

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

- a. Por que uma distribuição de freqüência não pode ter por base os intervalos de classe 0–50, 50–100, 100–150 e assim por diante?
- b. Construa uma distribuição de freqüência e um histograma dos dados usando limites de classes 0, 50, 100, ... e então faça comentários sobre as características interessantes.
- c. Construa uma distribuição de freqüência e um histograma dos logaritmos naturais relacionados às observações de vida útil e comente as características interessantes.
- d. Que proporção das observações de vida útil dessa amostra é inferior a 100? Que proporção das observações é igual ou maior que 200?
28. Construa um gráfico de pontos para a série de dados anexa. Os dados são mensais e foram obtidos durante o período de 1985–1989. Cada valor é a radiação solar média na faixa 385–530 nm como porcentagem da radiação total (“Global Energy in the Different Spectral Bands at Dhahran, Saudi Arabia,” *J. Solar Energy Engr.*, 1991, p. 290–294). Comente sobre algumas características interessantes dos dados.

20,9	19,6	20,4	20,3	20,8	20,6	20,5	20,4
19,9	19,8	19,5	20,2	16,5	18,3	18,7	19,6
20,0	20,0	19,5	19,6	19,1	18,8	18,3	17,6
17,2	17,8	18,7	19,0	19,0	18,6	18,8	19,0
18,5	18,3	17,5	16,9	17,0	17,8	18,1	18,8
18,9	18,9	19,1	18,8	18,4	17,8	17,0	16,8
17,9	18,4	19,0	19,4	19,7	19,5	19,5	19,5
19,0	18,7	18,1	17,9				

29. Considere os dados a seguir sobre os tipos de queixas de saúde (J = inflamação de articulações, F = fadiga, B = dor nas costas, M = fadiga muscular, C = tosse, N = irritação nasal/coriza, O = outros) feitas por agricultores. Obtenha as freqüências e as freqüências relativas das diversas categorias e desenhe um histograma. (Os dados são consistentes com as porcentagens fornecidas no artigo “Physiological Effects of Work Stress and Pesticide Exposure in Tree Planting by British Columbia Silviculture Workers,” *Ergonomics*, 1993, p. 951–961.)

O	O	N	J	C	F	B	B	F	O	J	O	O	M
O	F	F	O	O	N	O	N	J	F	J	B	O	C
J	O	J	J	F	N	O	B	M	O	J	M	O	B
O	F	J	O	O	B	N	C	O	O	O	M	B	F
J	O	F	N										

30. Um **Diagrama de Pareto** é uma variação de um histograma para dados categorizados resultantes de um estudo de controle de qualidade. Cada categoria representa um tipo diferente de não-conformidade de produto ou problema de produção. As categorias são ordenadas de forma que aquela com maior freqüência seja exibida na extremidade esquerda, seguida pela categoria com a segunda maior freqüência e assim por diante. Suponha que as informações a seguir tenham sido obtidas sobre não-conformidades em pacotes de circuitos: componentes com falha, 126; componentes incorretos, 210; soldas insuficientes, 67; soldas em excesso, 54; falta de componentes, 131. Construa um Diagrama de Pareto.

31. A **freqüência acumulada** e a freqüência relativa acumulada de um determinado intervalo de classe são a soma das freqüências e freqüências relativas, respectivamente, desse intervalo e de todos os intervalos abaixo dele. Se, por exemplo, houver quatro intervalos com freqüências 9, 16, 13 e 12, as freqüências acumuladas serão 9, 25, 38 e 50 e as freqüências relativas acumuladas serão 0,18, 0,50, 0,76 e 1,00. Calcule as freqüências acumuladas e as freqüências relativas acumuladas para os dados do Exercício 24.

32. Uma carga de incêndio ( $\text{MJ/m}^2$ ) é a energia térmica que pode ser liberada por metro quadrado de área de piso pela combustão de seu conteúdo e da estrutura em si. O artigo “Fire Loads in Office Buildings” (*J. of Structural Engr.*, 1997, p. 365–368) forneceu as seguintes porcentagens acumuladas (lidas de um gráfico) relativas a cargas de incêndio em uma amostra de 388 salas:

Valor	0	150	300	450	600
% Acumulada	0	19,3	37,6	62,7	77,5
Valor	750	900	1050	1200	1350
% Acumulada	87,2	93,8	95,7	98,6	99,1
Valor	1500	1650	1800	1950	
% Acumulada	99,5	99,6	99,8	100,0	

- a. Construa um histograma de freqüência relativa e comente as características interessantes.  
 b. Que proporção das cargas de incêndio é inferior a 600? Maior ou igual a 1200?  
 c. Que proporção das cargas está entre 600 e 1200?

## 1.3 | Medidas de localização

Os resumos visuais de dados são excelentes ferramentas para obter impressões e idéias iniciais. Uma análise mais formal de dados freqüentemente exige o cálculo e a interpretação de medidas-resumo numéricas simples. Isto é, a partir dos dados, tentamos extrair diversos números simples, que servem para caracterizar o conjunto de

dados e indicar algumas informações consideráveis. Nossa preocupação principal será com os dados numéricos. Alguns comentários sobre dados categorizados serão apresentados no final da seção.

Suponha, então, que nosso conjunto de dados é do formato  $x_1, x_2, \dots, x_n$ , onde cada  $x_i$  é um número. Que características de tal conjunto de números são de maior interesse e merecem ênfase? Uma característica importante de um conjunto de números é sua localização e, em particular, seu centro. Esta seção apresenta métodos de descrição da localização de um conjunto de dados. Na Seção 1.4, apresentaremos os métodos de medida da dispersão de um conjunto de números.

A mídia

Para um determinado conjunto de números  $x_1, x_2, \dots, x_n$ , a medida mais familiar e útil do centro é a *média* do conjunto. Como quase sempre temos os vários  $x_i$  constituindo uma amostra, freqüentemente chamaremos a média aritmética de *média amostral* e a representaremos por  $\bar{x}$ .

## DEFINIÇÃO

A **média amostral**  $\bar{x}$  das observações  $x_1, x_2, \dots, x_n$ , é dada por

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

O numerador de  $\bar{x}$  pode ser escrito mais informalmente como  $\sum x_i$ , onde a soma se dá sobre todas as observações da amostra.

Para informar  $\bar{x}$ , recomendamos o uso de precisão decimal de um dígito a mais do que a precisão dos  $x_i$ . Dessa forma, se as observações forem distâncias de parada com  $x_1 = 125$ ,  $x_2 = 131$  e assim por diante, podemos ter  $\bar{x} = 127,3$  pés.

### Exemplo 1.13

As trincas em aço e ferro causadas por fadiga de corrosão cáustica foram estudadas em decorrência de falhas em rebites de caldeiras de aço e em rotores a vapor. Considere as observações a seguir sobre  $x$  = comprimento da trinca ( $\mu\text{m}$ ) como resultado de testes de fadiga por corrosão devido a cargas constantes em amostras de barras de tração lisas durante um período de tempo fixo. (Os dados são consistentes com um histograma e as quantidades-resumo do artigo “On the Role of Phosphorus in the Caustic Stress Corrosion Cracking of Low Alloy Steels”, *Corrosion Science*, 1989: 53-68.)

$$\begin{array}{ccccccccc} x_1 = 16,1 & x_2 = 9,6 & x_3 = 24,9 & x_4 = 20,4 & x_5 = 12,7 & x_6 = 21,2 & x_7 = 30,2 \\ x_8 = 25,8 & x_9 = 18,5 & x_{10} = 10,3 & x_{11} = 25,3 & x_{12} = 14,0 & x_{13} = 27,1 & x_{14} = 45,0 \\ x_{15} = 23,3 & x_{16} = 24,2 & x_{17} = 14,6 & x_{18} = 8,9 & x_{19} = 32,4 & x_{20} = 11,8 & x_{21} = 28,5 \end{array}$$

A Figura 1.14 mostra um diagrama de caule e folha dos dados. Um comprimento de trinca no início da faixa dos 20 parece ser “típica”.

0H	96	89				
1L	27	03	40	46	18	
1H	61	85				
2L	49	04	12	33	42	Caule: dígito das dezenas
2H	58	53	71	85		Folha: dígito das unidades e das dezenas
3L	02	24				
3H						
4L						
4H	50					

**Figura 1.14** Um diagrama de caule e folha dos dados dos comprimentos de trincas

Sendo  $\sum x_i = 444,8$ , a média amostral é

$$\bar{x} = \frac{444,8}{21} = 21,18$$

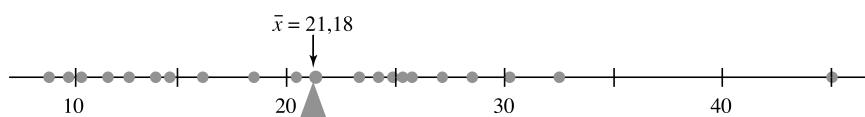
um valor consistente com as informações ilustradas pelo diagrama de caule e folha. ■

Uma interpretação física de  $\bar{x}$  demonstra como ela mede a localização (centro) de uma amostra. Imagine desenhar e definir a escala em um eixo horizontal e depois represente cada observação da amostra por um peso de uma libra colocado no ponto correspondente no eixo. O único ponto em que pode ser colocado um apoio para equilibrar o sistema de pesos é o correspondente ao valor de  $\bar{x}$  (veja a Figura 1.15).

Da mesma forma que  $\bar{x}$  representa o valor médio das observações de uma amostra, a média de todos os valores da população pode ser calculada. Essa média é denominada **média da população** e é representada pela letra grega  $\mu$ . Quando houver  $N$  valores na população (uma população finita),  $\mu = (\text{somatória dos } N \text{ valores da população})/N$ . Nos capítulos 3 e 4, forneceremos uma definição mais geral de  $\mu$  que se aplica a populações finitas e (conceitualmente) infinitas. Da mesma forma que  $\bar{x}$  é uma medida de localização de amostra importante e interessante,  $\mu$  é uma característica interessante e importante (freqüentemente a mais importante) de uma população. Nos capítulos sobre inferência estatística, apresentaremos métodos com base na média amostral para obtenção de conclusões sobre a média de uma população. Por exemplo: podemos usar a média amostral  $\bar{x} = 21,18$  calculada no Exemplo 1.13 como uma *estimativa de ponto* (um único número que é o “melhor” palpite) de  $\mu$ , o comprimento médio verdadeiro de todas as amostras tratadas como descrito.

A média sofre de uma deficiência que a torna uma medida de centro inadequada sob algumas circunstâncias: seu valor pode ser bastante afetado pela presença de um único *outlier* (uma observação incomumente grande ou pequena). No Exemplo 1.13, o valor  $x_{14} = 45,0$  obviamente é um *outlier*. Sem esta observação,  $\bar{x} = 399,8/20 = 19,99$ , o *outlier* aumenta a média em mais de 1  $\mu\text{m}$ . Se a observação 45,0  $\mu\text{m}$  fosse substituída pelo valor catastrófico de 295,0  $\mu\text{m}$ , um *outlier* realmente extremo, então  $\bar{x} = 694,8/21 = 33,09$ , que é maior que todas as observações, exceto uma.

Uma amostra de salários normalmente produz alguns poucos valores aberrantes (dos sortudos que possuem um salário astronômico) e o uso do salário médio como medida de localização freqüentemente será ilusório. Esses exemplos sugerem que procuremos uma medida menos sensível a valores fora da faixa que  $\bar{x}$ , assim, proporemos uma momentaneamente. Entretanto, apesar de  $\bar{x}$  ter essa falha potencial, ela ainda é a medida mais usada, em grande parte porque há muitas populações para as quais um *outlier* extremo na amostra seria altamente improvável. Ao obter uma amostra de uma tal população (a população normal ou em forma de sino, é o exemplo mais importante), a média amostral tenderá a ser estável e muito representativa.



**Figura 1.15** A média como ponto de equilíbrio de um sistema de pesos

## A mediana

A palavra *mediana* é sinônimo de “metade” e a mediana amostral é o valor do meio quando as observações são ordenadas da menor para a maior. Quando as observações estiverem representadas por  $x_1, \dots, x_n$ , usaremos o símbolo  $\tilde{x}$  para representar a mediana amostral.

## DEFINIÇÃO

A **mediana amostral** é obtida pela ordenação das  $n$  observações da menor para a maior (com os valores repetidos incluídos, de forma que cada observação da amostra seja exibida na lista ordenada). Assim,

$$\tilde{x} = \begin{cases} \text{O único valor médio se } n \text{ for ímpar} & = \left( \frac{n+1}{2} \right) \text{ enésimo valor ordenado} \\ \text{A média dos dois valores médios se } n \text{ for par} & = \text{média dos valores ordenados } \left( \frac{n}{2} \right) \text{ e } \left( \frac{n}{2} + 1 \right) \end{cases}$$

## Exemplo 1.14

O risco de desenvolvimento de deficiência de ferro é especialmente alto durante a gravidez. O problema na detecção dessa deficiência é que alguns métodos de determinação de nível de ferro podem ser afetados pelo próprio estado de gravidez. Considere os dados a seguir sobre a concentração do receptor de transferrina de uma amostra de mulheres com evidências laboratoriais de uma visível anemia por deficiência de ferro (“Serum Transferrin Receptor for the Detection of Iron Deficiency in Pregnancy,” *Amer. J. of Clinical Nutrition*, 1991: p. 1077-1081):

$$\begin{array}{llllll} x_1 = 15,2 & x_2 = 9,3 & x_3 = 7,6 & x_4 = 11,9 & x_5 = 10,4 & x_6 = 9,7 \\ x_7 = 20,4 & x_8 = 9,4 & x_9 = 11,5 & x_{10} = 16,2 & x_{11} = 9,4 & x_{12} = 8,3 \end{array}$$

A lista dos valores ordenados é

$$7,6 \quad 8,3 \quad 9,3 \quad 9,4 \quad 9,4 \quad 9,7 \quad 10,4 \quad 11,5 \quad 11,9 \quad 15,2 \quad 16,2 \quad 20,4$$

Como  $n = 12$  é par, tiramos a média  $n/2 =$  do sexto e sétimo valores ordenados:

$$\text{mediana amostral} = \frac{9,7 + 10,4}{2} = 10,05$$

Observe que, se a maior observação, 20,4, não tivesse aparecido na amostra, a mediana amostral resultante para as  $n = 11$  observações teria sido o único valor médio, 9,7 ( $(n+1)/2 =$  sexto valor ordenado). A média amostral é  $\bar{x} = \sum x_i/n = 139,3/12 = 11,61$ , que é um pouco maior que a mediana, por causa dos *outliers*, 15,2, 16,2 e 20,4. ■

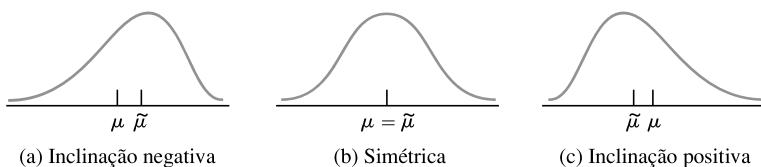
Os dados do Exemplo 1.14 ilustram uma propriedade importante de  $\tilde{x}$  em comparação com  $\bar{x}$ : a mediana amostral é muito insensível a muitos valores extremamente pequenos ou extremamente grandes. Se, por exemplo, aumentássemos os dois maiores  $x_i$  de 16,2 e 20,4 para 26,2 e 30,4, respectivamente,  $\tilde{x}$  não seria afetado. Dessa forma, no tratamento de valores de dados fora da faixa,  $\bar{x}$  e  $\tilde{x}$  são extremidades opostas de um espectro:  $\bar{x}$  é sensível mesmo a um único valor, enquanto  $\tilde{x}$  é insensível a um grande número de valores fora da faixa.

Como os valores grandes na amostra do Exemplo 1.14 afetam  $\bar{x}$  mais que  $\tilde{x}$ ,  $\tilde{x} < \bar{x}$  para esses dados. Apesar de  $\bar{x}$  e  $\tilde{x}$  fornecerem uma medida para o centro da amostra em um conjunto de dados, eles em geral não serão iguais, porque enfocam diferentes aspectos da amostra.

De forma análoga,  $\tilde{x}$  como valor médio na amostra é o valor médio da população, a **mediana da população**, representada por  $\tilde{\mu}$ . Como acontece com  $\bar{x}$  e  $\mu$ , podemos considerar o uso da mediana amostral  $\tilde{x}$  para fazer inferências de  $\tilde{\mu}$ . No Exemplo 1.14, podemos usar  $\tilde{x} = 10,05$  como estimativa da concentração da mediana em toda a população a partir da qual a amostra foi selecionada. Uma mediana normalmente é usada para descrever dados de salários ou rendimentos (porque ela não é influenciada por alguns grandes salários). Se a

mediana de uma amostra dos salários de engenheiros fosse  $\tilde{x} = \$ 66.416$ , poderíamos usá-la como base para concluir que o salário mediano dos engenheiros excede \$ 60.000.

A média da população  $\mu$  e a mediana  $\tilde{\mu}$  normalmente não serão idênticas. Se a distribuição da população tiver desvio positivo ou negativo, conforme ilustrado na Figura 1.16, então  $\mu \neq \tilde{\mu}$ . Quando esse for o caso, ao fazer inferências, devemos primeiro decidir quais características das populações são de maior interesse e então proceder de acordo.



**Figura 1.16** Três formatos diferentes para uma distribuição de população

## Outras medidas de localização Quartis, Percentis e Médias Aparadas

A mediana (de população ou amostra) divide o conjunto de dados em duas partes de mesmo tamanho. Para obter melhores medidas de localização, podemos dividir os dados em mais de duas partes. Grosso modo, os quartis dividem o conjunto em quatro partes iguais, sendo que as observações acima do terceiro quartil constituem o quarto superior do conjunto de dados, o segundo quartil é idêntico à mediana e o primeiro quartil separa o quarto inferior dos três quartos superiores. De forma similar, um conjunto de dados (amostra ou população) pode ser dividido mais detalhadamente usando percentis; o 99º percentil separa o 1% superior do restante, e assim por diante. A menos que o número de observações seja um múltiplo de 100, recomenda-se cuidado na utilização de percentis. Usaremos percentis no Capítulo 4 com alguns modelos de populações infinitas, de forma que adiaremos a discussão até lá.

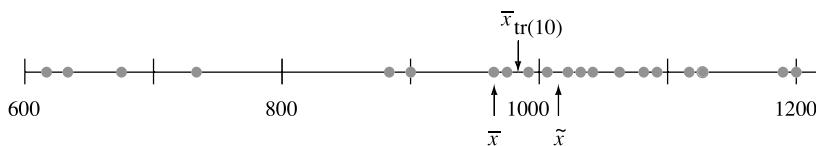
A média amostral e a mediana amostral são influenciadas por valores fora da faixa de uma forma bastante diferente: muito para a média e nada para a mediana. Como o comportamento extremo dos dois valores é indesejável, consideraremos medidas alternativas que não sejam tão sensíveis quanto  $\bar{x}$  e nem tão insensíveis como  $\tilde{x}$ . Para determinar essas alternativas, observe que  $\bar{x}$  e  $\tilde{x}$  são extremidades opostas da mesma “família” de medidas. Após o conjunto de dados ser ordenado,  $\tilde{x}$  é calculado desprezando-se todos os valores possíveis em cada extremidade sem eliminar nada (deixando apenas um ou dois valores centrais) e obtendo a média do que restou. Por outro lado, para calcular  $\bar{x}$ , nada é desprezado antes de se obter a média. Para fazer uma comparação, a média envolve desprezar 0% de cada extremidade da amostra, enquanto, para a mediana, o máximo possível é desprezado de cada extremidade. Uma **média aparada** é algo intermediário entre  $\bar{x}$  e  $\tilde{x}$ . Uma média aparada de 10%, por exemplo, seria calculada eliminando-se os 10% superiores e os 10% inferiores da amostra, obtendo-se, então, a média do restante.

### Exemplo 1.15

Considere as 20 observações a seguir, ordenadas da menor para a maior, cada uma representando a vida útil (em horas) de um determinado tipo de lâmpada incandescente:

612	623	666	744	883	898	964	970	983	1003
1016	1022	1029	1058	1085	1088	1122	1135	1197	1201

A média das 20 observações é  $\bar{x} = 965,0$  e  $\tilde{x} = 1009,5$ . A média aparada de 10% é obtida pela exclusão das duas menores observações (612 e 623) e as duas maiores (1197 e 1201) seguida do cálculo da média dos 16 valores restantes, para obter  $\bar{x}_{tr(10)} = 979,1$ . O efeito de truncar a média aqui é produzir um “valor central” ligeiramente acima da média ( $\bar{x}$  é trazido para baixo por alguns poucos valores de vida útil) e ainda consideravelmente abaixo da mediana. De forma similar, a média aparada de 20% faz uma média dos 12 valores do meio para obter  $\bar{x}_{tr(20)} = 999,9$ , mais perto ainda da mediana. (Veja a Figura 1.17.)



**Figura 1.17** Gráfico de pontos de vida útil (em horas) de lâmpadas incandescentes

Geralmente, o uso da média aparada com proporção de aparagem moderada (entre 5% e 25%) produzirá uma medida que não é nem tão sensível a *outliers* como a média nem tão insensível quanto a mediana. Por esse motivo, as médias truncadas têm sido objeto de crescente atenção dos estatísticos para propósitos descritivos e inferenciais. Mais será dito sobre médias aparadas quando a estimativa por pontos for discutida no Capítulo 6. Finalmente, se a proporção de aparagem for representada por  $\alpha$  e  $n\alpha$  não for inteiro, não será óbvio como calcular a média aparada  $100\alpha\%$ . Por exemplo: se  $\alpha = 0,10$  (10%) e  $n = 22$ , então  $n\alpha = (22)(0,10) = 2,2$  e não é possível aparar 2,2 observações de cada extremidade da amostra ordenada. Nesse caso, a média aparada de 10% seria obtida primeiro com a retirada das duas observações de cada extremidade e pelo cálculo de  $\bar{x}_{tr}$ , seguida pela retirada de três observações de cada extremidade e pelo cálculo de  $\bar{x}_{tr}$ , e então pela interpolação dos dois valores para obtenção de  $\bar{x}_{tr(10)}$ .

### Dados categorizados e proporção de amostras

Quando os dados são categorizados, uma distribuição de freqüência ou distribuição de freqüência relativa fornece um resumo tabular eficiente dos dados. Os indicadores numéricos naturais são, nessa situação, as freqüências individuais e as freqüências relativas. Por exemplo: se for feita uma pesquisa com indivíduos que possuem aparelhos de som para estudar a preferência de marca, cada indivíduo da amostra identificaria a marca do aparelho que possui. A partir disso poderíamos contar as pessoas que possuem aparelhos Sony, Pioneer, Marantz, entre outros. Considere a obtenção de uma amostra de uma população dicotômica, isto é, que consista em apenas duas categorias (votou ou não votou na eleição passada ou possui ou não um aparelho de som etc.). Se fizermos  $x$  representar o número da amostra na categoria 1, o número na categoria 2 será  $n - x$ . A freqüência relativa ou *proporção amostral* da categoria 1 será  $x/n$  e a proporção amostral da categoria 2 será  $1 - x/n$ . Vamos representar uma resposta da categoria 1 por 1 e uma resposta da categoria 2 por 0. Uma amostra de tamanho  $n = 10$  pode então resultar em 1, 1, 0, 1, 1, 1, 0, 0, 1, 1. A média dessa amostra numérica é (já que o número de ocorrências do número 1 =  $x = 7$ )

$$\frac{x_1 + \dots + x_n}{n} = \frac{1 + 1 + 0 + \dots + 1 + 1}{10} = \frac{7}{10} = \frac{x}{n} = \text{proporção amostral}$$

Esse resultado pode ser generalizado e resumido conforme segue: *Se em uma situação de dados categorizados focarmos a atenção em uma determinada categoria e codificarmos os resultados da amostra de forma que 1 seja registrado como um indivíduo da categoria e 0 para um indivíduo fora dela, a proporção amostral de indivíduos da categoria será a média amostral da seqüência de 1s e 0s.* Assim, uma média amostral pode ser usada para resumir os resultados de uma amostra categorizada. Essas observações também se aplicam a situações em que as categorias são definidas por valores agrupados em uma amostra ou população numérica (por exemplo: podemos querer saber se os indivíduos possuem seu automóvel atual há pelo menos cinco anos em vez de estudarmos o tempo exato de posse).

De forma análoga à proporção amostral  $x/n$  de indivíduos que estão em uma determinada categoria, representemos por  $p$  a proporção dos indivíduos da população inteira que pertencem à categoria. Como acontece com  $x/n$ ,  $p$  é uma quantidade entre 0 e 1 e, enquanto  $x/n$  é uma característica da amostra,  $p$  é uma característica da população. A relação entre os dois é semelhante à relação entre  $\bar{x}$  e  $\mu$  e entre  $\bar{x}$  e  $\mu$ . Em particular, usaremos  $x/n$  para fazer inferências sobre  $p$ . Se, por exemplo, uma amostra de 100 proprietários de carros revelar que 22 possuem seus carros há pelo menos 5 anos, podemos usar  $22/100 = 0,22$  como uma estimativa pontual da proporção