Capítulo 1 | Introdução à estatística e análise de dados

1.2 O papel da probabilidade

Tabela 1.1 Dados para o Exemplo 1.2.

Sem nitrogênio	Com nitrogênio
0,32	0,26
0,53	0,43
0,28	0,47
0,37	0,49
0,47	0,52
0,43	0,75
0,36	0,79
0,42	0,86
0,38	0,62
0,43	0,46

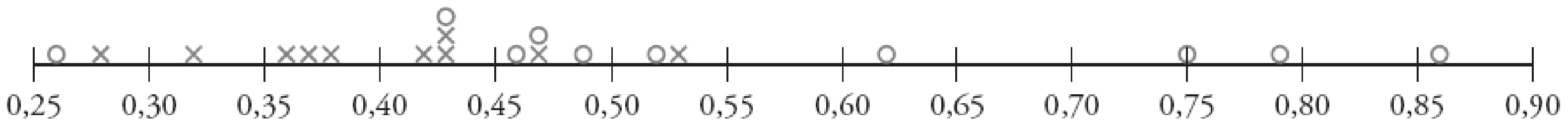


Figura 1.1 Dados do peso dos caules.

Então, para um problema estatístico, a amostra juntamente com a inferência estatística nos permitem chegar a conclusões sobre a população, com a inferência estatística fazendo claro uso dos elementos da probabilidade.

Pode-se dizer que problemas em probabilidade nos permitem tirar conclusões sobre as características de dados hipotéticos retirados da população, baseadas nas características conhecidas desta população.

Figura 1.2 Relação fundamental entre probabilidade e inferência estatística.

1.3 Procedimentos de amostragem e coleta de dados

Tabela 1.2 Dados para o Exemplo 1.3.

Revestimento	Umidade	Corrosão média em milhares de ciclos até a falha
Sem revestimento	20%	975
	80%	350
Revestimento químico	20%	1.750
	80%	1.550

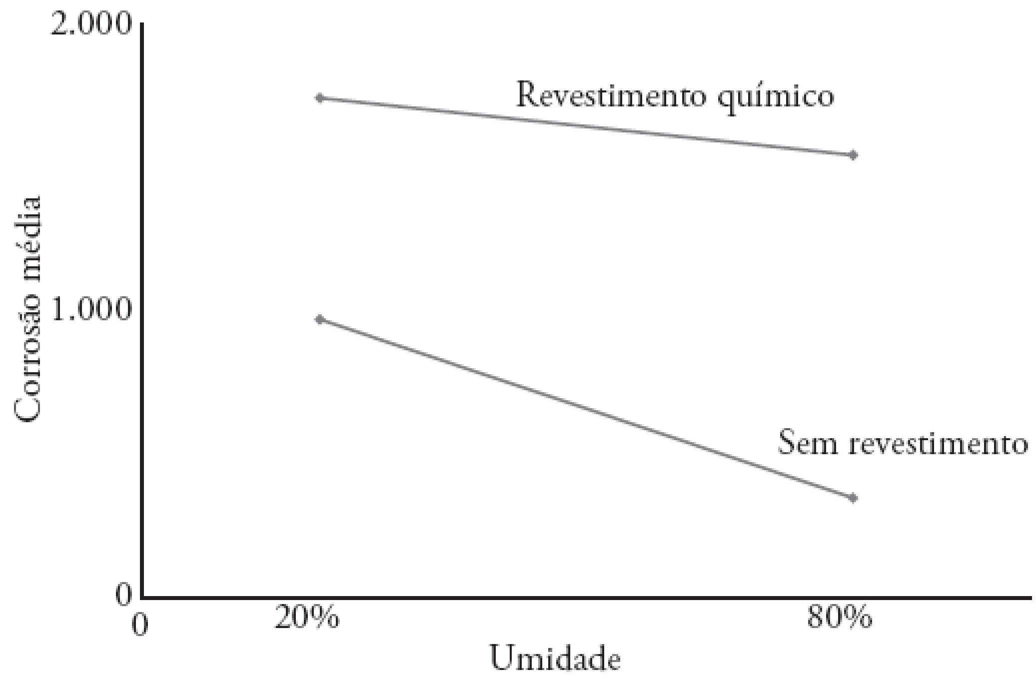


Figura 1.3 Resultados da corrosão para o Exemplo 1.3.

1.4 Medidas de localização: média e mediana amostrais

Definição 1.1

Suponha que as observações na amostra sejam $x_1, x_2, \ldots, x_n$. A *média amostral*, denotada por $\bar{x}$, é

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

Uma importante medida é a *mediana amostral. Seu propósito é* refletir a tendência central da amostra de modo que não seja influenciada por valores extremos ou discrepantes. Dado que as observações na amostra são $x_1, x_2, \ldots, x_n$, organizadas em ordem crescente, a mediana é

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{se } n \text{ for ímpar,} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{se } n \text{ for par.} \end{cases}$$

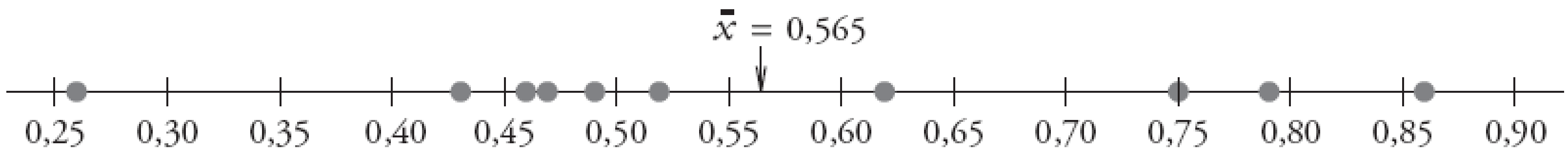


Figura 1.4 Média amostral como centróide das medidas de peso dos caules 'com nitrogênio'.

1.5 Medidas de variabilidade

Talvez a mais simples seja a *amplitude amostral* $X_{max} - X_{min}$. A amplitude pode ser muito útil.

Definição 1.2
A *variância amostral*, denotada por s^2, é dada por:

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

O *desvio-padrão amostral*, denotado por s, é a raiz quadrada positiva de s^2, isto é, $s = \sqrt{s^2}$.

1.7 Modelagem estatística, inspeção científica e diagnósticos gráficos

Tabela 1.3 Resistência à tensão.

Porcentagem de algodão	Resistência à tensão
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

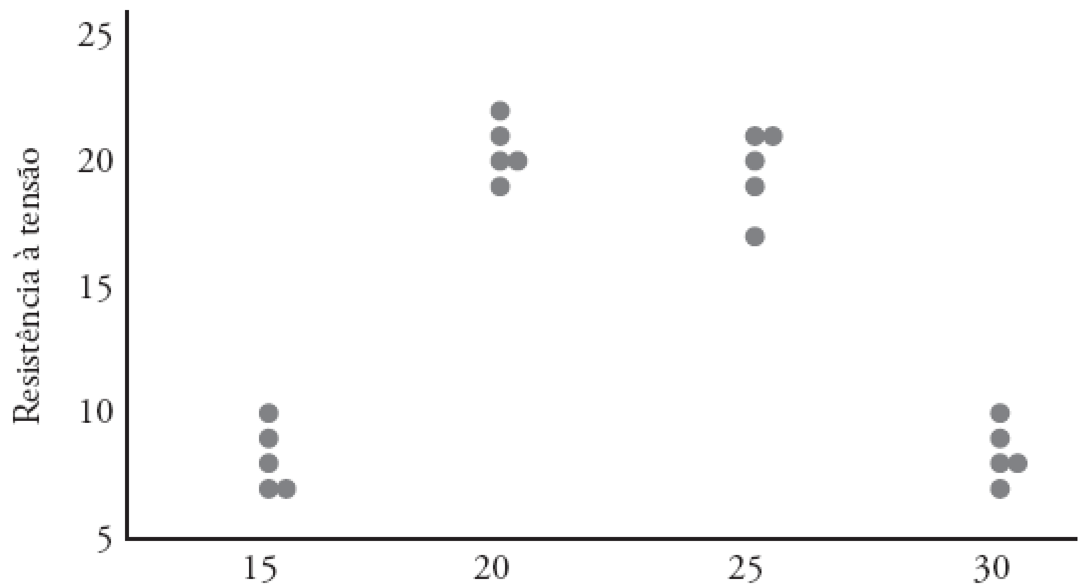


Figura 1.5 Gráfico da resistência à tensão e porcentagens de algodão.

1.8 Métodos gráficos e descrição de dados

Tabela 1.4 Vida útil das baterias de carro.

2,2	4,1	3,5	4,5	3,2	3,7	3,0	2,6
3,4	1,6	3,1	3,3	3,8	3,1	4,7	3,7
2,5	4,3	3,4	3,6	2,9	3,3	3,9	3,1
3,3	3,1	3,7	4,4	3,2	4,1	1,9	3,4
4,7	3,8	3,2	2,6	3,9	3,0	4,2	3,5

Tabela 1.5 Diagrama de ramo-e-folhas da vida das baterias.

Ramo	Folhas	Freqüência
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

Tabela 1.6 Diagrama de ramo-e-folhas duplo de vida útil das baterias.

Ramo	Folhas	Freqüência
1•	69	2
2*	2	1
2•	5669	4
3*	001111222333444	15
3•	5567778899	10
4*	11234	5
4•	577	3

Capítulo 1 | Introdução à estatística e análise de dados

Tabela 1.7 Distribuição de freqüências relativas da vida útil das baterias.

Intervalo de classe	Ponto médio de classe	Freqüência, f	Freqüência relativa
1,5–1,9	1,7	2	0,050
2,0–2,4	2,2	1	0,025
2,5–2,9	2,7	4	0,100
3,0–3,4	3,2	15	0,375
3,5–3,9	3,7	10	0,250
4,0–4,4	4,2	5	0,125
4,5–4,9	4,7	3	0,075

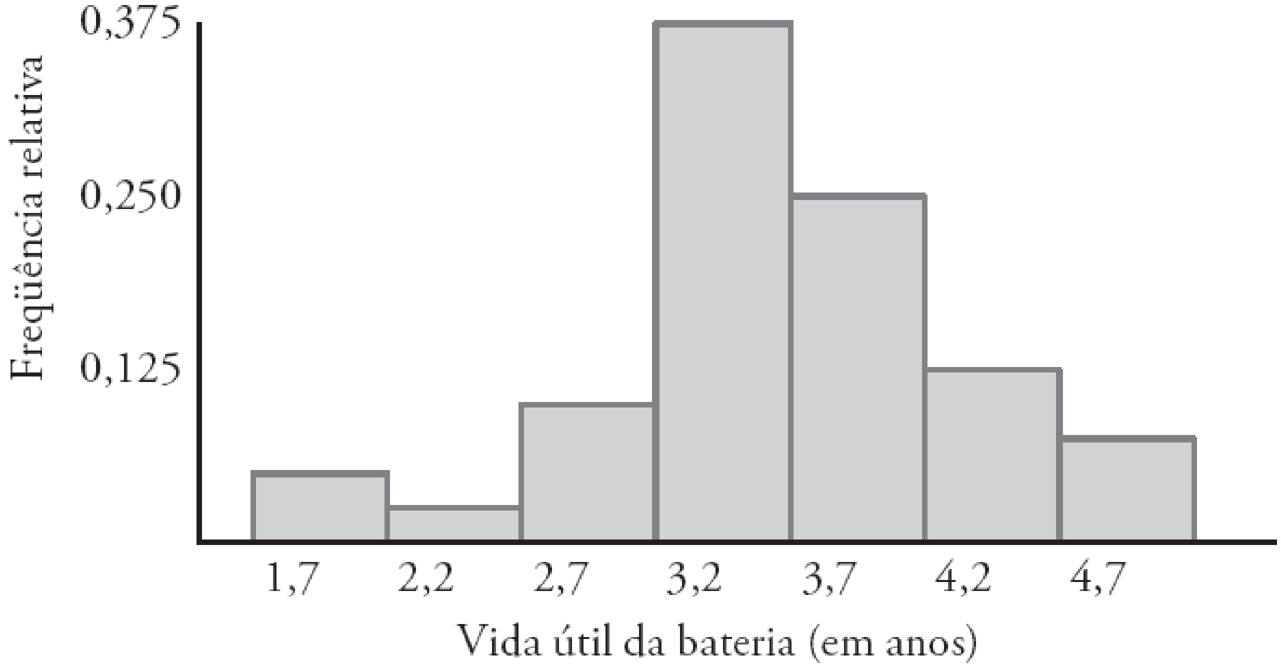


Figura 1.6 Histograma de freqüências relativas.

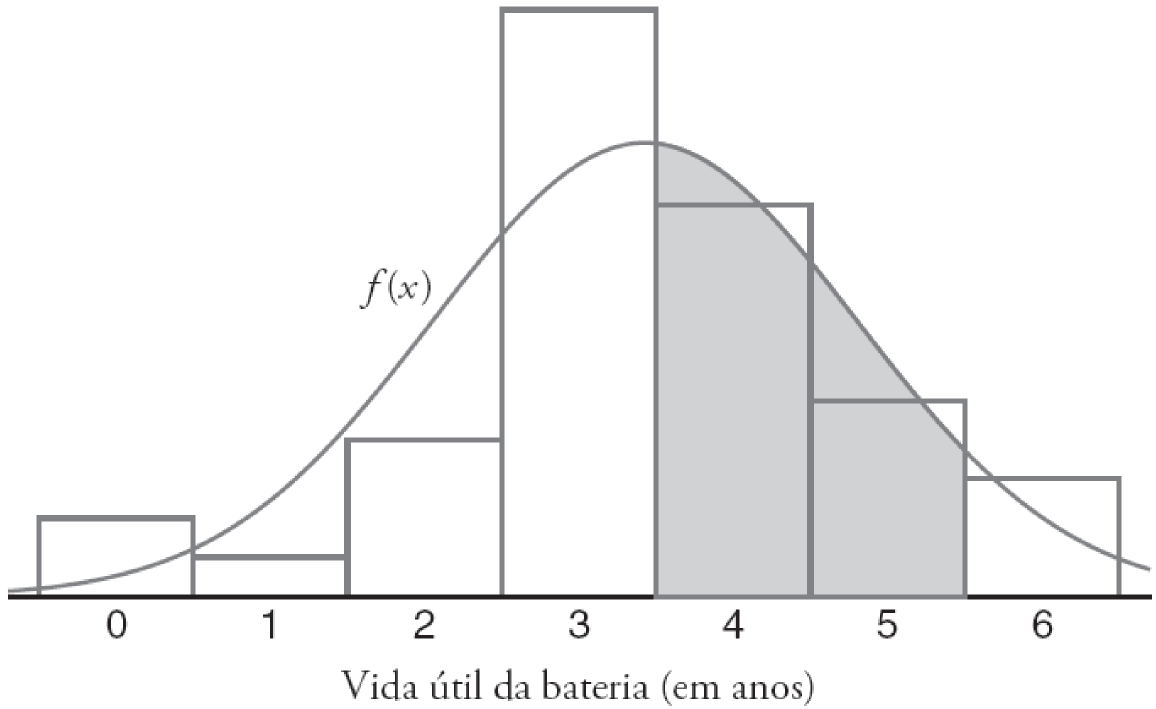


Figura 1.7 Estimativa da distribuição de freqüências.

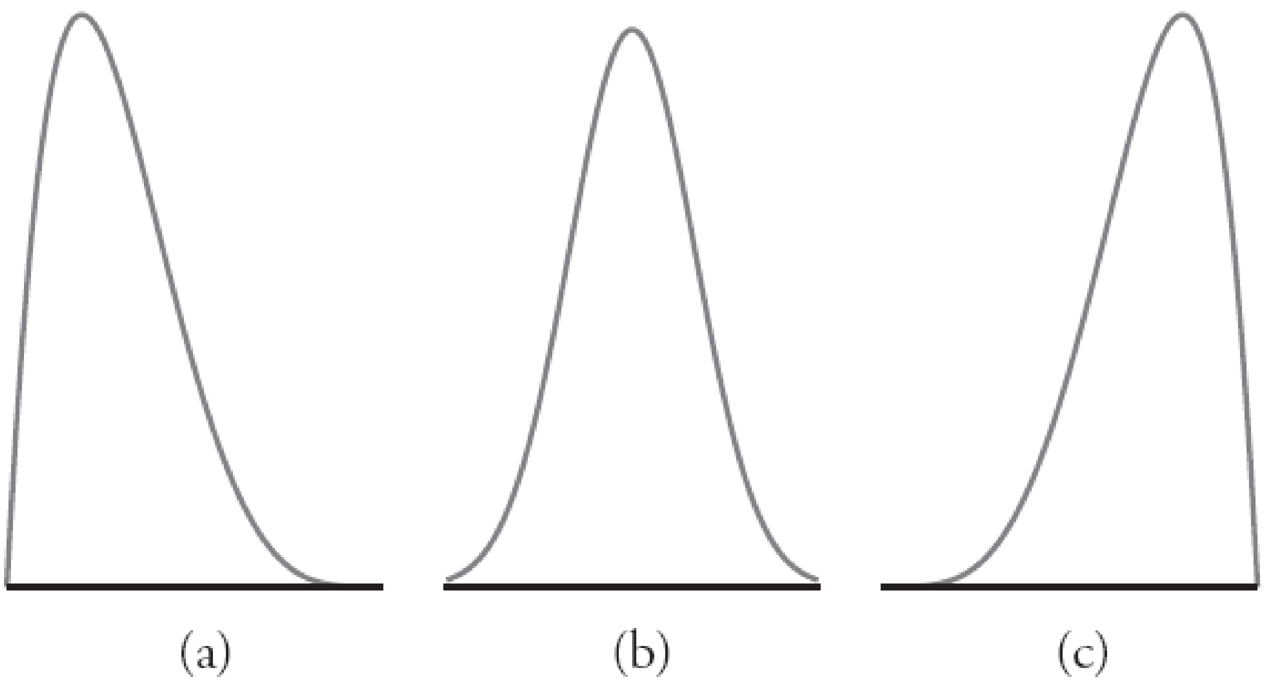


Figura 1.8 Assimetria dos dados.

Diagrama de Caixa com Bigodes

Construímos uma caixa, com referência de uma escala. Os limites da caixa são os primeiro (Q1) e terceiro quartil (Q3), e indicamos a mediana (segundo quartil). Por fim, indicamos os bigodes que ligam a caixa até os valores extremos da amostra.
Podemos estender os bigodes até o limite de uma vez e meia a distância interquartílica (DIQ=Q3-Q1).

1,09	1,92	2,31	1,79	2,28	1,74	1,47	1,97
0,85	1,24	1,58	2,03	1,7	2,17	2,55	2,11
1,86	1,9	1,68	1,51	1,64	0,72	1,69	1,85
1,82	1,79	2,46	1,88	2,08	1,67	1,37	1,93
1,4	1,64	2,09	1,75	1,63	2,37	1,75	1,69

Ordenando os dados, temos:

0,72	1,4	1,64	1,69	1,79	1,88	2,03	2,28
0,85	1,47	1,64	1,7	1,79	1,9	2,08	2,31
1,09	1,51	1,67	1,74	1,82	1,92	2,09	2,37
1,24	1,58	1,68	1,75	1,85	1,93	2,11	2,46
1,37	1,63	1,69	1,75	1,86	1,97	2,17	2,55

0,72	1,4	1,64	1,69	1,79	1,88	2,03	2,28
0,85	1,47	1,64	1,7	1,79	1,9	2,08	2,31
1,09	1,51	1,67	1,74	1,82	1,92	2,09	2,37
1,24	1,58	1,68	1,75	1,85	1,93	2,11	2,46
1,37	1,63	1,69	1,75	1,86	1,97	2,17	2,55

N=40

0,72	1,4	1,64	1,69	1,79	1,88	2,03	2,28
0,85	1,47	1,64	1,7	1,79	1,9	2,08	2,31
1,09	1,51	1,67	1,74	1,82	1,92	2,09	2,37
1,24	1,58	1,68	1,75	1,85	1,93	2,11	2,46
1,37	1,63	1,69	1,75	1,86	1,97	2,17	2,55

Daí, a mediana é (1,75+1,79)/2=1,77

0,72	1,4	1,64	1,69	1,79	1,88	2,03	2,28
0,85	1,47	1,64	1,7	1,79	1,9	2,08	2,31
1,09	1,51	1,67	1,74	1,82	1,92	2,09	2,37
1,24	1,58	1,68	1,75	1,85	1,93	2,11	2,46
1,37	1,63	1,69	1,75	1,86	1,97	2,17	2,55

N'=20

O primeiro quartil será Q1=(1,63+1,64)/2=1,635

0,72	1,4	1,64	1,69	1,79	1,88	2,03	2,28
0,85	1,47	1,64	1,7	1,79	1,9	2,08	2,31
1,09	1,51	1,67	1,74	1,82	1,92	2,09	2,37
1,24	1,58	1,68	1,75	1,85	1,93	2,11	2,46
1,37	1,63	1,69	1,75	1,86	1,97	2,17	2,55

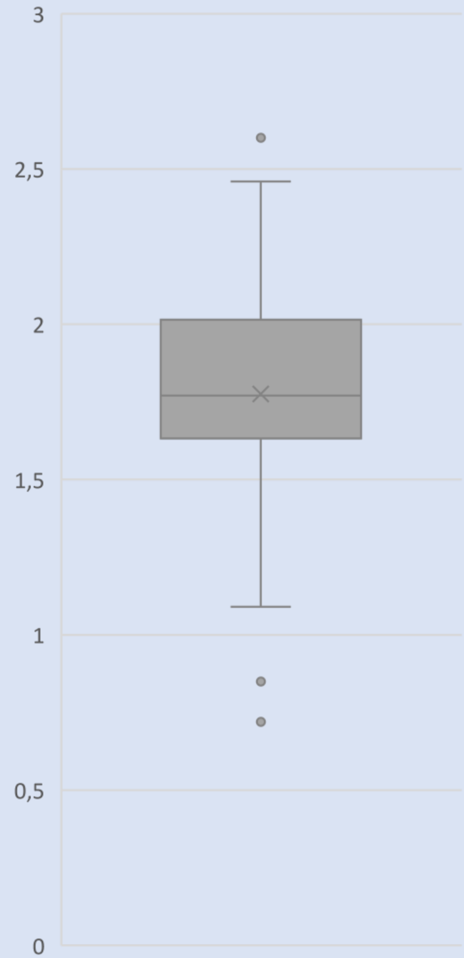
N'=20

O terceiro quartil será Q3=(1,97+2,03)/2=2,00

A distância Interquartílica será Q3-Q1=2,000-1,635=0,365

Daí os bigodes representarão os dados desde
Q1 –1,5* DIQ=1,635-1,5*0,365=1,0875
E até
Q3+1,5*DIQ=2,000+1,5*0,365=2,5475
Os valores fora do intervalo são considerados valores extremos

0,72	1,4	1,64	1,69	1,79	1,88	2,03	2,28
0,85	1,47	1,64	1,7	1,79	1,9	2,08	2,31
1,09	1,51	1,67	1,74	1,82	1,92	2,09	2,37
1,24	1,58	1,68	1,75	1,85	1,93	2,11	2,46
1,37	1,63	1,69	1,75	1,86	1,97	2,17	2,55


3
2,5
2
1,5
1
0,5
0