Amostragem Aleatória e Descrição de Dados - parte II

Marcos Oliveira Prates

2012/02

1. Diagrama de Ramo e Folhas
2. Distribuições de Frequência e Histograma
3. Box-plot ou Diagrama de Caixa
4. Gráficos Sequenciais Temporais
5. Gráficos de Probabilidade

Objetivos

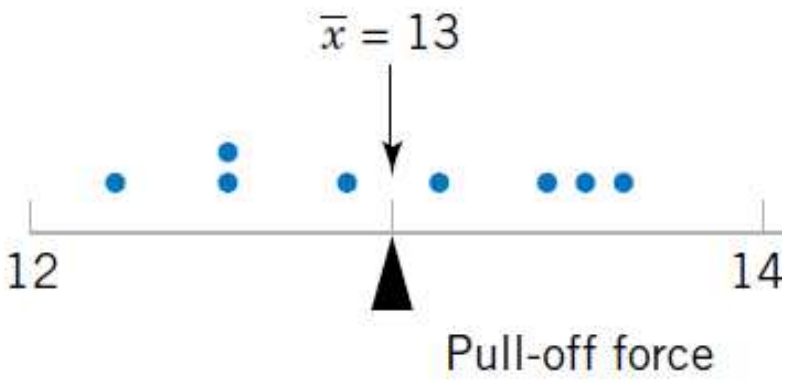
Ao final deste capítulo você deve ser capaz de:

- Construir e interpretar disposições gráficas dos dados.
- Iremos trabalhar aqui com os seguintes gráficos:
  - gráfico de ramo e folhas;
  - histograma;
  - diagrama de caixa ou box-plot;
  - gráfico de probabilidades.

Diagrama de Ramo e Folhas

- O **diagrama de pontos** é uma representação útil quando a amostra é pequena.
- Quando esse número é alto, outras representações são mais úteis.
- Pois nos fornecem mais informações sobre os dados.

Figure 6-1 The sample mean as a balance point for a system of weights.



- Os dados abaixo são a resistência à compressão de 80 corpos de prova.

Table 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

- Os dados não fornecem muita informação.
- Questões como:
  - "Qual a porcentagem de corpos com resistência abaixo de 120?"
- Não são fáceis de responder apenas olhando a tabela.
- Como a amostra é grande a construção do grafo de pontos é ineficiente.
- Devemos usar outro tipo de representação gráfica.

- Uma possibilidade é o **diagrama de ramos e folhas**.
- Ele é uma boa maneira de obter uma apresentação visual informativa dos dados.
- Pode ser usado quando os dados

$$x_1, \quad x_2, \ldots, \quad x_n$$

têm pelos menos dois dígitos.

Etapas para construir o Diagrama de Ramos e Folhas

- Divida cada número x_i em duas partes:
  - um **ramo** consistindo de um ou mais dígitos;
  - uma **folha** consistindo dos dígitos restantes.
- Liste os valores do ramo em uma coluna vertical.
- Ao lado do ramo, registre a folha de cada observação.
- Escreva as unidades para os ramos o folhas no grafo.

Exemplo:

- Considere que os dados representam a porcentagem de ítens defeituosos em um lote de pastilhas.
- Os dados variam entre 0 e 100.
- Se uma observação é 76% podemos dividir esse número:
  - ramo 7 e folha 6.
- Devemos escolher pouco ramos comparado com o número de observações.
- Senão perdemos muito informação.
- O ideal é escolher entre 5 e 20 ramos.

Exemplo:

- Considere os dados de resistência dos corpos de prova.
- Selecionamos os seguinte valores dos ramos:

$$7,\ 8, 9\ , \ldots , 24\ .$$

- Colocamos na última coluna a frequência do número de folhas associadas a cada ramo.

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)

- A maioria das resistências está entre 120 e 200.
- O valor central está entre 150 e 160.
- A distribuição é aproximadamente simétrica.

- Para algumas bases de dados:
  - pode ser interessante prover mais ramos.
- Podemos dividir o ramo 5 em dois novos:
  - 5L $\Rightarrow$ com as folhas 0,1,2,3,4;
  - 5U $\Rightarrow$ com as folhas 5,6,7,8,9.
- Isso dobra o número de ramos originais.
- Podemos ainda aumentar em quatro vezes o número de ramos originais.

Exemplo:

- A figura mostra o diagrama para dados de rendimento de um processo químico.

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

(a)

Stem	Leaf
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

(b)

Stem	Leaf
6z	1
6t	3
6f	4 5 5
6s	6
6e	
7z	0 1 1
7t	3
7f	5
7s	7
7e	8 8 9
8z	1
8t	3
8f	4 4
8s	7
8e	8 8
9z	
9t	2 3
9f	5
9s	
9e	

(c)

Diagrama de Ramo e Folhas usando Minitab

Character Stem-and-Leaf Display

Stem-and-leaf of Strength
N = 80 Leaf Unit = 1.0

1	7	6
2	8	7
3	9	7
5	10	1 5
8	11	0 5 8
11	12	0 1 3
17	13	1 3 3 4 5 5
25	14	1 2 3 5 6 8 9 9
37	15	0 0 1 3 4 4 6 7 8 8 8 8
(10)	16	0 0 0 3 3 5 7 7 8 9
33	17	0 1 1 2 4 4 5 6 6 8
23	18	0 0 1 1 3 4 6
16	19	0 3 4 6 9 9
10	20	0 1 7 8
6	21	8
5	22	1 8 9
2	23	7
1	24	5

Figure 6-6 A stem-and-leaf diagram from Minitab.

- Ordena as folhas de maneira crescente.

Diagrama de Ramo e Folhas usando R

```
> rf
 [1] 5.50 5.61 4.88 5.07 5.26 5.55 5.36 5.29 5.58 5.65 5.57 5.53 5.62 5.29 5.40
> stem(rf)

  The decimal point is 1 digit(s) to the left of the |

  48 | 8
  50 | 7
  52 | 6996
  54 | 003578
  56 | 125
```

Distribuições de Frequência e Histograma

Distribuição de Frequência

- É um sumário mais compactado dos dados em relação ao diagrama de ramo e folhas.
- Precisamos dividir a faixa dos dados em intervalos.
- Esses intervalos são chamados **intervalos de classe**.
- Sempre que possível esses intervalos devem ter a mesma largura.

- O número de intervalos depende do número de observações e de seu espalhamento.
- Uma regra prática é:
  - se temos um conjunto com n dados;
  - o número de intervalos de classe será aproximadamente

$$\sqrt{n}\,.$$

Exemplo:

- Considere os dados de resistência à compressão.
- O número total de dados é 80.
- Temos então
$$\sqrt{80} = 8,94 \approx 9$$
logo teremos 9 intervalos de classe.
- O maior e o menor valor desses dados são:
$$245 \quad \text{e} \quad 76.$$
- Pegamos o limite inferior de 70 (menor que 76) e o limite superior de 250 (maior que 245).
- A faixa de valores é 250-70=180.
- Cada intervalo de classe tem comprimento
$$\frac{180}{9} = 20.$$

Exemplo: (continuação)

- A tabela a seguir mostra a distribuição de frequência desses dados.
- A segunda linha mostra as **frequências relativas**.
- A terceira linha mostra as **frequências relativas acumuladas**.

Table 6-4 Frequency Distribution for the Compressive Strength Data in Table 6-2

Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

Histograma

- É uma disposição visual das frequências.
- Fornece uma impressão visual da forma da distribuição.
- Nos dá também informação sobre a tendência central e dispersão dos dados.

Construindo um histograma

- Marque os limites dos intervalos de classe no eixo horizontal.
- Marque e nomeie o eixo vertical com as frequências relativas.
- Acima de cada intervalo de classe desenhe um retângulo:
  - a altura do retângulo deve ser igual a frequência (ou frequência relativa).

- A seguir os histograma dos dados de resistência à compressão.
- A distribuição é simétrica, em forma de sino.
- A distribuição normal seria um modelo razoável.

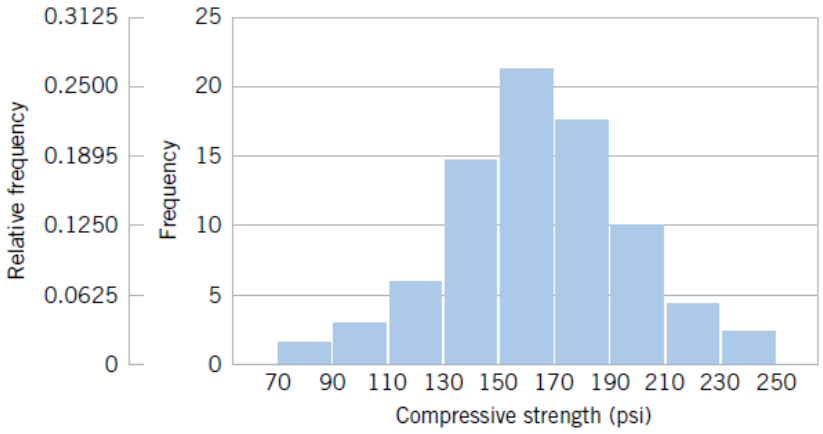


Figure 6-7 Histogram of compressive strength for 80 aluminum-lithium alloy specimens.

- Um histograma pode ter intervalos com larguras diferentes.
- Se existem *outliers* as observações se distribuem em poucos intervalos.
- Seria melhor usar mais intervalos na região em que estão a maioria dos dados.
- Nesse caso a **área** do retângulo é proporcional à frequencia

$$\text{altura do retângulo} = \frac{\text{frequência do intervalo}}{\text{largura do intervalo}} .$$

Histogramas usando Minitab

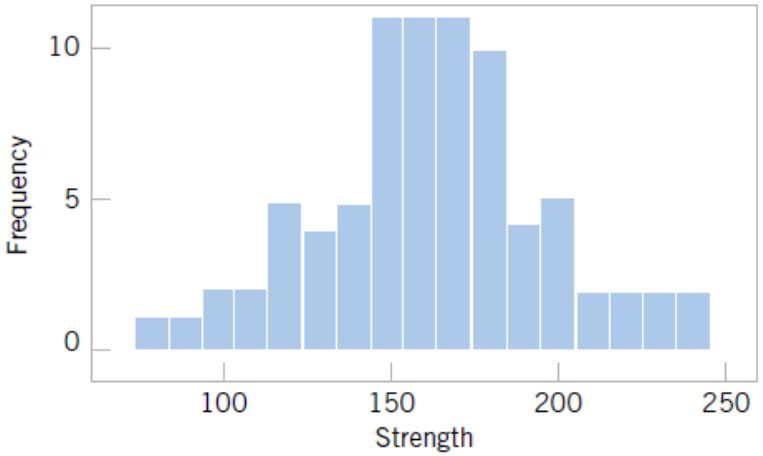


Figure 6-8 A histogram of the compressive strength data from Minitab with 17 bins.

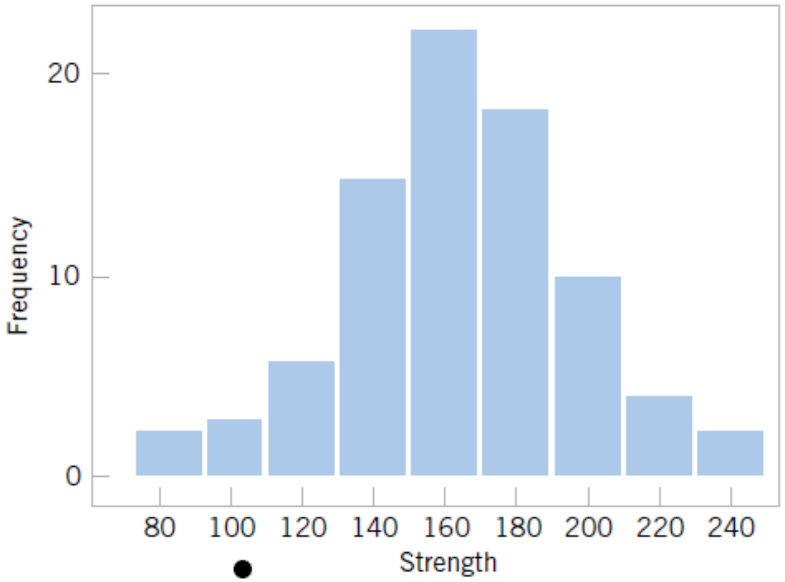


Figure 6-9 A histogram of the compressive strength data from Minitab with nine bins.

Histogramas usando R

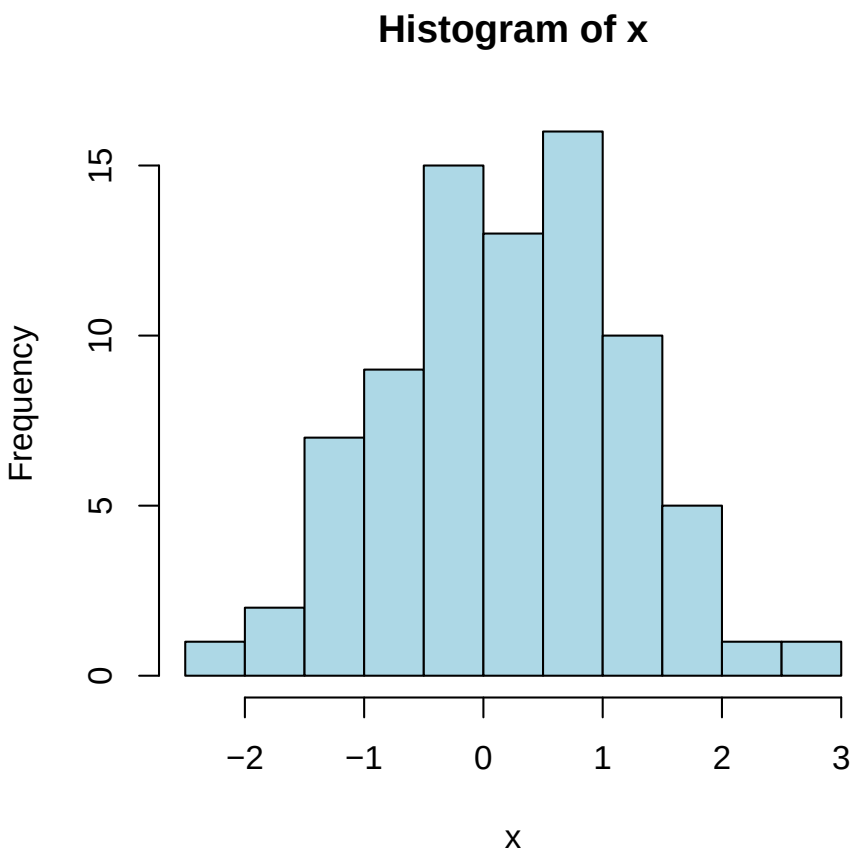


Gráfico de Frequência Cumulativa

- É uma variação do histograma.
- A altura do intervalo é definida como
  - o número de observações que é menor ou igual ao limite superior do intervalo.

- Veja abaixo o gráfico de frequência cumulativa para os dados de resistência à compressão.
- Existem aproximadamente 70 observações com força de compressão menor ou igual a 200.

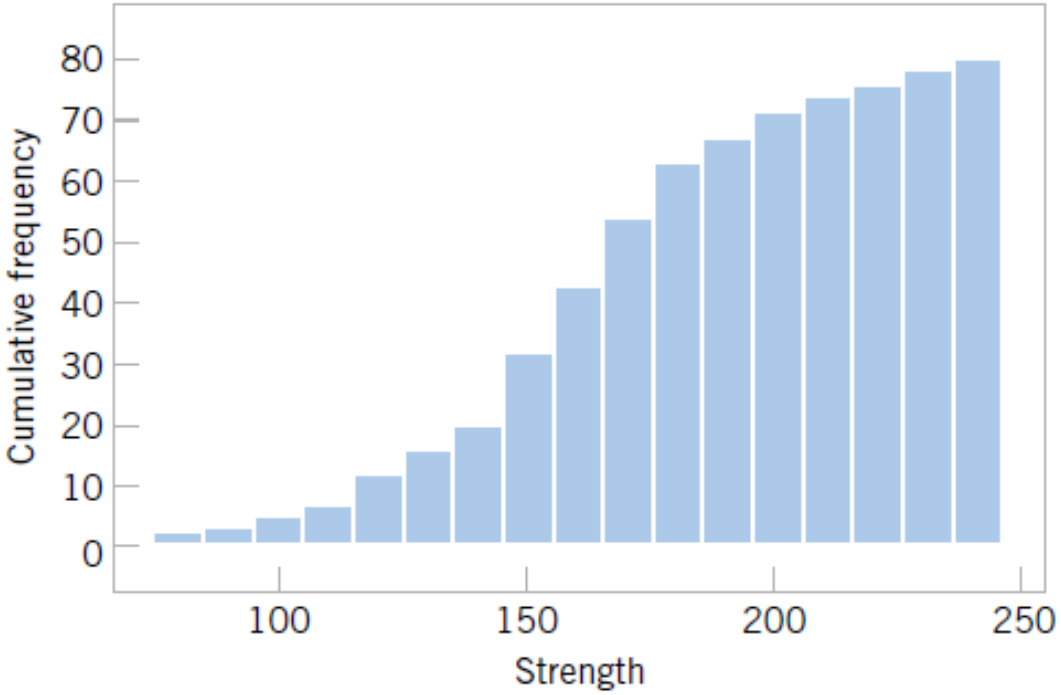


Figure 6-10 A cumulative distribution plot of the compressive strength data from Minitab.

- Quando o tamanho da amostra é grande:
  - o histograma indica a forma geral da distribuição.
- Essa distribuição pode ser:
  - simétrica;
  - assimétrica positiva;
  - assimétrica negativa.
- Essa classificação vai depender da posição relativa entre:
  - a mediana ($\tilde{x}$) e a média ($\bar{x}$).

- A figura abaixo ilustra os tipos de assimetria.

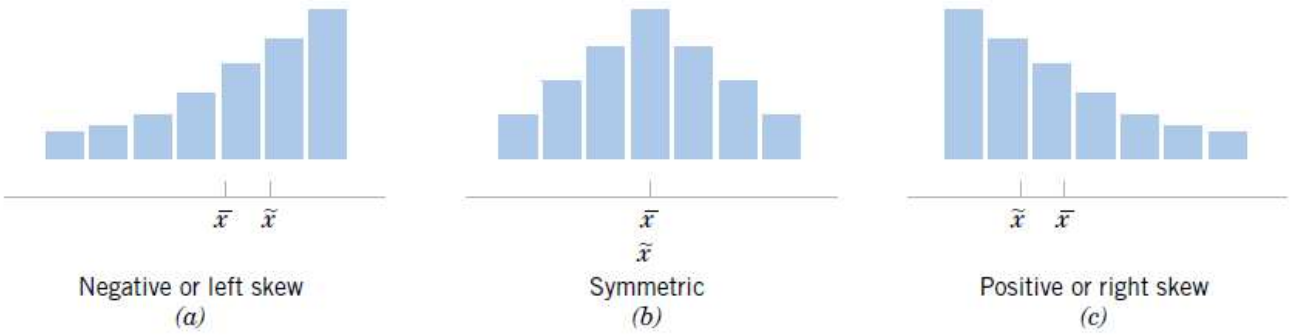
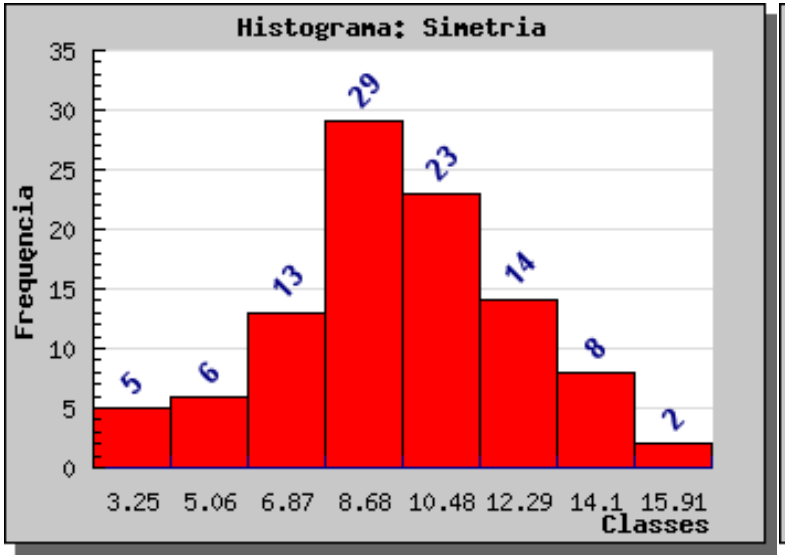
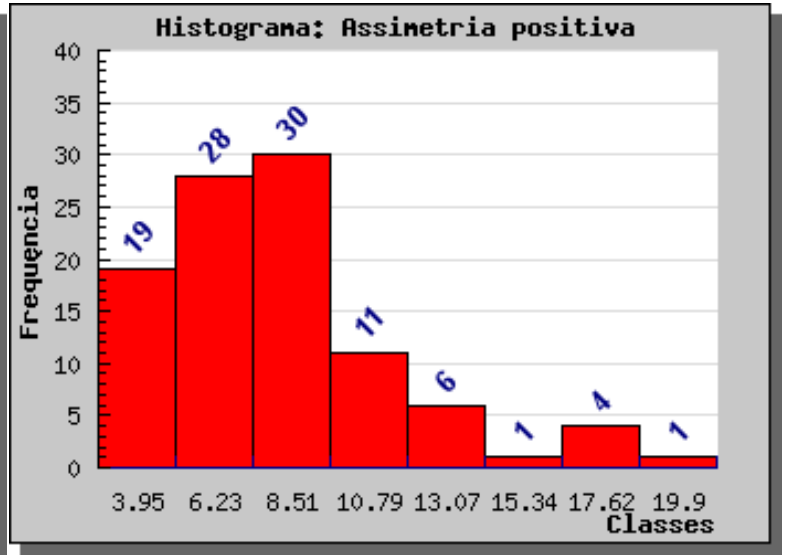
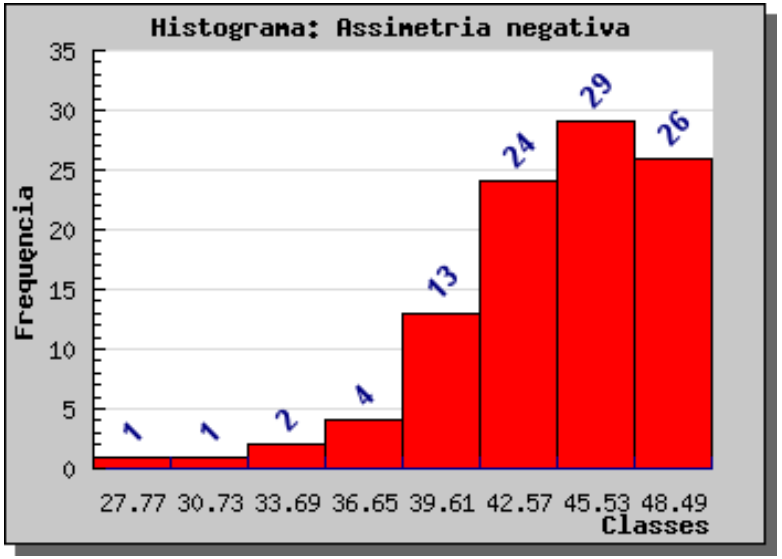


Figure 6-11
Histograms for symmetric and skewed distributions.


Histograma: Simetria
Frequencia
35
30
25
20
15
10
5
0
5
6
13
29
23
14
8
2
3.25 5.06 6.87 8.68 10.48 12.29 14.1 15.91
Classes


Histograma: Assimetria positiva
Frequencia
40
35
30
25
20
15
10
5
0
19
28
30
11
6
1
4
1
3.95 6.23 8.51 10.79 13.07 15.34 17.62 19.9
Classes


Histograma: Assimetria negativa
Frequencia
35
30
25
20
15
10
5
0
1
1
2
4
13
24
29
26
27.77 30.73 33.69 36.65 39.61 42.57 45.53 48.49
Classes

Box-plot ou Diagrama de Caixa

- O box-plot descreve uma variável.
- Ele apresenta várias características dessa variável:
  - centro;
  - dispersão;
  - assimetria;
  - identificação de observações não usuais ou *outliers*.

Elementos do box-plot:

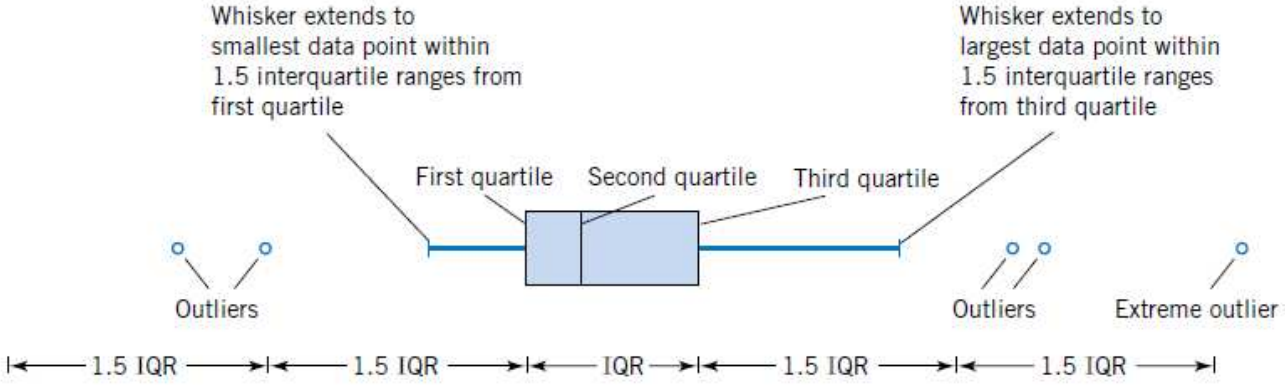


Figure 6-13 Description of a box plot.

Exemplo:

- Box-plot para os dados de resitência a compressão.

- A distribuição é razoávelmente simétrica.
- Há dois *outliers*.
- Os limites da linha são dados por

$$q_3 + 1,5IQR =$$

$$181 + 1,5(181 - 143,5) = 237,25$$

$$q_1 - 1,5IQR =$$

$$143,5 - 1,5(181 - 143,5) = 87,25$$

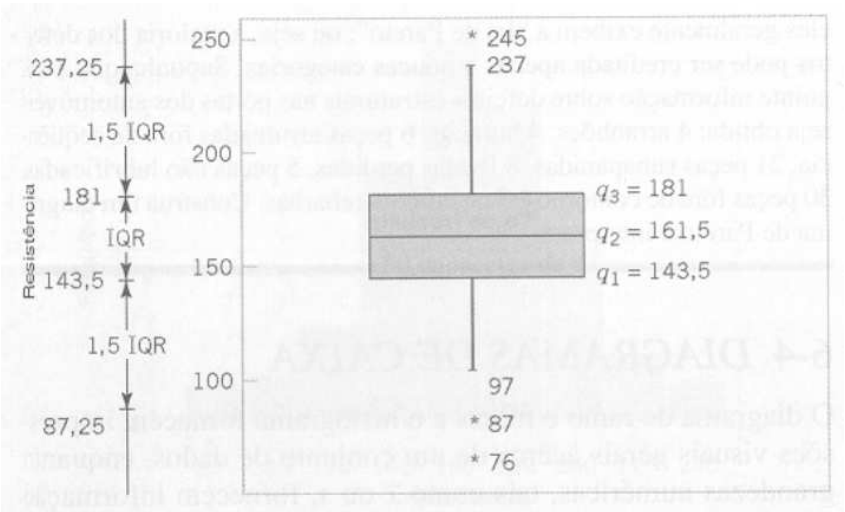


Figura 6-14 Diagrama de caixa para os dados de resistência à compressão da Tabela 6-2.

- O box-plot é muito útil para comparação de grupos.
- Ele tem alto impacto visual.
- Permite comparar rapidamente diferenças como:
  - localização;
  - simetria;
  - dispersão.

Exemplo:

- Índice de qualidade de fabricação em três plantas distintas.

- A qualidade na primeira planta é maior.
- A assimetria na terceira planta maior.
- A variabilidade na segunda planta é maior.
- Nenhuma apresenta *outliers*.

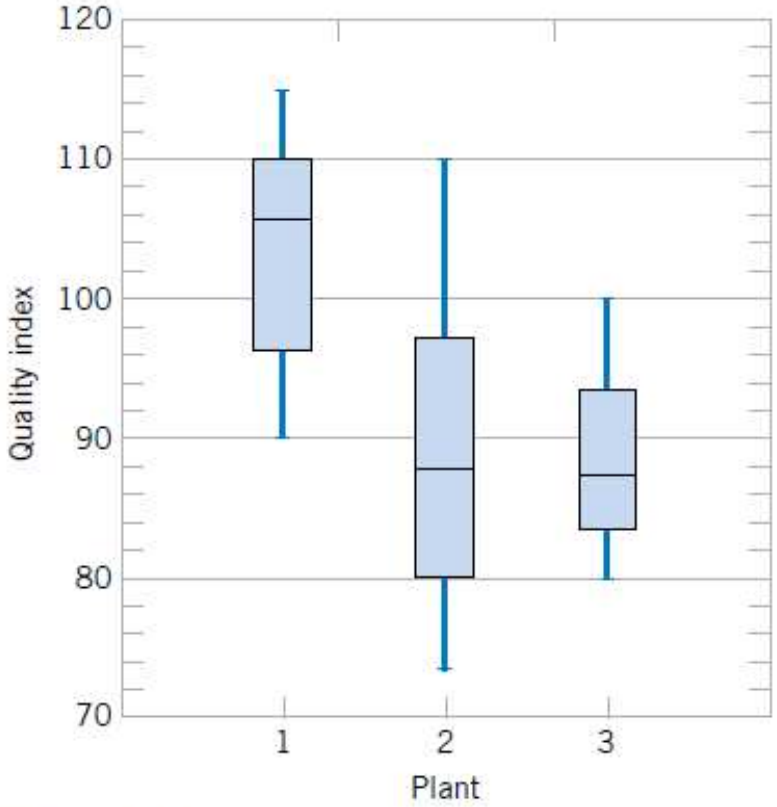


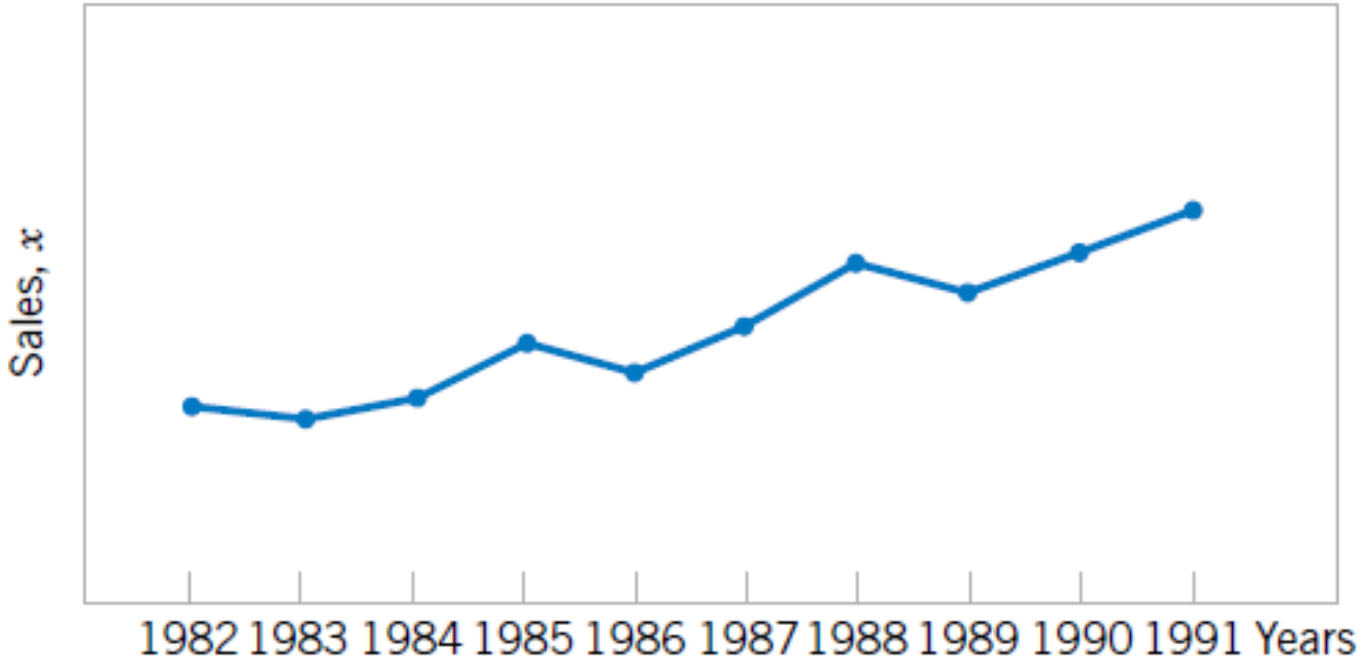
Figure 6-15 Comparative box plots of a

Gráficos Sequenciais Temporais

- O tempo é um fator importante para variabilidade dos dados.
- Uma **série temporal** ou **sequência temporal** é
  - um conjunto de dados registrados na ordem que ocorreram.
- Em um **gráfico de séries temporais**:
  - o eixo vertical denota o valor da variável;
  - o eixo horizontal denota o tempo.
- Podemos identificar características como:
  - tendências;
  - ciclos.

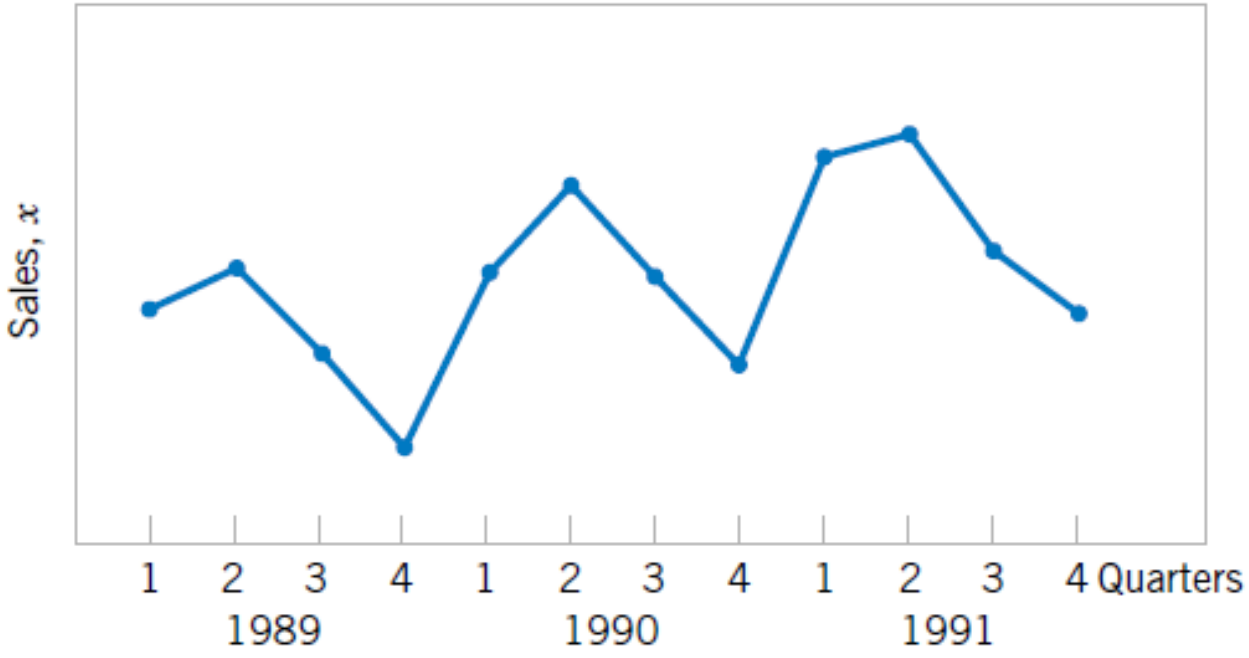
Exemplo:

- Vendas anuais em uma companhia nos últimos 10 anos.
- As vendas apresentam uma tendência crescente.
- Existe certa variabilidade em torno dessa tendência.



Exemplo:

- Vendas da companhia nos últimos três anos, registradas por trimestre.
- As vendas apresentam variabilidade cíclica.
- As vendas no primeiro e segundo semestre maiores que no terceiro e quarto.



Gráficos de Probabilidade

- Como sabemos se uma distribuição é boa para modelar os dados?
- Geralmente precisamos supor que as observações seguem uma distribuição.
- A distribuição pode dar informação sobre o mecanismo físico de geração dos dados.
- Exemplo: tempo de falha segue uma distribuição exponencial
  - a taxa de falha é constante ao longo do tempo.

- O histograma dá uma idéia da distribuição dos dados:
  - apenas se amostra é grande.

Gráfico de Probabilidade

- Determina se os dados obedecem uma distribuição hipotética.
- Baseado no exame visual dos dados.
- Se os pontos caem em torno de uma reta a distribuição é adequada.

Exemplo:

- Dez observações sobre o tempo de vida de baterias são realizadas.

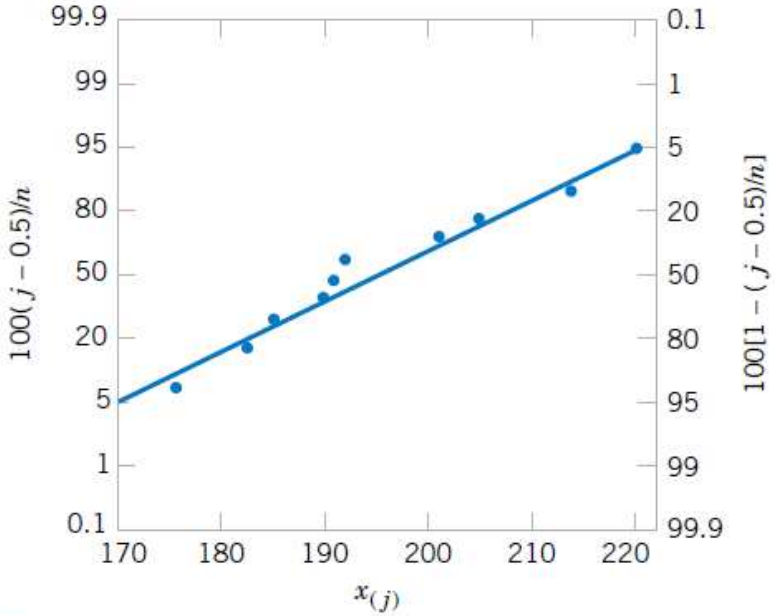


Figure 6-19 Normal probability plot for battery life.

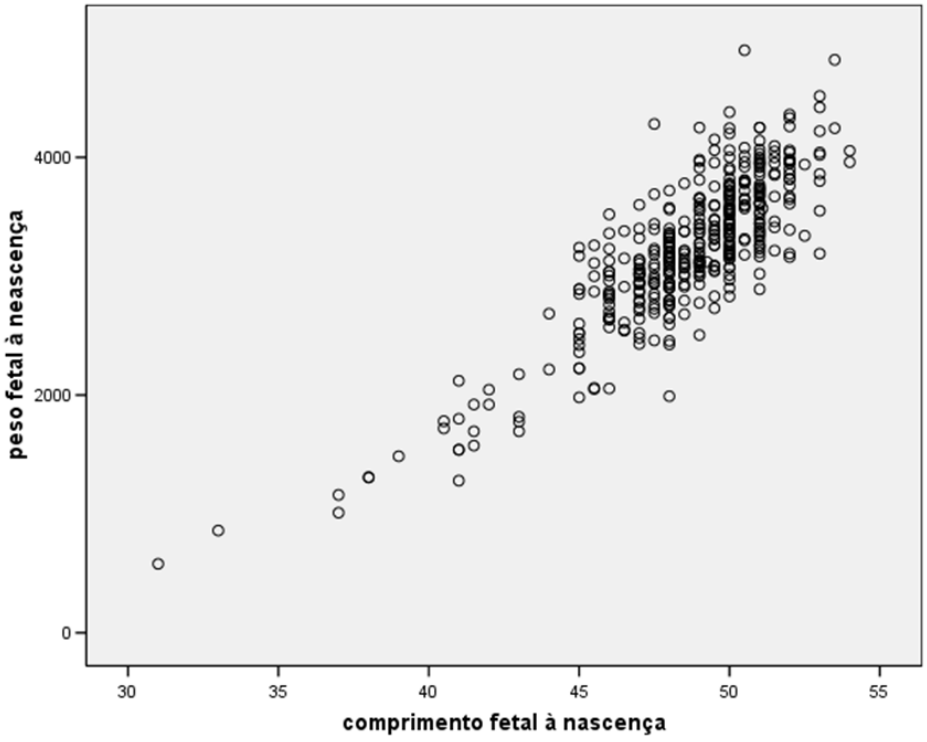
- Os pontos caem aproximadamente em torno da linha.

Gráfico de Dispersão

- Representa duas variáveis.
- O objetivo é verificar a relação entre elas.
- Se uma variável aumenta (diminui) quando a outra aumenta (diminui):
  - correlação positiva.
- Se uma variável diminui (aumenta) quando a outra aumenta (diminui):
  - correlação negativa.

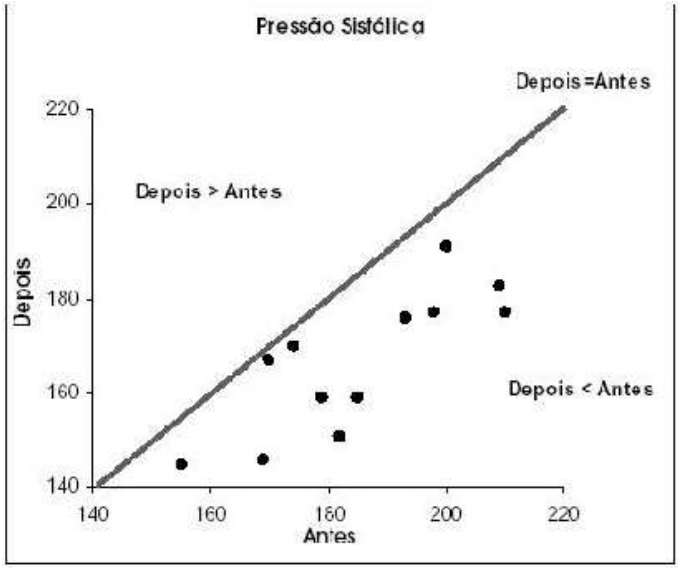
Exemplo:

- Pesos e comprimentos de 414 recém-nascidos.

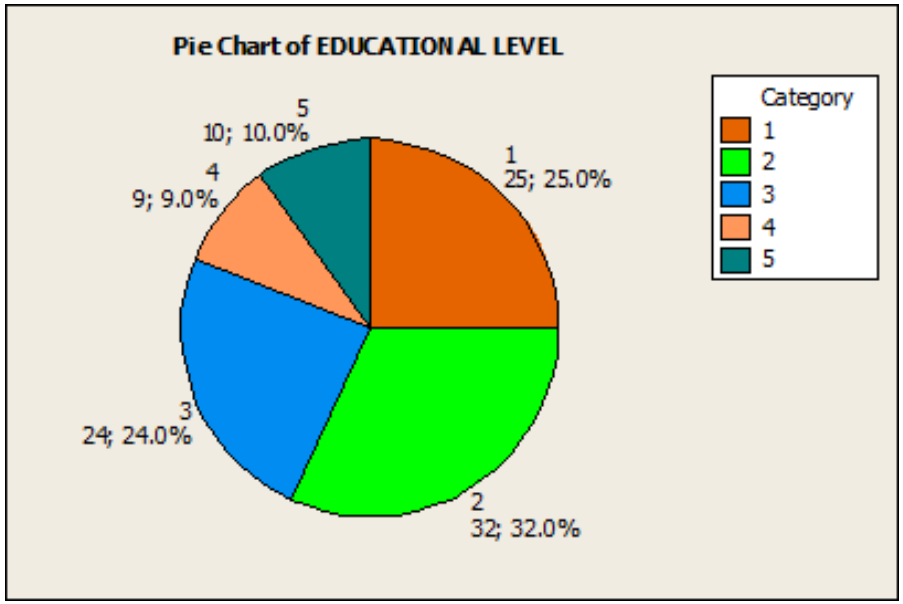


Exemplo:

- Um remédio é utilizado para reduzir a pressão sistólica.
- 12 pacientes têm a pressão medida antes e depois da medicação.



- Todos pontos abaixo da curva $\Rightarrow$ para todos pacientes a pressão diminuiu.

Gráfico de setores ou de pizza

- Diagrama circular.
- Mostra a importância relativa das proporções.
- Trabalha com porcentagens.
- Cada setor tem a área proporcional a porcentagem naquela categoria