, quanto mais lento for o tráfego, mais ele custa, devido aos atrasos. Mas, talvez *gastando-se* $500 por pessoa em melhorias na autoestrada, a velocidade irá aumentar. Se estivermos examinando esta opção, podemos escolher traçar o *Custo do Congestionamento por pessoa* como uma variável explanatória e a *Velocidade na Autoestrada* como a resposta.

[1] Os eixos são também chamados de "ordenada" e "abscissa" – mas nunca lembramos qual é qual, porque os estatísticos geralmente não usam esses termos. Na estatística (e em todos os programas estatísticos), os eixos, em geral, são chamados de "*x*" (abscissa) e "*y*" (ordenada) e são rotulados com os nomes das variáveis correspondentes.

Quem	Trimestres econômicos
O quê	*Vendas* trimestrais da Home Depot e *Novas Residências* trimestrais (não ajustadas) dos Estados Unidos.
Unidades	*Vendas* ($bilhões) e *Novas Residências* (milhares)
Quando	Maio de 1995 – Junho 2004
Onde	Estados Unidos
Por quê	Para examinar a associação entre vendas e novas residências.

As variáveis x e y são, às vezes, referidas como variáveis **independente** e **dependente**, respectivamente. A ideia é que a variável y *depende* da variável x e a variável x age *independentemente* para fazer y responder. Essas designações, entretanto, entram em conflito com outros usos dos mesmos termos em estatística. Assim, usaremos "variável explanatória" ou "previsora" e "variável resposta" quando discutimos papéis, mas geralmente iremos nos referir apenas a *variável x* e *variável y*.

7.3 Entendendo a correlação

Em geral, uma economia forte é acompanhada de grande consumo por parte da população. Será que isso também se aplica à indústria de materiais de construção e decoração? Durante o período de 1995 a 2005, novas residências, sazonalmente ajustadas, aumentaram drasticamente (Figura 7.1, p. 200). Vamos examinar um diagrama de dispersão dos dados não ajustados para ver se existe uma associação entre o crescimento de *Novas Residências* e *Vendas* da Home Depot. Não deve ser surpresa descobrir que existe uma relação positiva entre os dois. Como você deve suspeitar, quanto maior o número de casas construídas, mais altas são as vendas da Home Depot.

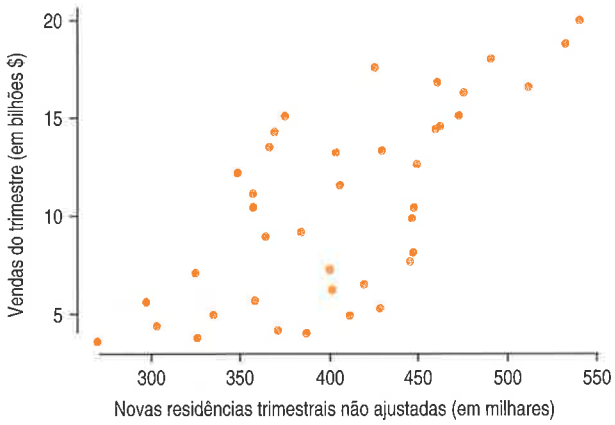


Figura 7.3 As vendas trimestrais da Home Depot ($ bilhões) e novas residências (milhares) nos Estados Unidos de 1995 a 2004.

Existe uma associação positiva evidente, e o diagrama de dispersão parece razoavelmente linear, mas quão forte é a associação? Se você tivesse de dar um número (digamos, entre 0 e 1) para a força da relação, qual seria? Sua medida não deveria depender da escolha das unidades para as variáveis. Afinal, se as vendas tivessem sido registradas em euros, em vez de dólares, ou novas residências em milhões de unidades, em vez de milhares, o diagrama de dispersão seria o mesmo. A direção, a forma e a força não irão mudar, portanto, nossa medida da força da associação também não deveria.

Visto que as unidades não importam, por que não removê-las? Para tanto, podemos padronizar ambas as variáveis, tornando as coordenadas de cada ponto em um par de escores – z, z_x e z_y. No Capítulo 6, vimos que, para padronizar valores, subtraímos a média de cada variável e dividimos pelo seu desvio padrão:

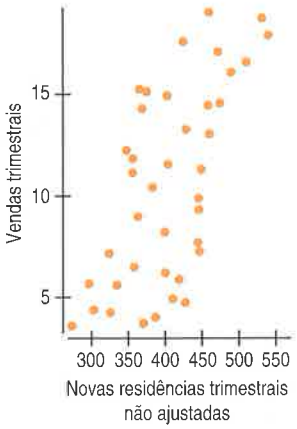
$$(z_x, z_y) = \left(\frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y}\right).$$

O diagrama de dispersão resultante parece quase o mesmo (caso você não leia as legendas dos eixos).

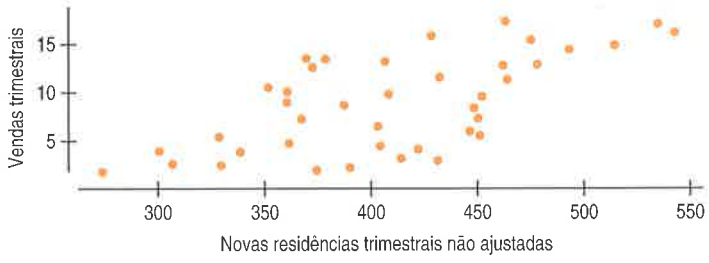
Figura 7.4 Vendas trimestrais padronizadas (z_{vendas}) *versus* novas residências padronizadas ($z_{Novas\ Residências}$).

Como a padronização torna as médias de ambas variáveis 0, o centro do novo diagrama de dispersão agora é a origem, e as escalas em ambos os eixos são unidades de desvio padrão.

Mas espere. Os dois gráficos não são *exatamente* iguais. Você consegue ver a diferença? Primeiro, o padrão linear subjacente é mais evidente no gráfico padronizado. Isso ocorre porque a padronização torna as escalas dos eixos iguais. É natural tornar o comprimento de um desvio padrão o mesmo vertical e horizontalmente. Quando trabalhamos nas unidades originais, tínhamos liberdade para deixar o gráfico alto e estreito



ou baixo e largo



conforme nossa vontade. Normalmente, fazemos o eixo *x* mais longo que o eixo *y* por razões estéticas,[2] mas escalas iguais são uma maneira neutra de desenhar diagramas de dispersão e fornecem uma impressão mais precisa da força da associação.

[2] A escolha estética para a razão entre os dois eixos está relacionada à razão áurea dos antigos gregos.

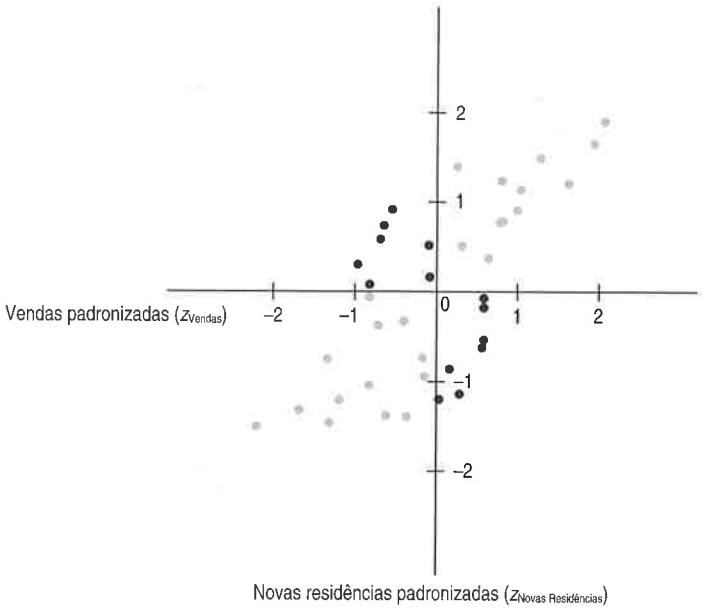


Figura 7.5 Nesse gráfico de dispersão dos escores-z, os pontos são coloridos pela forma como afetam a associação: laranja para positivo e cinza para negativo.

Visto que padronizar as variáveis não muda a força da associação, podemos trabalhar com os escores-z para entender como mensurar a força. Quais pontos no diagrama de dispersão dos escores-z (Figura 7.5) dão a impressão de uma associação positiva? Numa associação positiva, y tende a aumentar à medida que x aumenta. Portanto, os pontos na parte superior direita e na parte inferior esquerda da Figura 7.5 reforçam a relação positiva. Para esses pontos, z_x e z_y têm o mesmo sinal. Se os multiplicássemos, o produto, $z_x z_y$ seria positivo. Pontos longe da origem (que fazem a associação parecer mais negativa) têm um produto mais negativo.

Pontos com escores-z de zero em ambas as variáveis não influem, porque $z_x z_y = 0$. Esses pontos seriam encontrados nos eixos da Figura 7.5. Para transformar esses produtos em uma medida da força da associação, podemos simplesmente somar os produtos de $z_x z_y$ para cada ponto no diagrama de dispersão.

$$\sum z_x z_y$$

Isso resume a direção *e* a força da associação para todos os pontos. Se a maioria dos pontos estiver no quadrante onde os escores-z têm os mesmos sinais, a soma será positiva. Se a maioria estiver nos quadrantes onde os escores-z têm sinais opostos, ela será negativa.

Entretanto, quanto mais dados tivermos, maior o *tamanho* dessa soma. Para ajustar isso, dividimos a soma por $n - 1$.[3] Essa razão é chamada de **coeficiente de correlação**.

$$r = \frac{\sum z_x z_y}{n - 1}.$$

Para as *Vendas Trimestrais* da Home Depot e *Novas Residências,* o coeficiente de correlação é 0,70.

[3] Sim, o mesmo $n - 1$ que usamos para calcular o desvio padrão, que explicaremos nos próximos capítulos.

Existem fórmulas alternativas para o coeficiente de correlação em termos das variáveis x e y. Eis as duas mais comuns:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}.$$

Essas fórmulas talvez funcionem bem para calcular a correlação manualmente, mas a forma usando escores-z é melhor para entender o que a correlação significa. Se você quiser aprender como mudar da fórmula usando os escores-z para as outras duas fórmulas, leia o Quadro da Matemática.

ALERTA DE NOTAÇÃO:

A letra r sempre é usada para o coeficiente de correlação, portanto, você não pode aplicá-la de outra forma em estatística. Sempre que você vir um "r", com certeza trata-se de uma correlação.

QUADRO DA MATEMÁTICA

Padronizar as variáveis facilita o nosso entendimento da expressão para a correlação.

$$r = \frac{\sum z_x z_y}{n - 1}$$

Às vezes, no entanto, você verá outras fórmulas. Lembrar como a padronização funciona nos ajuda a transitar de uma fórmula para outra.

Visto que: $$z_x = \frac{x - \bar{x}}{s_x}$$

e $$z_y = \frac{y - \bar{y}}{s_y},$$

podemos substituir esses valores na expressão para a correlação acima e teremos:

$$r = \left(\frac{1}{n - 1}\right)\sum z_x z_y = \left(\frac{1}{n - 1}\right)\sum \frac{(x - \bar{x})}{s_x}\frac{(y - \bar{y})}{s_y} = \sum \frac{(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}.$$

Esta é uma versão. E como já conhecemos a fórmula para o desvio padrão

$$s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}},$$

podemos usar a substituição para escrever:

$$\begin{aligned} r &= \left(\frac{1}{n - 1}\right)\sum \frac{(x - \bar{x})}{s_x}\frac{(y - \bar{y})}{s_y} \\ &= \left(\frac{1}{n - 1}\right)\frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}\sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}} \\ &= \left(\frac{1}{n - 1}\right)\frac{\sum(x - \bar{x})(y - \bar{y})}{\left(\frac{1}{n - 1}\right)\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(y - \bar{y})^2}} \\ &= \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}. \end{aligned}$$

Esta é a outra versão comum. Se alguma vez você tiver de calcular a correlação manualmente, é mais fácil começar com uma dessas. Mas se você quer lembrar como a correlação funciona, fique com a primeira fórmula do quadro matemático.

Encontrando o coeficiente de correlação manualmente

Para encontrar o coeficiente de correlação manualmente, usaremos uma fórmula em unidades originais, em vez de escores-z. Isso economizará o trabalho de padronizar cada valor dos dados antes. Comece com o resumo estatístico para ambas as variáveis: $\bar{x}$, $\bar{y}$, s_x e s_y. Depois, encontre os desvios como fizemos para o desvio padrão, mas agora em x e y: $(x - \bar{x})$ e $(y - \bar{y})$. Para cada par de dados, multiplique estes desvios: $(x - \bar{x}) \times (y - \bar{y})$. Some o produto para todos os pares de dados. Por fim, divida a soma pelo produto de $(n - 1) \times s_x \times s_y$, a fim de encontrar o coeficiente de correlação. Lá vamos nós.

Suponha que os pares dos dados sejam:

x	6	10	14	19	21
y	5	3	7	8	12

Então $\bar{x} = 14$, $\bar{y} = 7$, $s_x = 6{,}20$ e $s_y = 3{,}39$.

Desvios em x	Desvios em y	Produto
$6 - 14 = -8$	$5 - 7 = -2$	$-8 \times -2 = 16$
$10 - 14 = -4$	$3 - 7 = -4$	16
$14 - 14 = 0$	$7 - 7 = 0$	0
$19 - 14 = 5$	$8 - 1 = 1$	5
$21 - 14 = 7$	$12 - 7 = 5$	35

Some os produtos: $16 + 16 + 0 + 5 + 35 = 72$

Finalmente, dividimos por $(n - 1) \times s_x \times s_y = (5 - 1) \times 6{,}20 \times 3{,}39 = 84{,}07$

A razão é o coeficiente de correlação:

$$r = 72/84{,}07 = 0{,}856.$$

Condições da correlação

A **correlação** mensura a força da associação *linear* entre duas variáveis *quantitativas*. Antes de usar a correlação, você deve verificar três *condições*:

- **Condição de variáveis quantitativas:** A correlação se aplica somente às variáveis quantitativas. Não use a correlação com dados categóricos mascarados como quantitativos. Certifique-se de que você conheça as unidades das variáveis e o que elas mensuram.

- **Condição de linearidade:** Claro, é possível *calcular* o coeficiente de correlação para qualquer par de variáveis. No entanto, a correlação mensura somente a força da associação *linear* e será enganosa se a relação não for linear o suficiente. O que é "linear o suficiente"? Essa pergunta pode parecer muito informal para uma condição estatística, mas é importante. Não podemos verificar se um relacionamento é linear ou não. Poucas relações entre variáveis são perfeitamente lineares, mesmo em teoria, e diagramas de dispersão de dados reais nunca o são. Quão não linear um diagrama de dispersão deve ser para falhar na condição? Você deve considerar essa questão. Você acha que a relação subjacente é curva? Se for, então resumir sua força com a correlação será um equívoco.

- **Condição do valor atípico:** Observações incomuns podem distorcer a correlação, fazendo que uma pequena correlação pareça grande ou, por outro lado, escondendo uma correlação grande. Podem até mesmo dar a uma associação positiva um coeficiente de correlação negativo (ou vice versa). Quando você tiver um valor atípico, em geral, é recomendável relatar a correlação com e sem o ponto.

É fácil verificar cada uma dessas condições com um diagrama de dispersão. Muitas correlações são relatadas sem o suporte de dados ou de um gráfico. Você também deve pensar sobre as condições. Seja cauteloso ao interpretar (ou ao aceitar interpretações de outros sobre) a correlação quando não pode verificar as condições por si próprio.

TESTE RÁPIDO

O preço trimestral das ações das empresas de semicondutores Cypress e Intel tem uma correlação de 0,86 para os anos de 1992 a 2002.

1. Antes de tirar conclusões a partir da correlação, o que você gostaria de ver? Por quê?
2. Se o seu colega de trabalho coletar os mesmos preços em euros, como isso irá mudar a correlação? Você terá de saber a taxa de câmbio entre euros e dólares americanos para tirar conclusões?
3. Se você padronizar os preços das duas ações, como isso irá afetar a correlação?
4. Em geral, se, num dado dia, o preço da Intel for relativamente baixo, é provável que o preço da Cypress também seja relativamente baixo?
5. Se, num dado dia, o preço da ação da Intel for alto, o preço da ação da Cypress também será alto?

EXEMPLO ORIENTADO — *Gastos do consumidor*

Uma grande empresa de cartões de crédito envia um incentivo para os seus melhores clientes na esperança de que eles utilizem mais seus cartões. Ela quer saber com que frequência pode oferecer o incentivo. As ofertas frequentes irão resultar em aumentos frequentes no uso do cartão de crédito? Para examinar essa questão, um analista tomou uma amostra aleatória de 184 clientes do seu segmento de uso mais alto e investigou os débitos durantes dois meses em que os incentivos foram oferecidos.

PLANEJAR

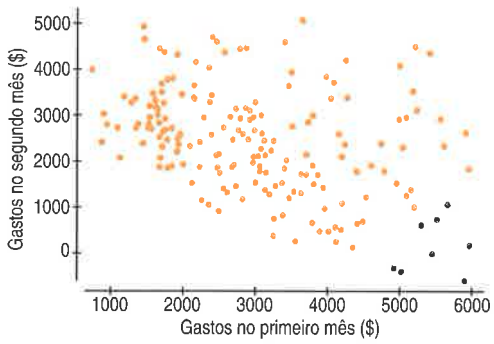
Configuração: Declare seu objetivo. Identifique as variáveis quantitativas a serem examinadas. Relate o espaço de tempo no qual os dados foram coletados e defina cada variável (utilize as cinco perguntas).

Nosso objetivo é investigar a associação entre a quantia que os clientes gastaram nos dois meses em que receberam o incentivo. Os clientes foram selecionados aleatoriamente dentre o segmento com a mais alta utilização do cartão. As variáveis mensuradas são o total de gastos nos cartões de crédito (em $) durante os dois meses de interesse.

✓ ***Condição da variável quantitativa:*** *Ambas as variáveis são quantitativas. Ambos os gastos são mensurados em dólares.*

Faça o diagrama de dispersão e rotule claramente os eixos para determinar a escala e as unidades.

Visto que temos duas variáveis quantitativas mensuradas na mesma situação, podemos fazer um diagrama de dispersão.



Verifique as condições.

✓ ***Condição de linearidade:*** *O diagrama de dispersão é linear o suficiente.*

✓ ***Condição do valor atípico:*** *Não existem valores atípicos óbvios.*

FAZER

Mecânica: Uma vez que as condições foram satisfeitas, calcule a correlação utilizando a tecnologia.

O valor da correlação é – 0,391

O coeficiente de correlação negativo confirma a impressão do diagrama de dispersão.

continua...

continuação

RELATAR	**Conclusão:** Descreva a direção, a forma e a força do gráfico, juntamente com qualquer ponto ou característica incomum. Tenha certeza de declarar sua interpretação num contexto apropriado.	MEMORANDO: Re: Gastos com o cartão de crédito Examinamos alguns dos dados do programa de incentivo. Em particular, analisamos os gastos efetuados nos dois primeiros meses do programa. Observamos que havia uma associação negativa entre os débitos no segundo mês e os débitos no primeiro mês. A correlação era de 0,391, que é apenas moderadamente forte e indica uma variação substancial. Concluímos que, embora o padrão observado seja negativo, esses dados não permitem encontrar a causa desse comportamento. É provável que alguns clientes tenham sido estimulados pela oferta a aumentar seus gastos no primeiro mês, mas depois retornaram ao antigo padrão de gastos. É possível que outros clientes não tenham mudado de comportamento até o segundo mês do programa, aumentando seus gastos a partir daí. Sem informações sobre o padrão de gastos dos clientes antes do programa de incentivo será difícil deduzir mais conclusões. Sugerimos uma pesquisa adicional e, além disso, que a próxima tentativa envolva um período maior de tempo, a fim de determinar se os padrões observados persistem.

Propriedades da correlação

Visto que a correlação é muito utilizada como uma medida de associação, é recomendável lembrar algumas das suas propriedades básicas. Eis uma lista útil de fatos sobre o coeficiente de correlação:

- **O sinal de um coeficiente de correlação fornece a direção da associação.**
- **A correlação é sempre um número entre –1 e +1.** A correlação *pode* ser exatamente igual a –1, 0 ou +1, mas cuidado. Esses valores são incomuns em dados reais, porque significam que todos os pontos dos dados caem *exatamente* sobre uma linha reta.
- **A correlação trata *x* e *y* simetricamente.** A correlação entre x e y é a mesma correlação entre y e x.
- **A correlação não tem unidades.** Esse fato é importante quando as unidades dos dados são um tanto vagas (satisfação do cliente, eficiência do trabalhador, produtividade e assim por diante).
- **A correlação não é afetada por mudanças no centro ou escala de ambas as variáveis.** Mudar as unidades ou a base das variáveis não afeta o coeficiente de correlação, porque a correlação depende somente dos escores-z.
- **A correlação mensura a força da associação *linear* entre duas variáveis.** As variáveis podem estar fortemente associadas, mas ainda ter uma pequena correlação se a associação não for linear.
- **A correlação é sensível às observações incomuns.** Um único valor atípico pode transformar uma correlação pequena em uma grande e vice-versa.

Quão forte é forte? Tenha cuidado ao usar os termos "fraco", "moderado" ou "forte", pois não há consenso sobre o que esses termos significam exatamente. Uma mesma correlação pode ser forte em um contexto e fraca noutro. É empolgante descobrir uma correlação de 0,7 entre um índice econômico e os preços de ações, mas uma correlação de "somente" 0,7 entre um tratamento com uma droga e a pressão alta seria percebida como um fracasso por uma empresa farmacêutica. Usar termos gerais como "fraco", "moderado" ou "forte" para descrever uma associação linear pode ser útil, mas tenha certeza de relatar a correlação e mostrar um diagrama de dispersão para que os outros possam fazer seus próprios julgamentos.

Tabelas de correlação

Às vezes, você verá as correlações entre cada par de variáveis de um conjunto de dados organizadas em uma tabela. As linhas e colunas da tabela nomeiam as variáveis e as células apresentam as correlações entre cada par de variáveis.

Tabela 7.1 Correlação para outras variáveis mensuradas mensalmente durante o período de 1995 a 2005. Preço final = preço da ação da Home Depot ao final de cada mês; Taxa de juros = taxa de juros preferencial e mais comum do banco; e Taxa de desemprego em percentual

	Preço final	Taxa de juros	Taxa de desemprego
Preço final	1,000		
Taxa de juros	–0,214	1,000	
Taxa de desemprego	–0,445	–0,679	1,000

As tabelas de correlação são compactas e fornecem muita informação resumida apenas num rápido olhar. Elas podem ser uma maneira eficiente de começar a analisar um grande conjunto de dados. As células diagonais da tabela de correlação sempre mostram correlações de exatamente 1,000, e a metade superior da tabela é simetricamente igual à metade inferior (você sabe por quê?), por isso, tradicionalmente apenas a metade inferior é exibida. Uma tabela como esta pode ser conveniente, mas certifique-se de que existem linearidade e observações atípicas, caso contrário, as correlações na tabela podem ser enganosas ou insignificantes. É possível ter certeza, observando a Tabela 7.1, de que as variáveis são linearmente associadas? As tabelas de correlação normalmente são produzidas por pacotes de *software* estatísticos. Felizmente, esses pacotes, em geral, oferecem maneiras simples de criar todos os diagramas de dispersão que você precisa observar.[4]

*7.4 Linearizando diagramas de dispersão

Depois do Índice Dow Jones, o S&P 500 é o índice mais observado das ações dos Estados Unidos. Desde sua introdução, em 1957, o índice S&P, composto por grandes empresas controladas publicamente, tem experimentado um período de crescimento extraordinário. Em 2 de janeiro de 1957, o S&P 500 manteve-se em 46,2 e alcançou um patamar de 1527,46 em 24 de março de 2000 (veja a Figura 7.6).

Caso você ouvisse que a correlação entre o *Tempo* e o *Índice S&P 500* é de 0,76, pode pensar que houve uma associação linear forte. No entanto, o diagrama de séries temporais dos dados mostra uma conclusão diferente. O crescimento foi relativamente modesto até meados de 1980, quando o índice começou crescer numa taxa mais rápida, atingindo seu pico em março de 2000. (É interessante ver, também, que o "*crash*" de 1987 agora aparece como um minúsculo desvio no crescimento geral.)

[4] Uma tabela de dispersão organizada igual à tabela de correlação às vezes é chamada de *diagrama de dispersão matricial*, ou DDM, e é facilmente criada usando um pacote estatístico.

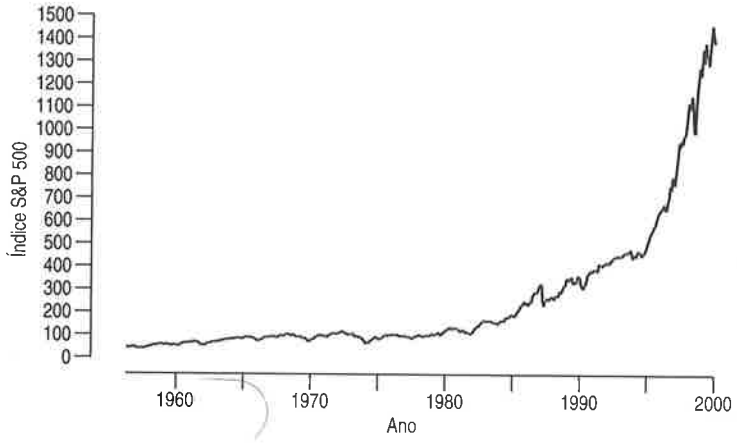


Figura 7.6 O diagrama de séries temporais do Índice S&P 500 mostra um relacionamento curvilíneo.

Lembre-se de que a correlação mensura somente a força de uma associação "linear". No diagrama de séries temporais do Índice S&P 500, está claro que o Índice não está aumentando linearmente. E se analisarmos o *logaritmo* do *S&P 500* ao longo do *Tempo*?

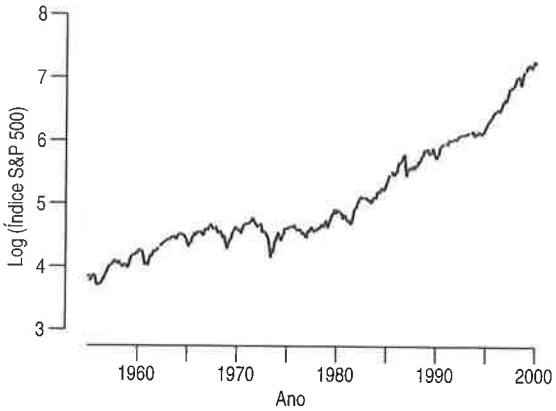


Figura 7.7 Transformar o S&P 500 com o logaritmo ajuda a linearizar o gráfico. Agora, a correlação é uma medida mais apropriada da associação.

Esse gráfico parece mais linear, portanto, a correlação agora é uma medida mais apropriada da associação. Outra vantagem desse gráfico é que os períodos de crescimentos diferentes estão claros. O que parecia ser um período de baixo crescimento no início dos anos 1960 agora é revelado como um período de crescimento normal – algo que estava escondido no gráfico original, devido ao crescimento enorme que ocorreu mais tarde. O período de 1970 a 1980, que sofreu desemprego e inflação altos, mostra pouco crescimento. Finalmente, o mercado pujante, que durou do início de 1980 até seu pico, em março de 2000, mostra um crescimento quase estável na escala logarítmica. O "*crash*" de 1987 parece ainda menos significativo quando visto no contexto desta extraordinária sequência de 20 anos. Índices como o S&P 500 geralmente são traçados numa escala logarítmica para facilitar a visualização do que está acontecendo.

Transformações simples como a do logaritmo, raiz quadrada ou recíproca podem, às vezes, linearizar a forma de um diagrama de dispersão. Os próximos capítulos irão discutir formas simples de reorganizar os dados.

7.5 Variáveis ocultas e causação

Um pesquisador da educação encontra uma forte associação entre a altura e a habilidade de leitura nos estudantes de escola primária em um levantamento de dados feito em todo o país. As crianças mais altas tendem a atingir escores mais altos em leitura. Isso significa que a altura dos estudantes é a *causa* dos seus escores altos? Independentemente de quão forte seja a correlação entre duas variáveis, não existe uma maneira simples de mostrar, a partir dos dados observados, que uma variável causa a outra. Uma correlação alta apenas aumenta a tentação de pensar e dizer que a variável *x causa* a variável *y*. Para ter certeza, vamos repetir.

Não importa quão forte seja a associação, não importa quão alto seja o valor de *r*, não importa quão linear a forma: não há maneira de concluir *exclusivamente* a partir de uma alta correlação que uma variável causa a outra. Sempre existe a possibilidade de que uma terceira variável – uma **variável oculta** – esteja afetando ambas as variáveis observadas. No exemplo do escore da leitura, talvez você já tenha adivinhado que a variável oculta é a idade da criança. As crianças mais velhas tendem a ser mais altas e ter maior habilidade de leitura. Entretanto, mesmo quando a variável oculta não for tão óbvia, resista à tentação de achar que uma correlação alta implica causa. Eis outro exemplo.

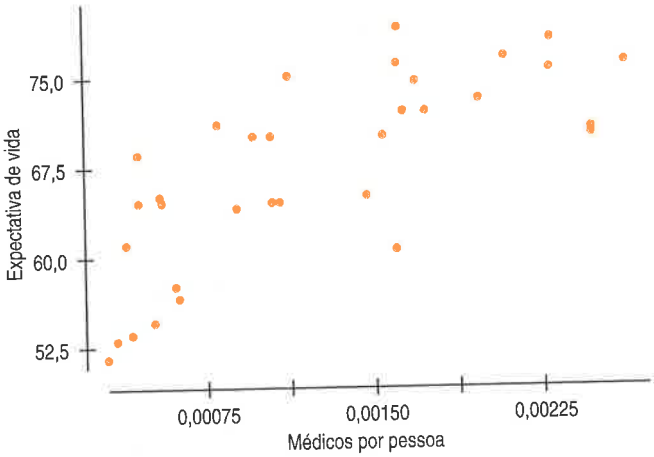


Figura 7.8 A expectativa de vida e número de médicos por pessoa em 34 países mostra um relacionamento linear forte e positivo, com uma correlação de 0,74.

O diagrama de dispersão mostra a *expectativa de vida* (média de homens e mulheres, em anos) para cada um dos 34 países do mundo *versus* o número de *médicos por pessoa* em cada país. A associação forte e positiva ($r = 0{,}74$) parece confirmar nossa expectativa de que mais *Médicos por Pessoa* melhoram os cuidados de saúde, refletindo em vidas mais longas e *Expectativa de Vida* maior. Talvez devêssemos enviar mais médicos a países em desenvolvimento para aumentar sua expectativa de vida.

Se aumentarmos o número de médicos, a expectativa de vida irá aumentar? Isto é, aumentar o número de médicos irá *causar* maior expectativa de vida? Poderia haver outra explicação para a associação? Veja outro diagrama de dispersão. A *Expectativa de Vida* ainda é a resposta, mas, desta vez, a variável previsora não é o número de médicos, mas o número de *Televisões por Pessoa* em cada país (veja Figura 7.9). A associação positiva neste diagrama de dispersão parece ainda mais *forte* que a associação do gráfico anterior. Se quisermos calcular a correlação, devemos linearizar o gráfico primeiro; no entanto, mesmo a partir deste gráfico, fica claro que as expectativas de vida altas estão associadas a mais televisões por pessoa. Devemos concluir que aumentar o número de televisores aumenta a expectativa de vida? Se sim, deveríamos

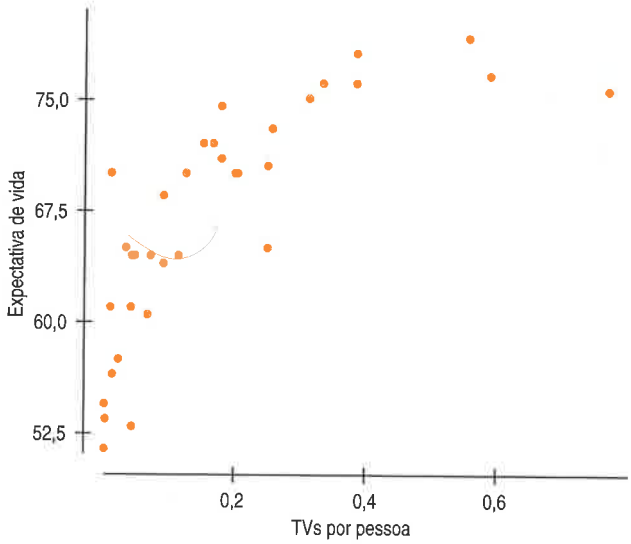


Figura 7.9 Expectativa de vida e número de televisores por pessoa mostra uma relação forte e positiva (embora claramente não linear).

enviar televisões, em vez de médicos, aos países em desenvolvimento. Essa associação com a expectativa de vida não somente é mais forte, mas televisores são mais baratos que médicos. O que está errado com este raciocínio? Talvez tenhamos nos apressado ao concluir que os médicos causam uma maior expectativa de vida. Talvez exista uma variável oculta aqui. Os países com altos padrões de vida têm expectativas de vida mais longas e mais médicos. Os altos padrões de vida poderiam causar mudanças nas outras variáveis? Se sim, então melhorar os padrões de vida poderia prolongar vidas, aumentar o número de médicos e aumentar o número de televisores. Deste exemplo, é possível perceber como é fácil cair numa armadilha de causalidade erroneamente inferida a partir de uma correlação. Pelo que sabemos, os médicos (ou televisores) *realmente* aumentam a expectativa de vida. Mas não podemos afirmar isso a partir de dados como esses, mesmo que quiséssemos. Resista à tentação de concluir, a partir de uma correlação, que x causa y, independentemente de quão óbvia essa conclusão lhe pareça.

O QUE PODE DAR ERRADO?

- **Não diga "correlação" quando você quer dizer "associação".** Quantas vezes você ouviu a palavra "correlação"? São grandes as chances de que em várias dessas situações o termo tenha sido usado equivocadamente. Trata-se de um dos termos estatísticos mais mal empregados – dada a quantidade de vezes que a estatística é mal empregada, isso significa muito. Um dos problemas é que muitas pessoas usam o termo *correlação* quando deveriam utilizar o termo mais geral *associação*. A associação é um termo deliberadamente vago, empregado para descrever a relação entre duas variáveis.

  A correlação é um termo preciso, usado para descrever a força e a direção de uma relação linear entre duas variáveis quantitativas.

- **Não correlacione variáveis categóricas.** Verifique a condição de variáveis quantitativas. Não faz sentido calcular a correlação para variáveis categóricas.

- **Tenha certeza de que a associação é linear.** Nem todas as associações entre variáveis quantitativas são lineares. A correlação pode deixar escapar até mesmo uma forte associação não linear. Uma empresa, preocupada com o fato de os consumidores usarem fornos com controle de temperatura imperfeitos, execu-

tou uma série de experimentos[5] para avaliar o efeito da temperatura na qualidade dos seus *brownies** congelados e desidratados. A empresa quer julgar a qualidade da sensibilidade dos *brownies* à variação das temperaturas do forno em torno da temperatura recomendada de 325 °F.** O laboratório relatou uma correlação de –0,05 entre os escores fornecidos por uma equipe de degustadores treinados e a temperatura, e relataram ao gerente que não existe uma relação. Antes de imprimir instruções na caixa informando aos clientes para não se preocupar com a temperatura, um estagiário perspicaz pediu para ver o diagrama de dispersão.

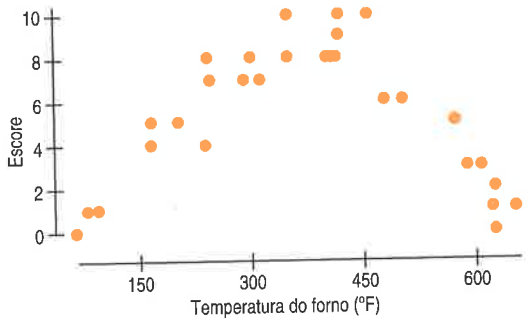


Figura 7.10 A relação entre o escore do sabor do *brownie* e a temperatura do forno é forte, mas não linear.

O gráfico, de fato, mostra uma associação forte – mas não linear. Não esqueça de verificar a Condição de Linearidade.

- **Tenha cuidado com os valores atípicos.** Você não pode interpretar um coeficiente de correlação com segurança sem verificar as observações atípicas. Veja um exemplo. A relação entre o QI e o tamanho do sapato de comediantes mostra uma correlação surpreendentemente forte e positiva de 0,50. Para verificar as suposições, analisamos o diagrama de dispersão.

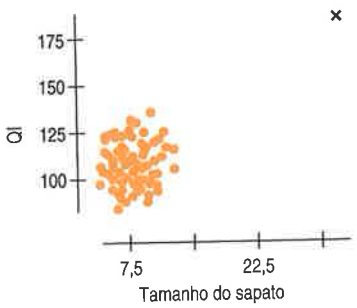


Figura 7.11 QI vs. tamanho do sapato.

- Deste "estudo", o que podemos afirmar sobre a relação entre as duas variáveis? A correlação é de 0,50. Mas a quem *realmente* pertence aquele ponto no lado superior direito? O valor atípico é *Bozo*, o Palhaço, famoso por seus sapatos grandes e amplamente reconhecido como um comediante "genial". Sem Bozo, a correlação está próxima de zero.

(continua)

[5] Experimentos projetados para avaliar o impacto de variáveis ambientais fora do controle da empresa na qualidade dos produtos foram defendidos pelo especialista japonês em qualidade Dr. Genichi Taguchi, a partir dos anos 1980, nos Estados Unidos.

* N. de T.: Bolo ou biscoito de chocolate recheado com nozes.

** N. de T.: Para obter a temperatura em graus celsius, basta usar a transformação: $^{0}C = \frac{5}{9}$ ($^{o}F - 32$).

Mesmo uma única observação incomum pode dominar o valor da correlação. É por esse motivo que você precisa verificar a Condição de Valores Atípicos.

- **Não confunda correlação com causa.** É grande a tentação de explicar uma correlação forte afirmando que a variável previsora *causou* a resposta à mudança. Nós humanos somos assim; temos a tendência de ver causas e efeitos em tudo. Só porque duas variáveis estão relacionadas não significa que uma *causa* a outra.

> **O câncer causa o fumo?** Mesmo que a correlação entre duas variáveis se deva a um relacionamento causal, a própria correlação não pode indicar o que causa o quê.
>
> Sir Ronald Aylmer Fisher (1890 – 1962) foi um dos maiores estatísticos do século XX. Fisher testemunhou perante um tribunal, pago pelas empresas de cigarro, que um relacionamento causal pode ter por base a correlação entre o fumo e o câncer.
>
> "É possível, então, que o câncer de pulmão ... seja uma das causas de fumar cigarros? Eu não acho que possa ser excluída ... a condição pré-câncer envolve leve quantidade de inflamação crônica ...
>
> Uma pequena irritação ... é comumente acompanhada pelo ato de acender um cigarro, compensando, assim, uma das pequenas mazelas da vida. E ... não é improvável que seja associada a fumar com mais frequência."
>
> Ironicamente, a prova de que o fumo é a causa de muitos cânceres veio de experimentos conduzidos seguindo os princípios do projeto experimental e da análise que Fisher desenvolveu.

Os diagramas de dispersão e os coeficientes de correlação *nunca* provam causa. Em parte, essa é a razão pela qual, por exemplo, os Estados Unidos tenham levado tanto tempo para colocar advertências de saúde nos cigarros. Embora houvesse muitas evidências de que o aumento do fumo estava *associado* a elevados níveis de câncer no pulmão, foram necessários anos de pesquisa para obter evidências de que o fumo realmente *causa* câncer. (As empresas de tabaco utilizaram esse fato a seu favor.)

- **Cuidado com as variáveis ocultas.** Um diagrama de dispersão dos danos (em dólares) causados a uma casa pelo fogo mostraria uma forte correlação com o número de bombeiros na cena. É óbvio que os danos não causam bombeiros. E os bombeiros realmente causam danos, borrifando água por tudo e cavando buracos, mas isso significa que não devemos chamar o corpo de bombeiros? É claro que não. Existe uma variável subjacente que leva a mais danos e mais bombeiros – o tamanho do incêndio. Uma variável que está escondida atrás da relação e a determina por simultaneamente afetar as outras duas variáveis é chamada de **variável oculta**. Em geral, você pode menosprezar reclamações feitas sobre os dados encontrando uma variável oculta.

ÉTICA EM AÇÃO

Uma agência de publicidade contratada por um conhecido fabricante de produtos de higiene dental (escovas de dente elétricas, irrigadores orais, etc.) juntou uma equipe criativa a fim de desenvolver ideias para uma nova campanha de publicidade. Trisha Simes foi escolhida para liderar o grupo, porque já havia trabalhado com o cliente anteriormente. Na primeira reunião, Trisha comunicou à equipe o desejo do cliente de se diferenciar dos seus concorrentes, mas sem focar sua mensagem nos benefícios estéticos de um bom cuidado dentário. Enquanto buscavam ideias, um membro da equipe, Brad Jonns, lembrou de uma recente transmissão da CNN que relatava a "correlação" entre o uso do fio dental nos dentes e a redução do risco de doença cardíaca. Percebendo o potencial de promover os benefícios à saúde de um cuidado dental adequado, a equipe concordou em investir na ideia. Na reunião seguinte, vários membros da equipe comentaram a sua surpresa em ter descoberto diversos artigos médicos, científicos e leigos que afirmavam que a boa higiene dental resultava em boa saúde. Um dos membros relatou que havia encontrado artigos que ligavam doenças na gengiva não somente a ataques cardíacos e derrames, mas à diabetes e até mesmo ao câncer. Enquanto Trisha tentava decifrar por que os concorrentes não tinham tirado proveito dessas pesquisas, sua equipe, animada, já havia começado a focar e projetar a campanha ao redor dessa mensagem.

PROBLEMA ÉTICO *A correlação não implica causa. A possibilidade de variáveis ocultas não é explorada. Por exemplo, é provável que quem tem mais cuidado consigo tenha menos risco de doenças cardíacas e também use o fio dental regularmente (relacionado ao Item C da ASA Ethical Guidelines).*

SOLUÇÃO ÉTICA *Eviete deduzir causa e efeito a partir de resultados da correlação.*

O que aprendemos?

Nos capítulos anteriores, aprendemos a ouvir a história contada pelos dados de uma única variável. Neste capítulo, voltamos nossa atenção à história mais complicada (e mais interessante) que podemos descobrir a partir da associação entre duas variáveis quantitativas.

Aprendemos a começar nossa investigação observando um diagrama de dispersão. Estamos interessados na *direção* da associação, na *forma* que ela toma e na sua *força*.

Aprendemos que, embora nem todo o relacionamento seja linear, quando o diagrama de dispersão for linear o suficiente, o *coeficiente de correlação* é um resumo numérico útil.

- O sinal da correlação indica a direção da associação.
- A magnitude da correlação revela a *força* da associação linear. Associações fortes têm correlações próximas a +1 ou −1 e associações muito fracas têm correlações próximas a zero.
- A correlação não tem unidades, portanto, deslocar ou mudar a escala dos dados, padronizar ou até mesmo trocar as variáveis entre si não tem efeito no valor do coeficiente de correlação.

Aprendemos que, para usar a correlação, é preciso verificar certas condições a fim de que a análise seja válida.

- Antes de encontrar ou falar sobre a correlação, sempre devemos verificar a Condição de Linearidade.
- E, como sempre, devemos tomar cuidado com as observações incomuns!

Finalmente, aprendemos a não cometer o erro de assumir que uma correlação alta ou uma associação forte é indício de uma relação de causa e efeito. Devemos tomar cuidado com as variáveis ocultas!

Termos

Associação

- **Direção:** Uma direção ou associação positiva significa que, em geral, à medida que uma variável aumenta, a outra também aumenta. Quando o crescimento de uma variável geralmente corresponde à diminuição da outra, a associação é negativa.
- **Forma:** A forma que mais nos interessa é a linear, mas você deve descrever outros padrões que percebe nos diagramas de dispersão.
- **Força:** Um diagrama de dispersão exibirá uma forte associação se existir pouca dispersão em torno da relação subjacente.

Coeficiente de correlação Medida numérica da direção e força de uma associação linear.

$$r = \frac{\sum z_x z_y}{n-1}$$

Diagrama de dispersão Gráfico que mostra a relação entre duas variáveis quantitativas mensuradas nos mesmos casos.

Valor atípico Ponto que não se encaixa no padrão geral visto no diagrama de dispersão.

Variável explanatória ou independente (variável *x*) Variável que determina, explica, prevê ou é de alguma forma responsável pela variável *y*.

Variável oculta Variável diferente de *x* e *y* que simultaneamente afeta ambas as variáveis, sendo responsável pela correlação entre as duas.

Variável resposta ou dependente (variável y) Variável que o diagrama de dispersão deve explicar ou prever.

Habilidades

PLANEJAR

- Reconhecer quando o interesse no padrão de uma possível relação entre duas variáveis quantitativas sugere um diagrama de dispersão.
- Ser capaz de identificar os papéis e colocar a variável resposta no eixo *y* e a variável explanatória no eixo *x*.
- Saber as condições para a existência da correlação e como verificá-las.
- Saber que as correlações estão entre –1 e +1 e que cada extremo indica uma associação linear perfeita.
- Entender de que maneira a magnitude da correlação reflete a força de uma associação linear como é vista num diagrama de dispersão.
- Saber que a correlação não tem unidades.
- Saber que o coeficiente de correlação não muda com a mudança do centro ou da escala de uma ou ambas as variáveis.
- Entender que causalidade não pode ser demonstrada por um diagrama de dispersão ou pela correlação.

FAZER

- Ser capaz de fazer um diagrama de dispersão manualmente (para um pequeno conjunto de dados) ou com tecnologia.
- Saber como calcular a correlação entre duas variáveis.
- Saber como ler uma tabela de correlações produzida por um programa estatístico.

RELATAR

- Ser capaz de descrever a direção, a forma e a força de um diagrama de dispersão.
- Estar preparado para identificar e descrever pontos que se desviam do padrão geral.
- Ser capaz de usar a correlação como parte da descrição de um diagrama de dispersão.

- Estar alerta a interpretações errôneas da correlação.
- Entender que encontrar uma correlação entre duas variáveis não indica uma relação causal entre elas. Ter cuidado acerca dos perigos de sugerir relacionamentos causais na descrição de correlações.

Ajuda tecnológica: diagramas de dispersão e correlação

Os pacotes estatísticos, em geral, ajudam a verificar se a correlação em um diagrama de dispersão é apropriada. Alguns pacotes facilitam o trabalho mais que outros.

Muitos pacotes permitem modificar ou melhorar um diagrama de dispersão, alterando as legendas dos eixos, a numeração dos eixos, os símbolos do gráfico ou as cores usadas. Algumas opções, como a escolha das cores e dos símbolos, podem ser usadas para exibir informações adicionais no diagrama de dispersão.

Excel

Para fazer um diagrama de dispersão com o "Assistente de Gráfico" do Excel:

- Clique no ícone **Assistente de Gráfico** na barra do menu. O Excel abre a caixa de diálogo "Assistente de gráfico".
- Verifique se o painel **Tipos Padrão** está selecionado e escolha **Dispersão (XY)** entre as opções oferecidas.
- Escolha o primeiro modelo (Dispersão – compara pares de valores) das cinco opções disponíveis e clique em **Avançar**.
- Se ele já não estiver selecionado, clique no painel **Intervalo de Dados** e entre com o intervalo dos dados no espaço disponível.
- Por convenção, sempre representamos as variáveis em colunas. O **Assistente de Gráfico** se refere às variáveis como **Séries**. Verifique se a opção **Coluna** está selecionada.
- O Excel insere a coluna da esquerda daquelas que você selecionou no eixo *x* do diagrama de dispersão. Se a coluna que você deseja ver no eixo *x* não é a coluna da esquerda na sua planilha, clique no painel **Séries** e edite a especificação dos eixos individualmente.
- Clique na tecla **Avançar.** A caixa de diálogo **Opções de Gráficos** (verifique na faixa azul superior da janela) aparece.
- Selecione o painel **Título,** se ele já não estiver ativo. Aqui, você especifica o título do diagrama e os nomes das variáveis que serão exibidas nos eixos.
- Digite o título do diagrama na caixa de edição **Título do Gráfico**.
- Digite o nome da variável *x* na caixa de edição **Eixo dos Valores (X)**. Observe que você deve nomear as colunas corretamente. Nomear outra variável não irá alterar o gráfico, apenas irá apresentar a legenda errada.
- Digite o nome da variável *y* na caixa de edição **Eixo dos Valores (Y)**.
- Clique na tecla **Avançar** para abrir a caixa de diálogo do local do gráfico.
- Selecione a opção de **Como Nova Planilha**.
- Clique no botão **Concluir.**

Geralmente, o diagrama de dispersão irá requerer uma mudança de escala. Por padrão, o Excel inclui a origem no diagrama mesmo quando os dados estão longe do zero. Você pode ajustar as escalas dos eixos. Para mudar a escala do eixo de um diagrama no Excel:

- Clique duas vezes no eixo. Uma caixa de diálogo **Formatar Eixo** aparece.
- Se o painel **Escala** não estiver ativo, selecione-o.
- Entre com os novos valores para o mínimo ou o máximo nos espaços fornecidos. Você pode arrastar a caixa de diálogo sobre o diagrama de dispersão para servir como um esquadro e ajudá-lo na leitura dos valores máximos e mínimos dos eixos.
- Clique no botão **OK** para ver o diagrama de dispersão na nova escala.
- Siga os mesmos passos para alterar a escala do outro eixo.

Calcule a correlação no Excel com a função **CORREL** do menu suspenso do procedimento **Inserir Função**. Escolha a categoria "Estatística" e em seguida procure "**CORREL**" na lista do painel que abrir.

Na caixa de diálogo que aparece, entre com o intervalo das células, colocando uma variável em cada espaço fornecido. Não se preocupe com a ordem, pois ela é indiferente para o cálculo do coeficiente de correlação.

Para fazer um diagrama de dispersão usando o suplemento DDXL, selecione as duas variáveis a serem exibidas. Elas devem estar no formato colunas. Se a primeira linha tiver o título das colunas (nome da variável), inclua-a. Do menu do DDXL, escolha **Charts and Plots.** Do menu da caixa de diálogo da função, escolha **Scatterplot**. Arraste a variável *x* para área **X-Axis Variable** e a variável *y* para a área **Y-Axis Variable.** Se você tiver uma coluna que nomeie cada caso, arraste-a para a área **Label Variable**. Clique em **OK.**

Excel 2007

Para fazer um diagrama de dispersão no Excel 2007:

- Selecione a coluna dos dados para usar no diagrama de dispersão. Você pode selecionar mais que uma coluna segurando a tecla CTRL (*Control*) enquanto clica.
- No item de menu (painel) **Inserir**, clique no ícone **Dispersão** e selecione o gráfico **Dispersão Somente com Marcadores** das opções apresentadas (é o diagrama localizado no canto superior esquerdo).

Para tornar o gráfico mais útil para a análise dos dados, ajuste a apresentação da seguinte forma:

- Selecione o gráfico e clique no painel **Layout** da aba **Ferramentas de Gráfico**. Em seguida, clique no ícone **Linhas de Grade** e escolha a opção **Linhas de Grade Horizontais Principais** entre as duas que aparecem.
- Em **Linhas de Grade Horizontais Principais,** selecione **Nenhuma.** Isso irá remover as linhas de grade do diagrama de dispersão. Você também pode explorar o item **Mais Opções** do **Linhas de Grade.**
- Para mudar a escala dos eixos, clique sobre os números de cada eixo do gráfico e, em seguida, clique em **Formatar Seleção** no painel ***Layout,*** da aba **Ferramentas de Gráfico**.
- Selecione a opção **Fixo,** em vez da opção Automático e digite os valores Mínimo e Máximo que você julga mais adequados para o diagrama de dispersão. Você pode obter esse mesmo menu após clicar sobre os eixos com o botão direito do mouse, e então escolher a opção **Formatar Eixo** do menu flutuante que aparece.

O Excel 2007 automaticamente coloca a coluna da esquerda das duas colunas que você seleciona no eixo *x* e a coluna da direita no eixo *y*. Você pode trocá-las, se quiser.
Para trocar as variáveis *X* e *Y*:

- Clique no gráfico para ter acesso à aba **Ferramentas de Gráfico**.
- Clique em **Selecionar Dados** do painel **Design**.
- Em **Entradas de Legenda (Série)**, clique em **Editar.**
- Marque e delete tudo na caixa de entrada **Valores de X da Série** e selecione novos dados da planilha (tenha cuidado ao selecionar a coluna para não selecionar inadvertidamente o título da coluna, o que não funcionaria aqui).
- Faça o mesmo com a caixa de entrada **Valores de Y da Série**.
- Pressione **OK** para fechar a caixa de diálogo **Editar**, depois pressione **OK** novamente para fechar a caixa de diálogo **Fonte de Dados.**

Para fazer um diagrama de dispersão usando o suplemento DDXL, selecione as duas variáveis a serem exibidas. Elas devem ser colunas. Se a primeira linha tem nome das colunas, inclua-a. Do menu do DDXL, escolha **Charts and Plots.** Do menu das funções do diálogo, escolha **Scatterplot.** Arraste a variável *x* na área **X-Axis Variable** e a variável *y* na área **Y-Axis Variable.** Se você tiver uma coluna que nomeie cada caso, arraste-a para a área **Label Variable**. Clique em **OK.**

Minitab

Para fazer um diagrama de dispersão, escolha **Scatterplot** do menu **Graph.** Escolha "**Simple**" para o tipo de gráfico. Clique em **OK.** Entre com os nomes das variáveis para a variável Y e a variável X na tabela. Clique em **OK.**

Para calcular o coeficiente de correlação, escolha **Basics Statistics** do menu **Stat.** Do submenu **Basic Statistics**, escolha **Correlation.** Especifique os nomes de pelo menos duas variáveis quantitativas na caixa "**Variables**". Clique em **OK** para calcular a tabela da correlação.

SPSS

Para fazer um diagrama de dispersão no SPSS, clique no item de menu **Graphs**. Depois:

- Clique no submenu **Scatter/Dot...**.
- Escolha **Simple Scatter** dos tipos de gráficos e clique em **Define**.
- Transfira a variável que você quer como variável resposta para o espaço denominado **Y Axis**.
- Transfira a variável que você quer como variável previsora para o espaço denominado **X Axis**.
- Clique em **OK.**

Para calcular o coeficiente de correlação, escolha **Correlate** do menu **Analyze.** Do submenu **Correlate**, escolha **Bivariate.** Na caixa de diálogo **Bivariate Correlations**, use a tecla da seta para mover as variáveis para o painel denominado **Variables**. Certifique-se de que a opção **Pearson** está selecionada no campo **Correlations Coeficients**. Clique em **OK**.

JMP

Para criar um diagrama de dispersão e calcular a correlação, escolha **Fit Y by X** do menu **Analyze.** Na caixa de diálogo Fit Y by X, arraste a variável Y para a caixa "**Y, Response**" e arraste a variável X para a caixa "**X, Factor**". Clique em **OK.** Uma vez que o JMP fez o diagrama de dispersão, clique no triângulo vermelho próximo ao título do gráfico para abrir um menu de opções. Selecione **Density Ellipse** e selecione 0,95. O JMP desenha uma elipse em torno dos dados e revê o painel **Correlations.** Clique no triângulo azul próximo a **Correlation** para mostrar uma tabela contendo o coeficiente de correlação.

Data Desk

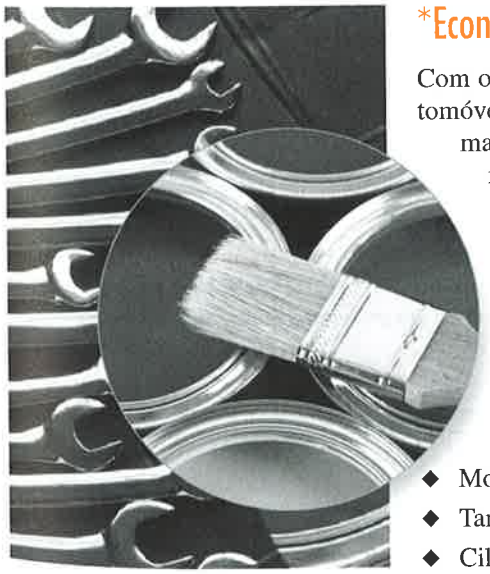
Para fazer um diagrama de dispersão de duas variáveis, selecione uma variável como Y e a outra como X e escolha **Scatterplot** do menu **Plot.** Depois, encontre o coeficiente de correlação selecionando **Correlation** do menu do diagrama de dispersão **HyperView**.

Alternativamente, selecione as duas variáveis e escolha **Pearson Product-Moment** do submenu **Correlations** do menu **Calc.**

Comentários

Preferimos que você observe primeiro o diagrama de dispersão e depois encontre a correlação. Mas se você encontrou a correlação primeiro, clique no valor da correlação para abrir um menu suspenso, que oferece uma opção para criar um diagrama de dispersão.

Projetos de estudo de pequenos casos



*Economia de combustível

Com o aumento constante no preço da gasolina, os motoristas e as fábricas de automóveis estão motivados a diminuir o consumo de combustível dos carros. Informações recentes fornecidas pelo governo dos Estados Unidos propõem algumas maneiras simples de fazer isso (veja www.fueleconomy.gov): evitar a aceleração rápida, evitar dirigir acima de 100 Km/h, reduzir o tempo que o carro fica em marcha lenta e reduzir o peso do veículo. Um peso extra de 100 libras pode aumentar o consumo de combustível em mais de 2%. Um executivo de *marketing* está estudando a relação entre o consumo de combustível dos carros (mensurado em milhas por galão) e seu peso, a fim de projetar uma campanha para um novo carro compacto. No conjunto de dados **ch07_MCSP_Fuel_Efficiency** você irá encontrar os dados das variáveis abaixo.

- Modelo do carro
- Tamanho do motor (L)
- Cilindros
- PSPF (Preço Sugerido Pelo Fabricante em $)
- Consumo na cidade (mpg)
- Consumo na autoestrada (mpg)
- Peso (libras)
- Tipo e país do fabricante

Descreva a relação entre Peso, PSPF e Tamanho do Motor com a eficiência do combustível (na cidade e autoestrada) em um relatório escrito. Certifique-se de transformar as variáveis, se necessário.

A economia dos Estados Unidos e os preços das ações da Home Depot

O arquivo **ch07_MCSP_Home_Depor** contém variáveis econômicas e dados do mercado de ações para a Home Depot, Inc. Os economistas, investidores e executivos de corporações usam medidas econômicas norte-americanas para avaliar o impacto das pressões inflacionárias e das flutuações da taxa de desemprego no mercado de ações. A inflação normalmente é acompanhada por intermédio das taxas de juros. Embora existam inúmeros tipos de taxa de juros, aqui incluimos os valores mensais da taxa de empréstimo principal dos bancos, em que a taxa é anunciada pela maioria dos 25 principais (baseado em avaliações) bancos comerciais segurados dos Estados Unidos. A taxa principal de juros geralmente é usada por bancos para avaliar empréstimos de negócios a curto prazo. Além disso, fornecemos as taxas de juros de seis meses para os Certificados de Depósitos (CDs), as taxas de desemprego (ajustadas sazonalmente) e a taxa das letras do Tesouro. Investigue a relação entre o *Preço de Fechamento* (*Closing Price*) para as ações da Home Depot e as seguintes variáveis de 2006 a 2008:[6]

- Taxa de Desemprego (%)
- Taxa Preferencial de Juros Bancários (Taxa de Juros em %)
- Taxa dos CDs (%)
- Taxa das letras do Tesouro (%)

Descreva a relação de cada uma dessas variáveis com o *Preço de Fechamento* (*Closing Price*) da Home Depot num relatório escrito. Certifique-se de usar os diagramas de dispersão e as tabelas de correlação na sua análise e de transformar as variáveis, se necessário.

EXERCÍCIOS

1. Associação. Suponha que você deva coletar dados para cada par de variáveis abaixo. Você quer fazer um diagrama de dispersão. Qual variável você usaria como variável explanatória e qual como variável resposta? Por quê? O que você esperaria ver no diagrama de dispersão? Discuta sua provável direção e forma.

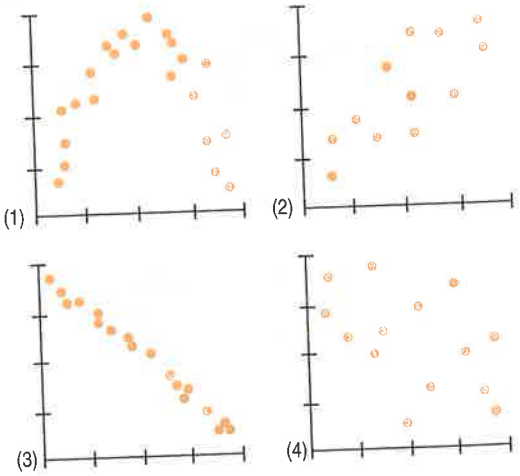
a) Contas de telefone celular: número de mensagens, custo.
b) Automóveis: consumo de combustível (mpg), volume das vendas (número de carros).
c) Para cada semana: vendas de casquinhas de sorvetes, vendas de condicionadores de ar.
d) Produto: Preço ($), demanda (número vendido por dia).

2. Associação, parte 2. Suponha que você deva coletar dados para cada par de variáveis. Você quer fazer um diagrama de dispersão. Qual variável você usaria como a variável explanatória e qual como a variável resposta? Por quê? O que você esperaria ver no diagrama de dispersão? Discuta a direção e forma provável de cada diagrama.

a) Camisetas numa loja: preço unitário, número vendido.
b) Imóveis: preço da casa, tamanho da casa (metros quadrados).
c) Economia: taxas de juros, número de requisições de empréstimos para compra da casa própria.
d) Empregados: salário, anos de experiência.

3. Diagramas de dispersão. Qual dos diagramas de dispersão exibe:

a) Pouca ou nenhuma associação?
b) Uma associação negativa?
c) Uma associação linear?

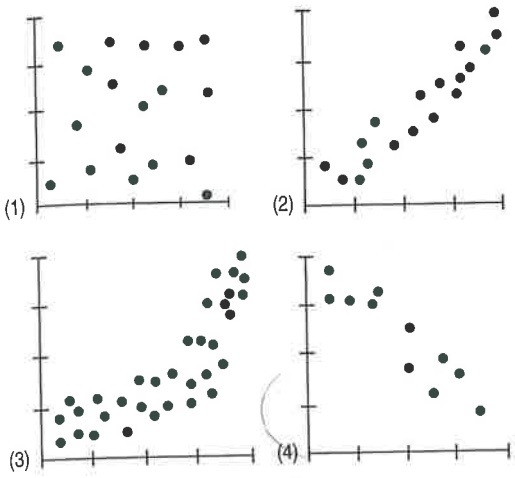


d) Uma associação moderadamente forte?
e) Uma associação muito forte?

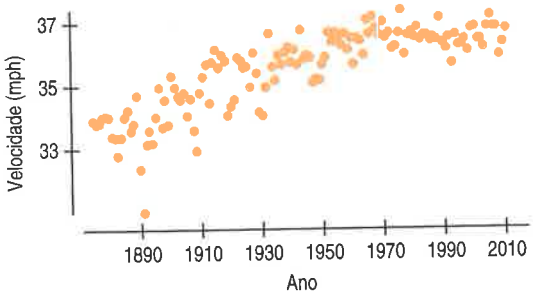
4. Diagramas de dispersão, parte 2. Qual dos diagramas de dispersão da próxima página exibe:

a) Pouca ou nenhuma associação?
b) Uma associação negativa?
c) Uma associação linear?
d) Uma associação moderadamente forte?
e) Uma associação muito forte?

[6] Fontes: Taxa de desemprego – Agência de Estatísticas do Trabalho Americana. Veja página do desemprego em www.bls.gov/cps/home.htm#data. Taxa de juros – Banco Central Americano (Federal Reserve). Veja www.federalreserve.gov/releases/H15/update/. Preços das ações da Home Depot no *site* HD/Investor Relations. Veja ir.homedepot.com/quote.cfm.

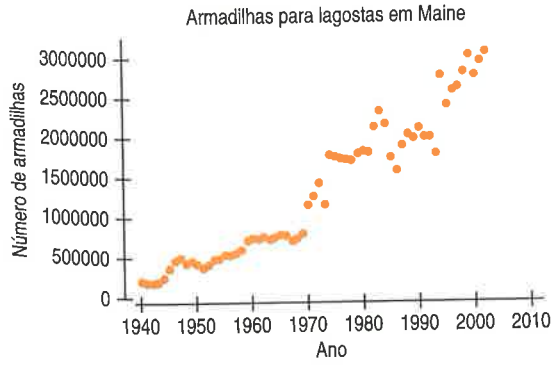


T **5. Kentucky Derby, 2007.** O cavalo mais rápido da história do Kentucky Derby foi Secretariat, em 1973. O diagrama de dispersão mostra as velocidades (em milhas por hora) dos cavalos vencedores em cada ano.



O que você vê? Na maioria dos eventos esportivos, os desempenhos têm melhorado e continuam a melhorar, portanto, antecipamos uma direção positiva. Mas e a forma? O desempenho aumentou na mesma taxa nos últimos 125 anos?

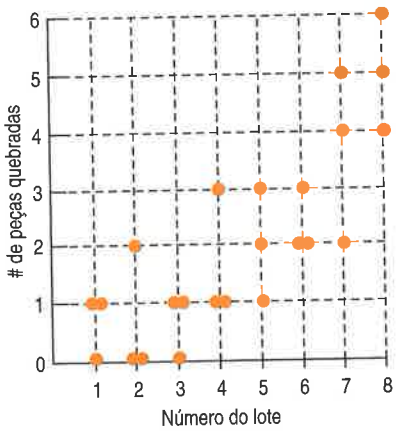
T **6. Armadilhas para lagosta.** Em muitas partes do mundo, a pesca da lagosta é um grande negócio. O gráfico mostra o crescimento do número de armadilhas para lagostas (legais) no estado de Maine, Estados Unidos, desde 1940.



a) O que você vê? Embora esperássemos uma tendência positiva, o que você pode dizer sobre a forma?

b) Você percebe o impacto da introdução das armadilhas de telas e de equipamentos eletrônicos nos barcos de lagosta no início dos anos 70? Qual efeito, se algum, isso pareceu ter?

7. Produção. Uma fábrica de cerâmica pode queimar oito lotes grandes de cerâmica por dia. Às vezes, algumas peças quebram durante o processo. Para entender melhor o problema, a fábrica registra o número das peças quebradas em cada lote ao longo de três dias e cria o diagrama de dispersão mostrado a seguir.

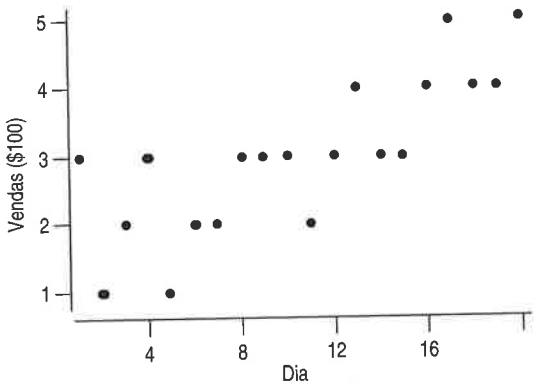


a) Faça um histograma mostrando a distribuição do número de peças quebradas nos 24 lotes de cerâmica examinados.

b) Descreva a distribuição mostrada no histograma. Que característica do problema é mais aparente no histograma do que no diagrama de dispersão?

c) Que aspecto do problema da fábrica é mais aparente no diagrama de dispersão?

8. Vendas de café. Os proprietários de uma nova cafeteria acompanharam as vendas dos primeiros 20 dias e exibiram os dados num diagrama de dispersão (por dia)

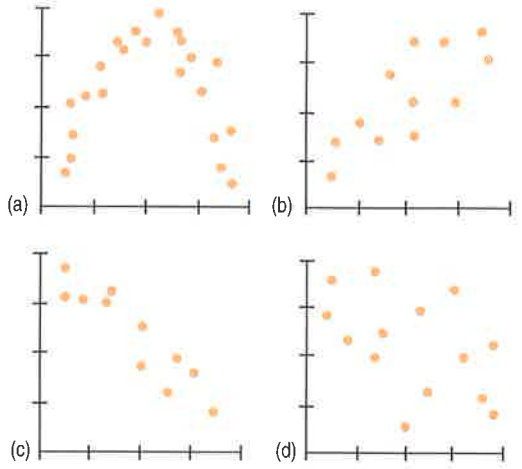


a) Faça um histograma das vendas diárias desde que a loja iniciou suas atividades comerciais.

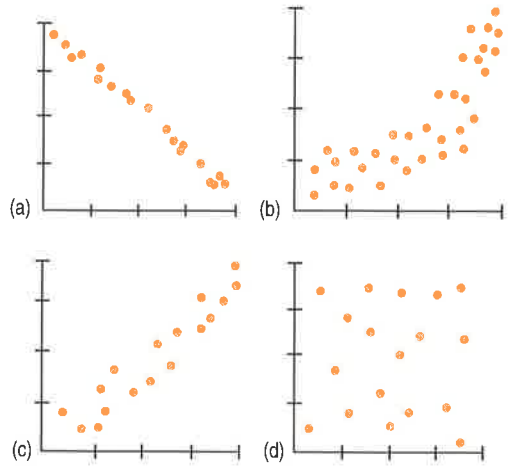
b) Apresente um fato que seja óbvio no diagrama de dispersão, mas não no histograma.

c) Apresente um fato que seja óbvio no histograma, mas não no diagrama de dispersão.

9. Associando. Veja vários diagramas de dispersão. As correlações calculadas são de –0,923, -0,487, 0,006 e 0,777. Atribua a cada diagrama uma das correlações.



10. Associando, parte 2. Eis vários diagramas de dispersão. As correlações calculadas são de -0,977, –0,021, 0,736 e 0,951. Associe uma correlação a cada um dos diagramas.



11. Empacotamento. Um CEO anuncia na reunião anual dos acionistas que a nova embalagem transparente para o produto mais importante da empresa foi um sucesso. Ele afirma: "Existe uma forte correlação entre a embalagem e as vendas". Critique essa afirmação do ponto de vista estatístico.

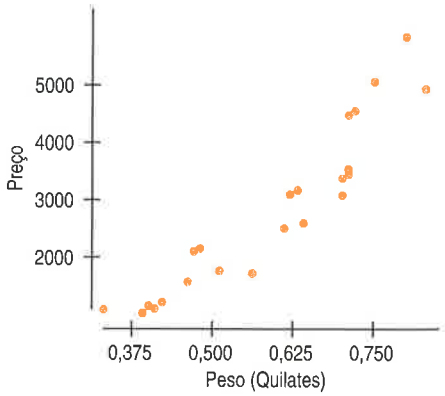
12. Seguros. As companhias de seguros mantém históricos de reclamações para que possam avaliar riscos e determinar as taxas de forma apropriada. O National Insurance Crime Bureau relata que os Honda Accords, Honda Civics e Toyota Camrys são os carros mais roubados, enquanto os Ford Tauruses, Pontiac Vibes e Buick LeSabres são os menos roubados. É razoável afirmar que existe uma correlação entre o tipo de carro que você tem e o risco de ele ser roubado?

13. Vendas de livros. Um analista considera a correlação entre a venda de livros (número de livros) numa livraria de uma faculdade e o dia do ano (1 = 1° de janeiro, ..., 365 = 31 de dezembro). O que você pode esperar da correlação entre as *Vendas de Livros* e o *Número do Dia*? Você acha que existe uma associação entre essas variáveis? Explique.

14. Vendas pela Internet. Um artigo em uma revista de negócios relatou que o comércio pela Internet explodiu recentemente, praticamente dobrando a cada três anos. Ele declarou que existia uma alta correlação entre as vendas feitas pela Internet e o *Ano*. Esse é um resumo apropriado? Explique.

T 15. Preços dos diamantes. O preço de um diamante depende da sua cor, corte, limpidez e peso em quilates. Aqui estão dados de um vendedor de diamantes de qualidade (portanto, presumimos bons cortes) para diamantes de melhores cores (D) e alta limpidez (VS1).

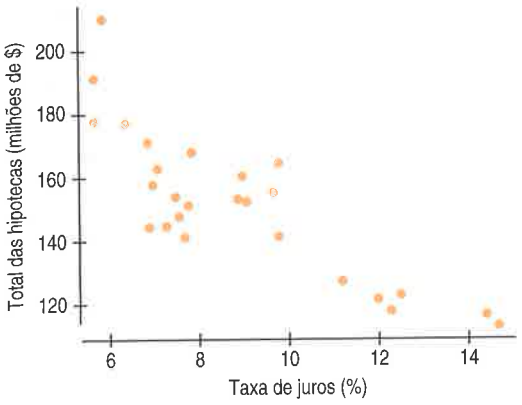
Quilate	Preço	Quilate	Preço
0,33	1079	0,62	3116
0,33	1079	0,63	3165
0,39	1030	0,64	2600
0,40	1150	0,70	3080
0,41	1110	0,70	3390
0,42	1210	0,71	3440
0,42	1210	0,71	3530
0,46	1570	0,71	4481
0,47	2113	0,72	4562
0,48	2147	0,75	5069
0,51	1770	0,80	5847
0,56	1720	0,83	4930
0,61	2500		



a) As suposições e condições estão satisfeitas para determinar a correlação?

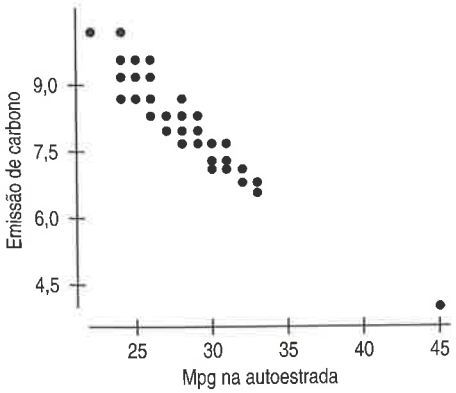
b) O coeficiente de correlação é 0,937. Usando essa informação, descreva a relação.

T 16. Taxas de juros e empréstimos. Desde 1980, a média dos juros dos empréstimos tem flutuado de um valor tão baixo quanto 6% a um tão alto quanto 14%. Existe um relacionamento entre a quantidade de dinheiro emprestado e a taxa de juros oferecida? Aqui está um diagrama de dispersão do *Total de Hipotecas* nos Estados Unidos (em milhões de dólares de 2005) versus *Taxa de Juros* em vários períodos nos últimos 26 anos. A correlação é de –0,84.



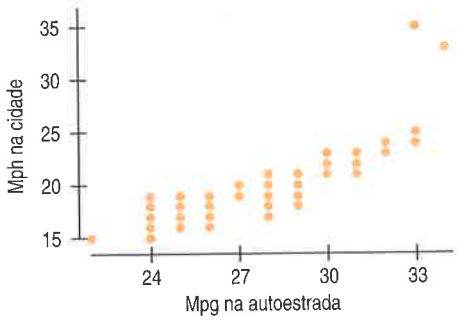
a) Descreva a relação entre *Total de Hipotecas* e *Taxas de Juros*.
b) Se padronizarmos as duas variáveis, qual seria o coeficiente de correlação entre as variáveis padronizadas?
c) Se mensurássemos o *Total de Hipotecas* em milhares de dólares, em vez de milhões de dólares, como mudaria o coeficiente de correlação?
d) Suponha que, em outro ano, as taxas de juros fossem de 11% e as hipotecas totalizassem $250 milhões. Como esses dados afetariam o coeficiente de correlação, se fossem incluídos?
e) Esses dados fornecem provas de que, se as taxas das hipotecas baixassem, as pessoas fariam mais hipotecas? Explique.

17. Emissões de carbono. O diagrama de dispersão mostra, para carros de 2008, a emissão de carbono (em toneladas de CO_2 por ano) *versus* o consumo em uma autoestrada, da nova Agência de Proteção ao Meio Ambiente (Environment Protection Agency – EPA), de 82 sedãs familiares, conforme registro do governo norte-americano (www.fueleconomy.gov/feg/byclass.htm). O carro com a maior Taxa de milhas por galão na autoestrada e a menor emissão de carbono é o Toyota Prius.



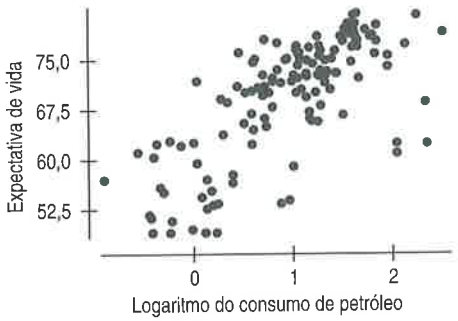
a) A correlação é –0,947. Descreva a associação.
b) As condições e suposições para o cálculo da correlação foram satisfeitas?
c) Usando a tecnologia, encontre a correlação dos dados quando o Prius não estiver incluído nos demais. Você pode explicar por que ela muda dessa forma?

18. Mpg da EPA. Em 2008, a EPA revisou seus métodos para estimar o consumo de combustível (milhas por galão) dos carros – fator de importância crescente nas vendas de carros. Como os novos valores estimados da mpg na autoestrada e na cidade estão relacionados entre si? Eis um diagrama de dispersão para 83 sedãs familiares, conforme registrado pelo governo. Esses são os mesmos carros do Exercício 17, exceto pelo Toyota Prius, removido dos dados, e por dois outros híbridos, o Nissan Altima e o Toyota Camry, incluídos nos dados (e são os carros com mpg mais alto na cidade).



a) A correlação dessas duas variáveis é 0,823. Descreva a associação.
b) Se os dois híbridos fossem removidos dos dados, você esperaria que a correlação aumentasse, diminuísse ou permanecesse a mesma? Tente usar a tecnologia. Relate e discuta o que você encontrar.

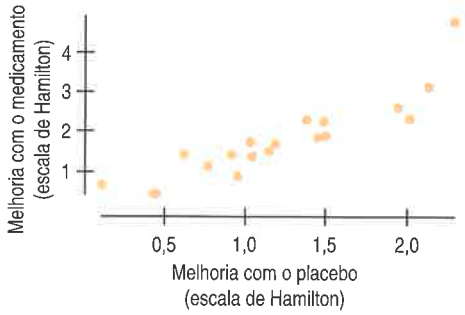
19. Consumo de petróleo. O diagrama de dispersão mostra o relacionamento entre a *Expectativa de Vida* e o logaritmo de *Consumo do Petróleo* em 137 países do mundo para os quais ambas as variáveis estão disponíveis.



a) É apropriado calcular a correlação? Explique.
b) A correlação é 0,80. Descreva a associação.

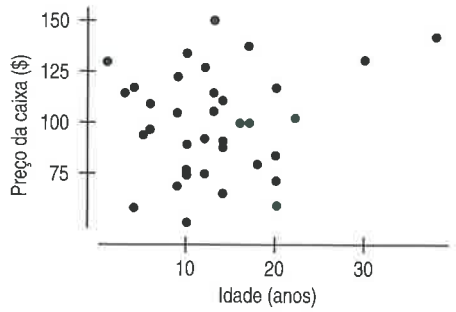
20. Antidepressivos. Quatorze anos após a Eli Lilly ter introduzido o Prozac, o mercado de antidepressivos cresceu a ponto de se tornar uma indústria de $12 bilhões, superada somente pelos medicamentos para o coração entre os remédios prescritos. No entanto, a eficácia desses remédios ainda é discutida. Um estudo comparou a eficácia de vários antidepressivos examinando os experimentos da Food and Drug Administration (FDA) nos quais eles foram aprovados. Cada um desses experimentos comparou a droga ativa com um placebo, uma pílula inerte fornecida a alguns dos pacientes. Em cada experimento, algumas pessoas tratadas com o placebo melhoraram, um fenômeno chamado de *efeito placebo*. Os níveis de depressão dos pacientes foram avaliados numa escala padrão para classificar quantitativamente a depressão, chamada de Escala de Classificação de Hamilton. As mudanças positivas nas classifica-

ções da escala registram melhoras nas condições dos pacientes. A escala de Hamilton é um padrão amplamente aceito, que foi usado em cada um desses estudos executados independentemente. O diagrama de dispersão compara os níveis médios de melhoria com os antidepressivos e com os placebos para os experimentos.



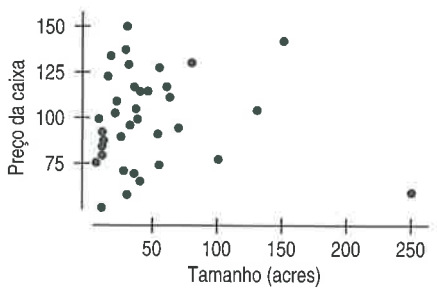
a) É apropriado calcular a correlação? Explique.
b) A correlação é 0,898. Explique o que descobrimos a partir dos resultados desses experimentos.

21. Vinhedos. Aqui está um diagrama de dispersão e correlação para *preço da caixa* de vinhos de 36 vinhedos dos Lagos Finger, região do estado de Nova York, e *idade* dos vinhedos.



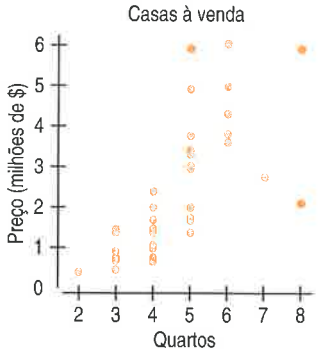
a) Verifique as condições e suposições para a correlação.
b) Parece que os vinhedos mais antigos têm vinhos com preços mais altos? Explique.
c) O que essa análise indica sobre os vinhedos no restante do mundo?

22. Vinhedos, novamente. Em vez da idade do vinhedo, considerado no Exercício 22, talvez o *tamanho* do vinhedo (em acres) esteja associado ao preço dos vinhos. Veja o diagrama de dispersão.



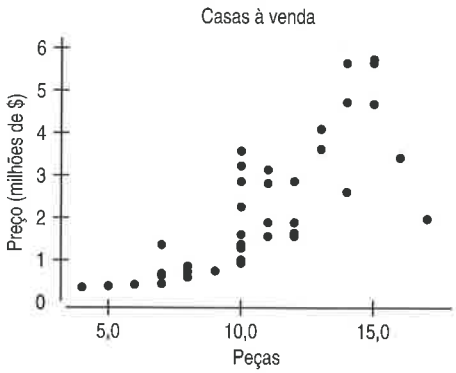
a) A correlação é –0,022. O preço fica menor com o aumento do tamanho do vinhedo? Explique.
b) Se o ponto que corresponde ao maior vinhedo fosse removido, que efeito isso teria na correlação?

23. Imóveis. Usando uma amostra aleatória de casas à venda, um comprador em potencial está interessado em examinar a relação entre o *Preço* e o *Número de Quartos*. O gráfico mostra o diagrama de dispersão para o *Preço* versus o *Número de Quartos*. A correlação é 0,723.



a) Verifique as condições e suposições para a determinação da correlação.
b) Descreva a relação.

24. Imóveis, novamente. Talvez o número total de peças da casa esteja associado ao preço da casa. Eis um diagrama de dispersão para as mesmas casas que examinamos no Exercício 23.



a) Existe uma associação?
b) Verifique as suposições e condições para a correlação.

25. Vendas regionais. A líder dos vendedores de um varejista de roupas está analisando se a empresa se sai melhor em algumas partes do país do que em outras. Ela examina um diagrama de dispersão do total das *Vendas* por *Estado* do ano passado, onde os estados estão numerados em ordem alfabética, *Alaska* = 01, ..., *Wyoming* = 50. A correlação é somente 0,045, assim, ela conclui que não existem diferenças nas vendas entre os 50 estados. Comente.

26. Recursos humanos. Numa empresa pequena, o Chefe do Setor Financeiro (CSF) está preocupado com o absenteísmo dos empregados e solicita ao chefe de recursos humanos que investigue a situação. As ocupações estão codificadas de 01 a 99, com 01 = empregado do almoxarifado e 99 = presidente. O gerente de recursos humanos

diagramou o número de ausências no ano passado por tipo de ocupação e encontrou uma correlação de –0,034 e nenhuma tendência óbvia. Ele, então, informa ao CSF que parece não existir relação entre ausências no trabalho e tipo de ocupação. Comente.

27. Financiamento público da educação. Todos os 50 estados dos Estados Unidos oferecem educação superior pública por meio de faculdades e universidades de quatro anos e faculdades de dois anos. O custo do ensino varia enormemente em diferentes estados para ambos os tipos. Existe uma relação entre as taxas cobradas pelo estado para os dois tipos? (Os dados para o ano de 2007-2008 são encontrados nas variáveis *Public.2yr* e *Public.4yr.* no conjunto de dados **CH06_Tuition**.)

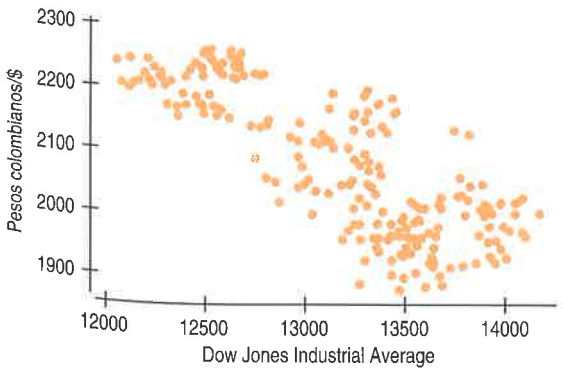
a) Examine o diagrama de dispersão do custo médio da instrução para faculdades de 4 anos contra o custo da instrução cobrado por faculdades de dois anos. Descreva a relação.
b) A direção da relação é a esperada?
c) A correlação é um resumo numérico apropriado da força da relação? Explique. Se for, encontre-a.

28. Educação superior pública e privada. No Exercício 27, examinamos a relação entre a média das taxas cobradas pelos 50 estados norte-americanos para faculdades de quatro anos *versus* faculdades de dois anos. Agora vamos analisar a relação entre as taxas cobradas por faculdades e universidades privadas de quatro anos (*Private.4yr*) com as instituições públicas de quatro anos (*Public.4yr*). Os dados estão em **Ch06_Tuition**.

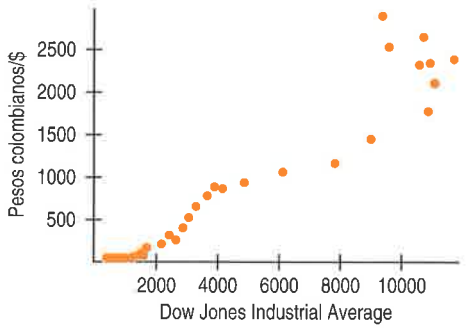
a) Você esperaria que a relação entre a taxa cobrada pelas faculdades e universidades privadas e públicas fosse tão forte quanto a relação entre instituições públicas de quatro e dois anos?
b) Examine o diagrama de dispersão e descreva a relação.
c) A correlação é um resumo numérico apropriado da força do relacionamento? Explique. Se for, encontre-a.

29. Peso colombiano. Em 2007, o peso colombiano foi uma das moedas globais que mais valorizou. Um estudante notou que, durante aquele ano, a taxa de câmbio do peso parecia se mover na direção oposta do índice Dow Jones (Dow Jones Industrial Average – DJIA), e a correlação calculada era de –0,81. O estudante conclui que o DJIA estava puxando para baixo o valor do peso. Eis um diagrama de dispersão do valor do peso colombiano (em pesos/dólar) *versus* o DJIA (www.measuringworth.co).

a) Descreva a relação.
b) A correlação é um resumo numérico apropriado da força da relação?
c) Comente a conclusão do estudante.

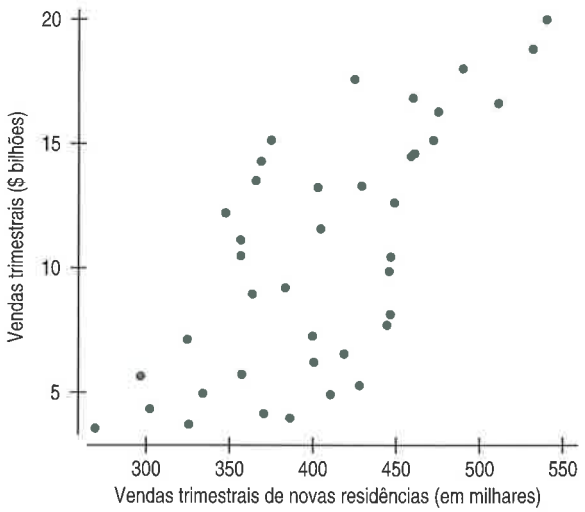


30. Peso colombiano, parte 2. No Exercício 29, um estudante comentou a correlação negativa entre a taxa de câmbio do peso colombiano e o índice Dow Jones durante o ano de 2007. Outro aluno da turma decidiu pesquisar a história da relação e examinou o diagrama de dispersão da taxa média anual de câmbio e o DJIA de 1928 a 2006 visto abaixo (www.measuringworth.com) (o *DJIA* alcançou 9000 pontos, pela primeira vez na história, em 1998). Ele conclui que a relação é positiva, porque a correlação é 0,97.



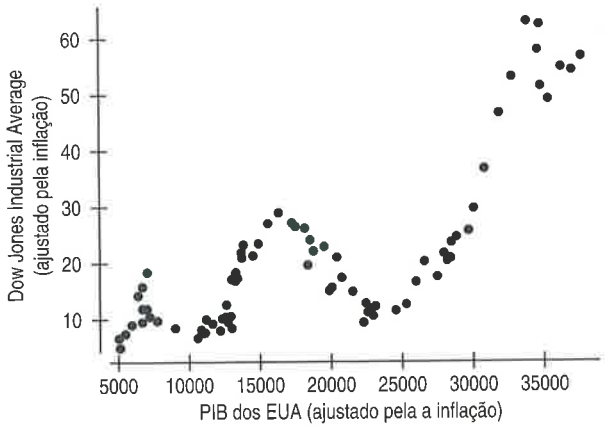
a) Descreva a relação.
b) A correlação é um resumo numérico apropriado da força da relação?
c) Comente a conclusão do estudante e explique como os dois alunos puderam chegar a conclusões tão diferentes.

31. Indústria da habitação. Um estudante, acompanhando a indústria da habitação, comparou as vendas trimestrais da Home Depot de 1995 a 2004 às vendas trimestrais de novas residências dos Estados Unidos e criou o diagrama de dispersão que vimos na Figura 7.3 (p. 204). Ele calculou a correlação como sendo 0,70 e escreveu um relatório que afirma que um crescimento de novas residências irá resultar em um aumento de vendas para a Home Depot.



a) Descreva a relação.
b) A correlação é um resumo numérico apropriado da força da relação?
c) Comente a conclusão do estudante.

32. Análise econômica. Um estudante está estudando a economia norte-americana e descobre que a correlação entre a inflação ajustada pelo índice Dow Jones e o Produto Interno Bruto (PIB) (também com a inflação ajustada) é 0,77 (www.measuringworth.com). A partir dessa informação, ele conclui que existe uma forte relação entre as duas séries e prevê que uma queda no PIB irá forçar uma queda no mercado de ações. Eis um diagrama de dispersão com o índice Dow Jones ajustado frente ao PIB (do ano de 2000, em $). Descreva a relação e comente as conclusões do estudante.



33. Erros de correlação em economia internacional. O professor da disciplina de Economia Internacional pede à sua classe para investigar fatores associados ao Produto Interno Bruto dos países. Cada estudante examina um fator diferente (como *Expectativa de Vida*, *Taxa de Analfabetismo*, etc.) para alguns países e relata à turma. Aparentemente, alguns dos alunos não entendem estatística muito bem, porque muitas das conclusões são incorretas. Explique os seus erros.

a) "Minha correlação de –0,772 mostra que praticamente não existe associação entre o *PIB* e a *Taxa de Mortalidade Infantil*."
b) "Havia uma correlação de 0,44 entre o *PIB* e o *Continente*."

34. Mais erros de correlação em economia internacional. Os estudantes da turma apresentada no Exercício 33 também escreveram estas conclusões. Explique os erros que eles cometeram.

a) "Havia uma correlação muito forte, de 1,22, entre a *Expectativa de Vida* e o *PIB*."
b) "A correlação entre a *Taxa de Alfabetização* e o *PIB* foi de 0,83. Isso mostra que países que querem aumentar seu padrão de vida devem investir pesado em educação."

35. Investimentos. Uma analista de investimentos, examinando a associação entre as vendas e os ativos das empresas, ficou surpresa quando calculou a correlação. Ela esperava encontrar uma associação muito forte, mas a correlação estava próxima de 0. Explique como um diagrama de dispersão poderá revelar a associação forte que ela antecipou.

36. Carros usados. Uma cliente que deseja comprar um carro usado acredita que há uma associação negativa entre a quilometragem de um carro usado e o seu preço. Porém, ela se surpreende quando encontra um valor próximo de 0 ao determinar a associação. Explique como um diagrama de dispersão poderia ajudá-la a entender a relação.

37. Consumo do petróleo, novamente. No Exercício 19, vimos que havia uma forte associação entre o logaritmo do consumo de petróleo e a expectativa de vida em vários países do mundo.

a) Isso significa que o consumo do petróleo é bom para a saúde?
b) O que pode explicar a forte correlação?

38. Idade e renda. As correlações entre *Idade* e *Renda* conforme mensuradas em 100 pessoas é $r = 0{,}75$. Explique se cada uma destas possíveis conclusões é ou não justificada.

a) Quando a *Idade* aumenta, a *Renda* também aumenta.
b) A forma da relação entre *Idade* e *Renda* é linear.
c) Não existem valores atípicos no diagrama de dispersão da *Renda* versus *Idade*.
d) Se mensurarmos *Idade* em anos ou meses, a correlação ainda será de 0,75.

39. Redução dos custos no transporte rodoviário. Os controladores devem estar de olho no peso dos caminhões nas principais autoestradas, mas parar e pesar os caminhões custa caro para os controladores e para os caminhoneiros. O Departamento de Transporte de Minnesota esperava economizar gastos ao calcular o peso de caminhões grandes sem ter de parar os veículos, mas usando uma escala do "peso em movimento", recentemente desenvolvida. Para testar o novo dispositivo, o departamento conduziu um teste de calibragem. Ele pesou vários caminhões parados (peso estático), assumindo que esse peso era correto. Depois, pesou novamente os caminhões enquanto estavam em movimento, para analisar quão bem a nova escala poderia estimar o peso real. Os dados estão na próxima tabela.

Peso de um caminhão (milhares de libras)

Peso em movimento	Peso estático
26,0	27,9
29,9	29,1
39,5	38,0
25,1	27,0
31,6	30,3
36,2	34,5
25,1	27,8
31,0	29,6
35,6	33,1
40,2	35,5

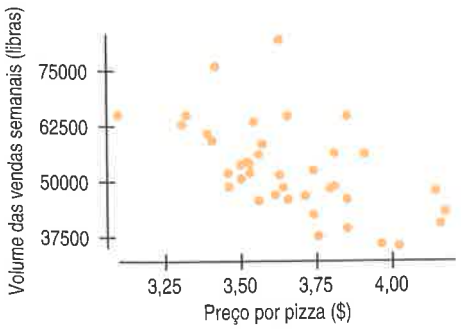
a) Faça um diagrama de dispersão para esses dados.
b) Descreva a direção, forma e força da relação.
c) Escreva algumas frases sobre o que o gráfico indica dos dados. (Observação: as frases devem ser sobre a pesagem dos caminhões, não sobre diagramas de dispersão.)
d) Encontre a correlação.
e) Se os caminhões fossem pesados em quilogramas (1 kg = 2,2 libras), como isso mudaria a correlação?
f) Alguns pontos se desviam do padrão geral? O que o diagrama indica sobre a possível recalibragem da escala do peso em movimento?

40. Fazendo economia de combustível em 2007. Em 2006, um estudo do Consumer Reports descobriu que 37% dos respondentes de todo o país pensavam em trocar seu carro atual por outro com maior economia de combustível. Eis as classificações de potência anunciadas e o consumo de gasolina esperada para vários veículos em 2007 (www.kbb.com/KBB/ReviewsAndRating).

Veículo	Potência	Consumo de combustível na estrada (mpg)
Audi A4	200	32
BMW 328	230	30
Buick Lacrosse	200	30
Chevy Cobalt	148	32
Chevy TrailBlazer	291	22
Ford Expedition	300	20
GMC Yukon	295	21
Honda Civic	140	40
Honda Accord	166	34
Hyundai Elantra	138	36
Lexus IS 350	306	28
Lincoln Navigator	300	18
Mazda Tribute	212	25
Toyota Camry	158	34
Volkswagen Beetle	150	30

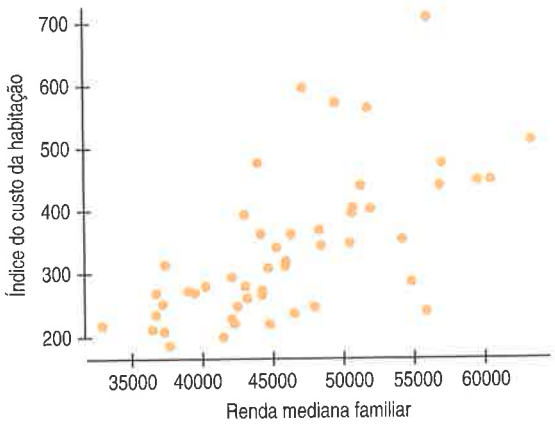
a) Faça um diagrama de dispersão para esses dados.
b) Descreva a direção, forma e força da relação.
c) Encontre a correlação entre a potência e o consumo (em mpg).
d) Escreva algumas frases relatando o que o gráfico informa sobre a economia de combustível.

41. Venda de *pizzas*. Eis um diagrama de dispersão das vendas semanais (em libras) para cada quarta semana de uma marca de *pizza* congelada *versus* o preço unitário da *pizza* para uma amostra de lojas na área de Dallas.



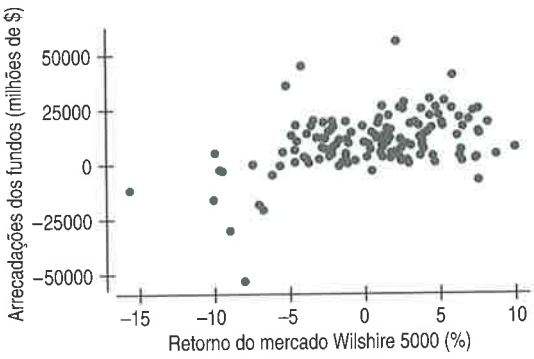
a) Verifique as suposições e as condições para a correlação.
b) Calcule a correlação entre as vendas e os preços.
c) Esse gráfico apoia a teoria de que, à medida que os preços caem, a demanda pelos produtos aumenta?
d) Se assumirmos que o número de libras de *pizza* por caixa é consistente e mensurarmos as vendas pelo número de caixas de *pizzas* vendidas em vez de libras, a correlação irá mudar? Explique.

42. Custos da habitação. A preocupação com uma possível "bolha no custo da habitação" tem levado muitos economistas a examinar os seus custos. O Office of Federal Housing Enteprise Oversight (www.ofheo.gov) coleta dados de vários aspectos dos custos da habitação em todos os Estados Unidos. Eis um diagrama de dispersão do *Índice do Custo da Habitação* versus a *Renda Mediana Familiar* para cada um dos 50 estados. A correlação é 0,65.



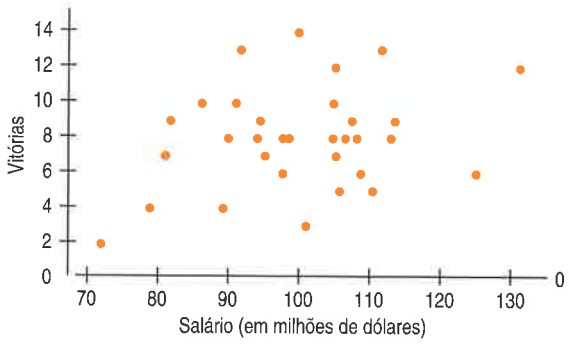
a) Descreva a relação entre o *Índice do Custo da Habitação* e a *Renda Mediana Familiar* por estado.
b) Se padronizarmos as duas variáveis, qual seria o coeficiente de correlação entre as variáveis padronizadas?
c) Se tivéssemos mensurado a *Renda Mediana Familiar* em milhares de dólares, em vez de dólares, como isso mudaria a correlação?
d) Washington, DC, tem um *Índice do Custo da Habitação* de 548 e uma renda mediana de aproximadamente $45000. Se incluíssemos DC no conjunto de dados, como isso afetaria o coeficiente de correlação?
e) Esses dados fornecem provas de que aumentar a renda mediana num estado aumenta o *Índice do Custo da Habitação* como consequência? Explique.

43. Fundos Mútuos. Eis um diagrama de dispersão mostrando a associação entre o dinheiro arrecadado nos fundos mútuos (o fundo arrecada em milhões de $) e um tipo específico de retorno de mercado (Índice Wilshire) para cada mês de 1990 a 2002.

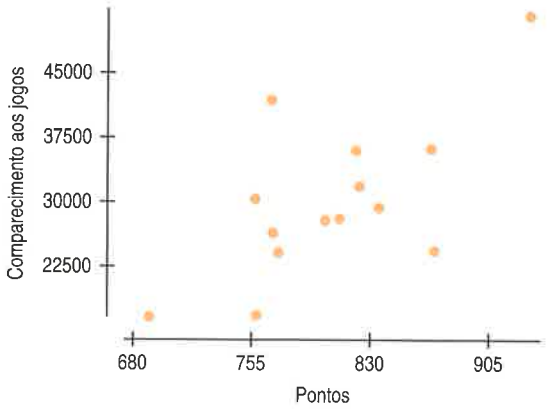


a) É apropriado calcular a correlação? Explique.
b) Identifique o valor atípico maior no diagrama de dispersão. Encontre e discuta a correlação após eliminar esse valor atípico.

44. Salários do futebol americano. Os pagamentos para os 32 times na National Football League (NFL – Liga Nacional de Futebol Americano) variam muito. Os altos salários levam a mais vitórias? Veja um diagrama de dispersão de vitórias *versus* salários das equipes para 2006.



45. Comparecimento em 2006. Os jogos da Liga Americana de Beisebol são disputados sob a regra do batedor designado, significando que arremessadores fracos não pegam o bastão. Os proprietários dos times de beisebol acreditam que essa regra leva a mais pontos, o que, por sua vez, gera um comparecimento maior. Existem indícios de que mais torcedores comparecem aos jogos se os times marcam mais pontos? Os dados coletados dos Jogos da Liga Americana durante a temporada de 2006 têm uma correlação de 0,667 entre *Pontos Marcados* e o *Número de Pessoas que Assistiram ao Jogo* (www.mlb.com).

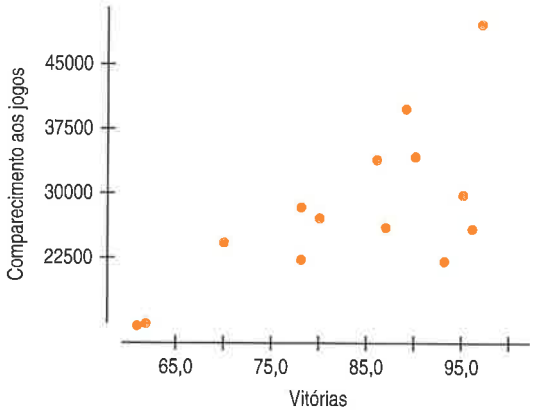


a) O diagrama de dispersão indica que é apropriado calcular a correlação? Explique.
b) Descreva a associação entre o comparecimento e os pontos marcados.
c) Essa associação prova que os proprietários estão certos que mais torcedores comparecerão aos jogos se os times marcarem mais pontos?

46. Segunda rodada de 2006. Talvez os torcedores estejam interessados apenas em times que vencem. As representações são baseadas nos times da Liga Americana para a temporada de 2008 (espn.go.com). As equipes que vencem são necessariamente aquelas que marcam mais pontos?

Correlação

	Vitórias	Pontos	Comparecimento
Vitórias	1,000		
Pontos	0,605	1,000	
Comparecimento	0,697	0,667	1,000



a) Os times vencedores sempre desfrutam de grandes comparecimentos nos jogos em casa? Descreva a associação.
b) O comparecimento está mais fortemente associado à vitória ou à marcação de pontos? Explique.
c) Quão forte está marcar mais pontos associado a vencer mais jogos?

47. Arrecadação de recursos. Analistas de uma organização filantrópica querem prever que tipo de pessoa tem maior probabilidade de doar dinheiro para a próxima campanha de arrecadação de recursos. Eles consideraram as variáveis *Estado Civil* (solteiro = 1, casado = 2, divorciado = 3, viúvo = 4) e *Doador* (Não = 0, Sim = 1) dos possíveis doadores, e encontraram uma correlação de 0,089 entre as duas variáveis. Comente sobre a conclusão de que o estado civil não tem associação quanto à pessoa responder ou não à campanha. O que a organização deveria ter feito com esses dados?

48. Juntando fontes de dados. Uma empresa está verificando seu banco de dados para ver se as duas fontes de dados utilizadas como amostra têm o mesmo código postal. A variável *fonte dos dados* = 1, se a fonte dos dados for *MetroMedia*; 2, se a fonte dos dados for *DataQwest*; 3, se for *RollingPoll*. A organização acha que a correlação entre um código postal de cinco dígitos e a *fonte dos dados* é –0,0229. Ela conclui que a correlação é baixa o suficiente para declarar que não existe dependência entre *código postal* e *fonte de dados*. Comente.

49. Produção de petróleo. A tabela mostra a produção de petróleo dos Estados Unidos de 1949 a 2005 (em milhares de barris por ano).

Ano	Produção (milhares de barris)	Ano	Produção (milhares de barris)
1949	1841940	1978	3178216
1950	1973574	1979	3121310
1951	2247711	1980	3146365
1952	2289836	1981	3128624
1953	2757082	1982	3156715
1954	2314988	1983	3170999
1955	2484420	1984	3249696
1956	2617283	1985	3274553
1957	2616901	1986	3168252
1958	2448987	1987	3047377
1959	2574590	1988	2979126
1960	2574933	1989	2778772
1961	2621758	1990	2684689
1962	2676189	1991	2707039
1963	2752723	1992	2624631
1964	2786822	1993	2499033
1965	2848514	1994	2431476
1966	3027763	1995	2394269
1967	3215742	1996	2366016
1968	3329042	1997	2354831
1969	3371751	1998	2281920
1970	3517450	1999	2146732
1971	3453914	2000	2130706
1972	3455368	2001	2117512
1973	3360903	2002	2097124
1974	32202585	2003	2073454
1975	3056779	2004	1983300
1976	2976180	2005	1890107
1977	3009265		

a) Encontre a correlação entre a produção e o ano.
b) Um repórter conclui que uma correlação baixa entre *ano* e *produção* mostra que a produção de petróleo tem permanecido estável por um período de 57 anos. Você concorda com essa interpretação? Explique.

T **50. Indústria aérea.** A indústria aérea cresceu rapidamente durante a década de 1995 a 2005. A tabela mostra o número de voos em cada um desses anos.

Ano	Voos
1995	5327435
1996	5351983
1997	5411843
1998	5384721
1999	5527884
2000	5683047
2001	5967780
2002	5271359
2003	6488539
2004	7129270
2005	7140596

a) Encontre a correlação de *Voos* e *Ano*.
b) Faça um diagrama de dispersão.
c) Aponte duas razões por que a correlação que você achou em a) não é um resumo adequado da força da associação. Você pode esclarecer essas violações das condições?

RESPOSTAS DO TESTE RÁPIDO

1 Sabemos que os escores são quantitativos. Devemos verificar se a *Condição de Linearidade* e a *Condição do Valor Atípico* foram satisfeitas, observando um diagrama de dispersão dos dois escores.

2 Não irá mudar.

3 Não irá mudar.

4 Eles provavelmente ficarão baixos. A correlação positiva significa que os preços baixos do fechamento para a Intel estão associados a preços baixos de fechamento para a Cypress.

5 Não, a associação geral é positiva, mas os preços de fechamento diários podem variar.