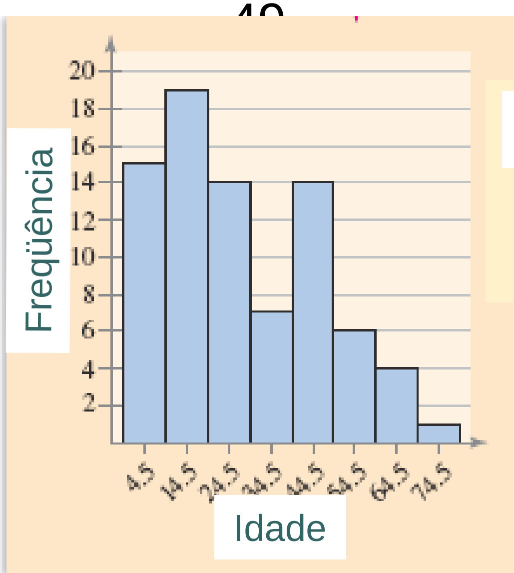
Estatística Descritiva

- Como construir uma distribuição de freqüências.
- Como construir gráficos de freqüências.
- Como encontrar medidas de tendência central.
- Como encontrar medidas de variabilidade.
- Como encontrar separatrizes

Motivação

- Idades de uma amostra com 80 residentes em Alaska:
  - 25, 5, 18, 12, 60, 44, 24, 22, 2, 7, 15, 39, 58, 53, 36, 42, 16, 20, 1, 5, 39, 51, 44, 23, 3, 13, 37, 56, 58, 13, 47, 23, 1, 17, 39, 13, 24, 0, 39, 10, 41, 1, 48, 17, 18, 3, 72, 20, 3, 9, 0, 12, 33, 21, 40, 68, 25, 40, 59, 4, 67, 29, 13, 18, 19, 13, 16, 41, 19, 26, 68, 49, 5, 26, 49, 26, 45, 41, 19, 49



$$\text{Média} = \frac{0+0+1+1+1+\cdots+67+68+68+72}{80} = \frac{2226}{80}$$

$$\text{Range} = 72 - 0 = 72 \text{ anos}$$

Distribuição de Freqüências

Dados Quantitativos

- Uma tabela de classes ou intervalos de valores de uma amostra com um número total de observações em cada classes.

Classe	Freqüência
1–5	5
6–10	8
11–15	6
16–20	8
21–25	5
26–30	4

Etapas para construção de uma distribuição de freqüências

- 1. Decida o tamanho do número de intervalos. Um bom tamanho é $[1, \sqrt{n}]$ onde n é o tamanho da amostra.
- 2. Determine a amplitude de cada intervalo. Divida o range dos valores pelo tamanho do número de intervalos. Arredonde até o próximo número.
- 3. Calcule os limites das classes. O valor mínimo dos dados pode ser o limite inferior da primeiro intervalo. Adicione o range para formar o limite máximo deste intervalo e obter os próximos intervalos. Os intervalos não podem sobrepor.
- 4. Conte as freqüências de cada classe.

Exemplo: Tempo (em min) gasto na Internet

- Conjunto de dados amostrais: lista do número de minutos de 50 assinantes.
  - 50 40 41 17 11 7 22 44 28 21 19 23 37 51 54 42 88 41 78 56 72 56 17 7 69 30 80 56 29 33 46 31 39 20 18 29 34 59 73 77 36 39 30 62 54 67 39 31 53 44

Construindo a distribuição de freqüências

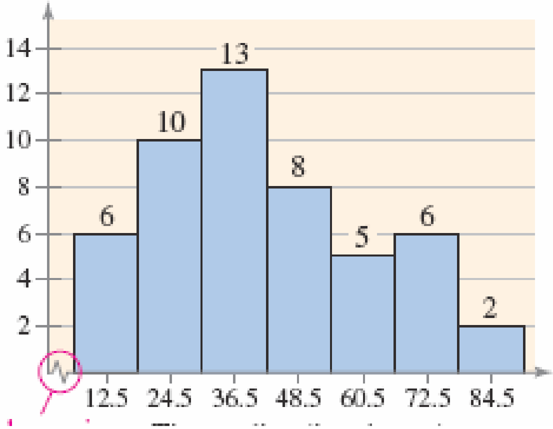
- 1. O número de intervalos é 7.
- 2. Os valores mínimo e máximo são 7 e 88, respectivamente. Logo a amplitude total é 81. A amplitude dos intervalos é 12.
- 3. Os limites inferior e superior do primeiro intervalo são 7 e 18, respectivamente.
- 4. Estabeleça a freqüência de cada classe.

Distribuição de freqüências

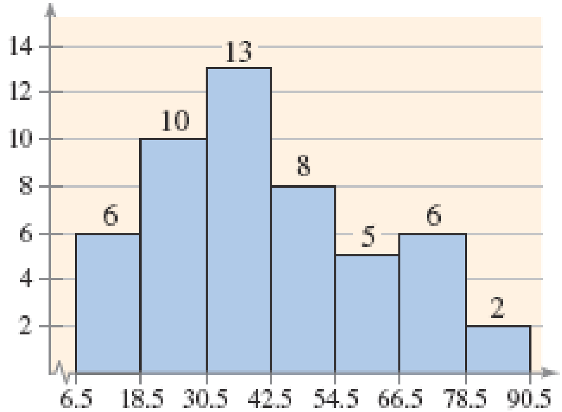
Classe	Freqüência	Ponto Médio	Freqüência Relativa	Freqüência Acumulado
7–18	6	12.5	0.12	6
19–30	10	24.5	0.2	16
31–42	13	36.5	0.26	29
43–54	8	48.5	0.16	37
55–66	5	60.5	0.1	42
67–78	6	72.5	0.12	48
79–90	2	84.5	0.04	50
	$\sum f = 50$		$\sum \frac{f}{n} = 1$	

Gráficos da distribuição de freqüências

Histograma usando os pontos médios



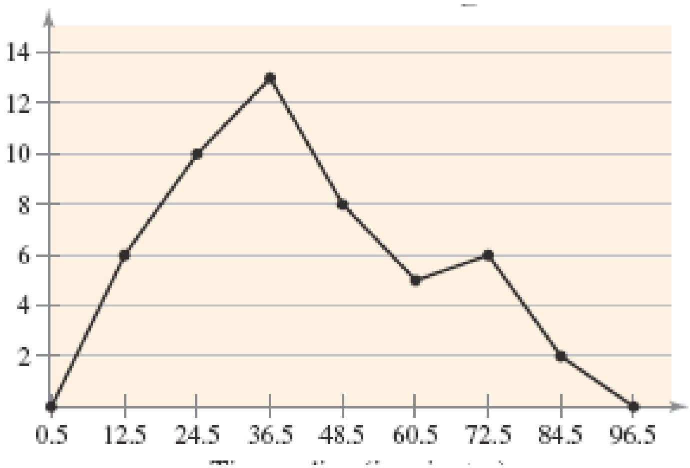
Histograma usando as fronteiras



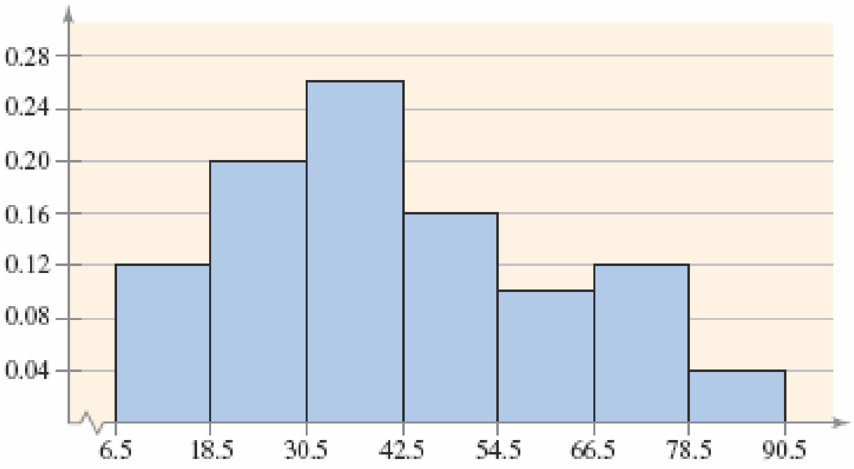
Classe	Fronteiras	Freqüência
7–18	6.5–18.5	6
19–30	18.5–30.5	10
31–42	30.5–42.5	13
43–54	42.5–54.5	8
55–66	54.5–66.5	5
67–78	66.5–78.5	6
79–90	78.5–90.5	2

Gráficos da distribuição de freqüências

Polígono de freqüências

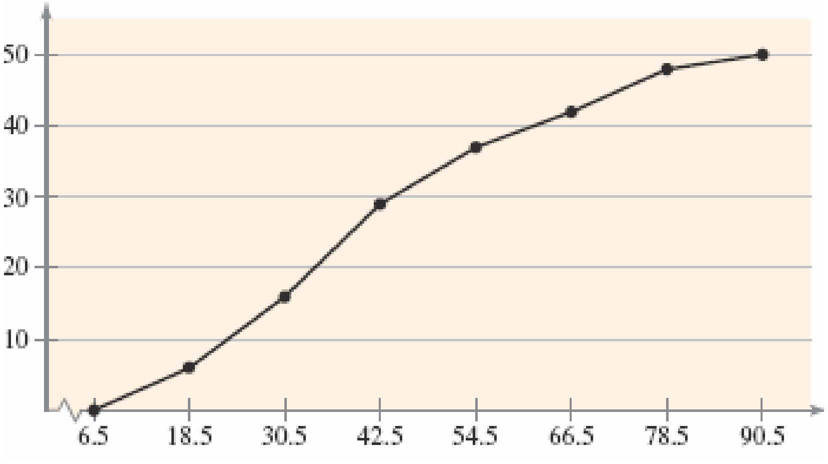


Histograma de freqüência relativa



Gráficos da distribuição de freqüências

Ogiva

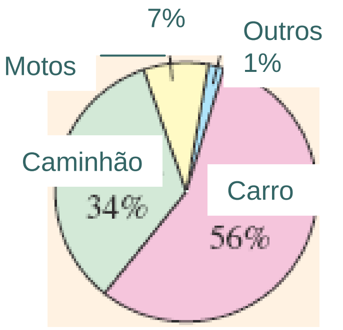


Limites Superiores	f	Freqüência Acumulada
18.5	6	6
30.5	10	16
42.5	13	29
54.5	8	37
66.5	5	42
78.5	6	48
90.5	2	50

Gráficos de dados qualitativos

	f	Freqüência Relativa	Angulo
Carros	20,269	0.56	202°
Caminhões	12,260	0.34	122°
Motos	3,067	0.08	29°
Outros	610	0.02	7°

Gráfico de Pizza



Gráficos de dados qualitativos

Causas de redução de ativos

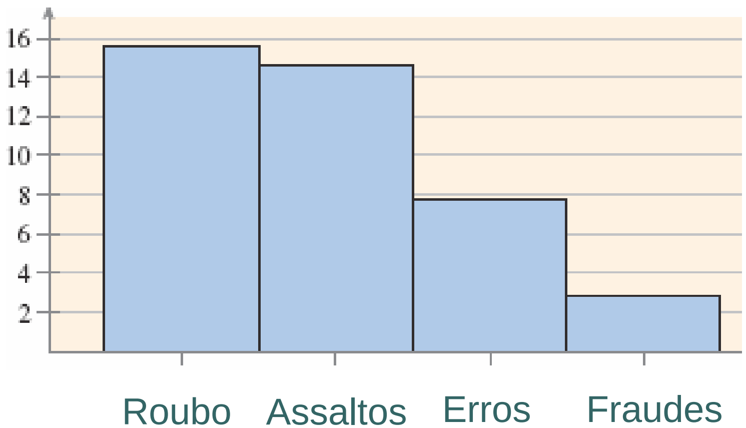
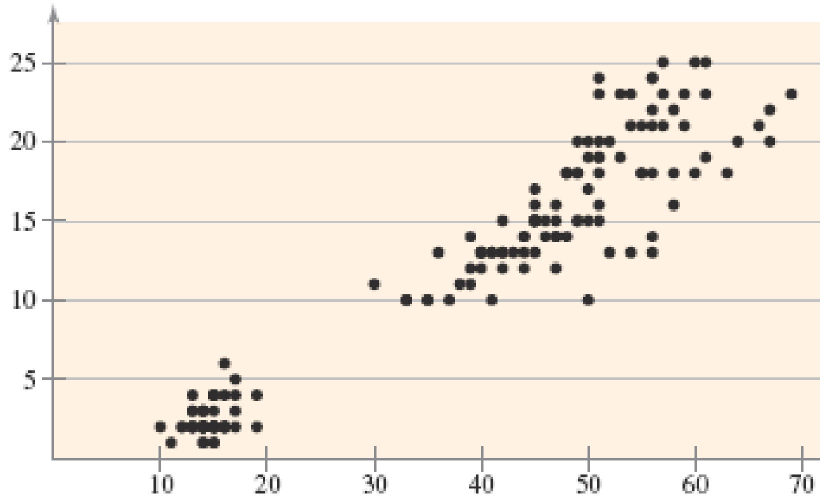


Gráfico de barras verticais

Gráfico de dados emparelhados

Conjunto de dados Iris

Largura da pétala

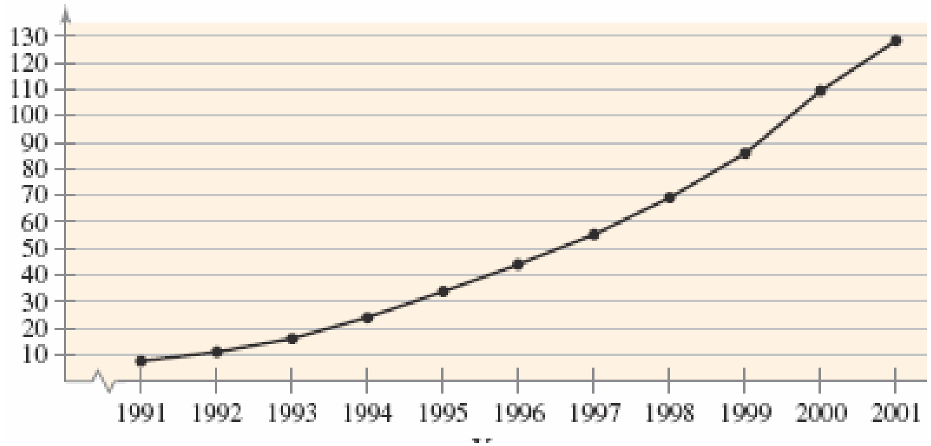


Comprimento da pétala

Gráfico de série temporal

Número de assinantes de telefones celulares (em milhões)

Ano	Assinantes	Conta Média
1991	7.6	72.74
1992	11.0	68.68
1993	16.0	61.48
1994	24.1	56.21
1995	33.8	51.00
1996	44.0	47.70
1997	55.3	42.78
1998	69.2	39.43
1999	86.0	41.24
2000	109.5	45.27
2001	128.4	47.37



Medidas de tendência central

- Média
  - Amostra $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - População $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Mediana
  - Valor que divide o conjunto em duas partes de iguais. Se o tamanho do conjunto é par , a mediana é a média entre os dois elementos mais centrais.
- Moda
  - Valor que tem a maior freqüência
- Em uma distribuição normal a média, a mediana e a moda são iguais.

Comparação entre Média, Moda e Mediana

- Vantagens e desvantagens:
  - Média: funciona bem com muitos métodos estatísticos
  - Mediana: costuma ser uma boa escolha se há alguns valores extremos.
  - Moda: apropriada para dados ao nível nominal

Exemplo

Idades em uma classe						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Valor aberrante

- Média= 23,75
- Mediana=21,5
- Moda= 20

Média ponderada

Fonte	Nota x	Peso w	xw
Média testes	86	0.50	43.0
Exame do meio	96	0.15	14.4
Laboratório	82	0.20	16.4
Trabalho de casa	98	0.10	9.8
	100	0.05	5.0
		$\sum w = 1$	$\sum (x \cdot w) = 88.6$

$$\bar{x} = \sum_{i=1}^{n} x_i w_i = 88{,}6$$

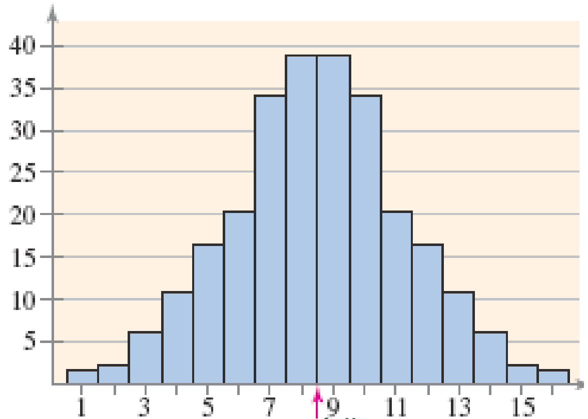
Média de dados agrupados

x	f	x.f
12.5	6	75.0
24.5	10	245.0
36.5	13	474.5
48.5	8	388.0
60.5	5	302.5
72.5	6	435.0
84.5	2	169.0
	$n = 50$	$\sum = 2089.0$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i f_i = 41{,}80$$

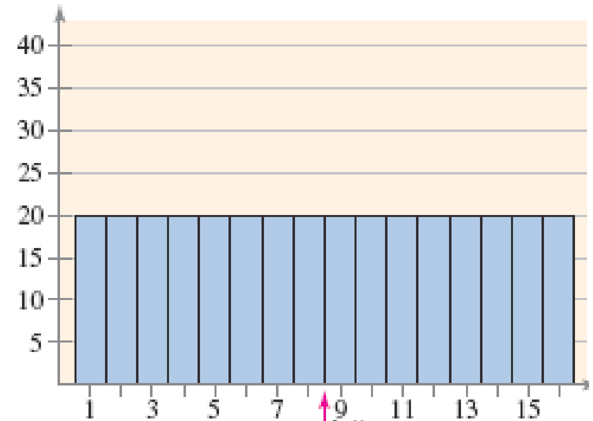
Aspectos das distribuições

Simétrica



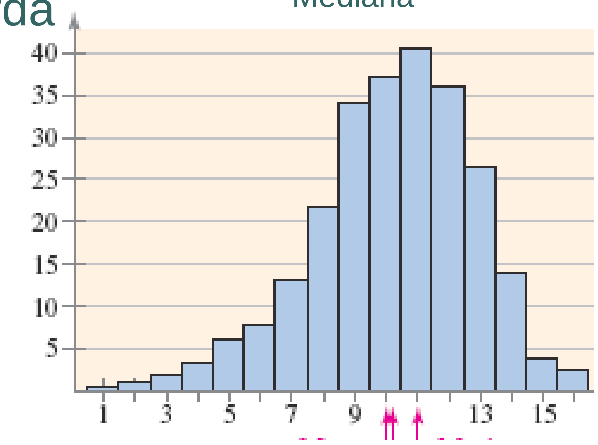
Média
Moda
Mediana

Uniforme

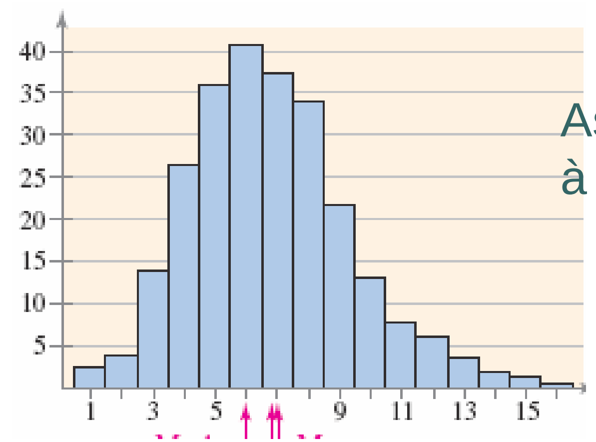


Média
Mediana

Assimétrica
à esquerda



Media < Mediana < Moda



Assimétrica
à direita

Moda < Mediana < Média

Aspectos das distribuições

- Assimetria *Sk*: mede o grau de deformação . Assume valores entre –1 e 1.

$$Sk = \frac{\bar{x} - Mo}{S}$$

  - onde Mo é a moda.
- Curtose: mede o grau de achatamento ou afilamento

$$K = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})^4}{S^4}$$

Usando Regra Empírica

- Usada para determinar a porcentagem de valores que precisam estar dentro de um número especificado de desvios-padrões da média.
- Para dados que tem uma distribuição na forma de um sino:
  - Aproximadamente 68% dos valores dos dados estarão dentro de um desvio padrão da média.
  - Aproximadamente 95% dos valores dos dados estarão dentro de dois desvios padrões da média.
  - Aproximadamente 99% dos valores dos dados estarão dentro de três desvios padrões da média.

Assimetria e Curtose

- Sk = 0 (Simétrica)

  Sk > 0 (Assimetria positiva)

  Sk < 0 (Assimetria negativa
  - Menores que 0,15 ➔ distribuição é simétrica
  - 0,15<IA<1,0 ➔ Distribuição é moderadamente assimétrica
  - Maior que 1,0 ➔ Distribuição é fortemente assimétrica
- K = 3 (Mesocúrtica) (Distribuição Normal)

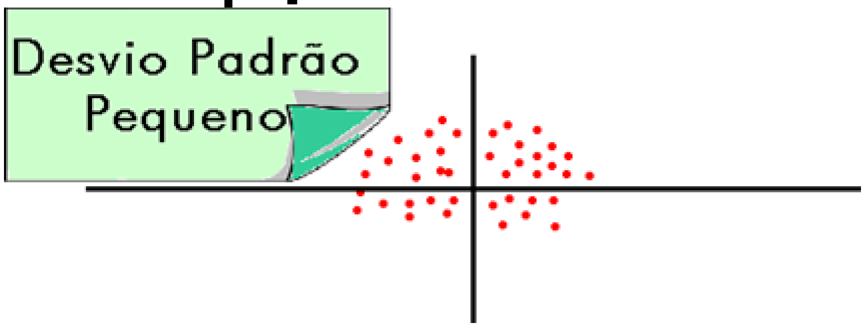
  K > 3 (Leptocúrtica)

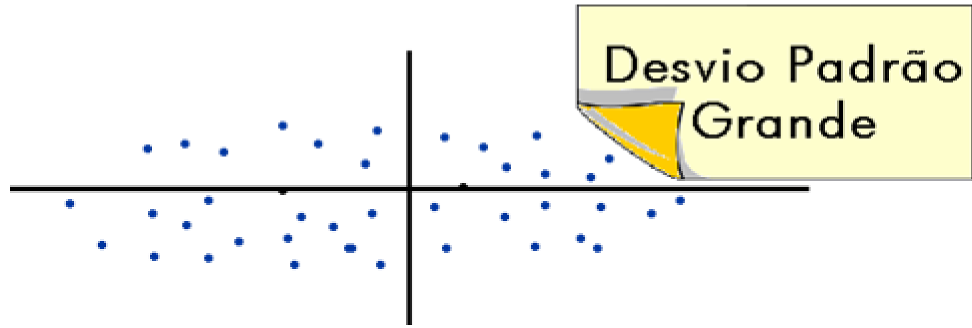
  K < 3 (Platocúrtica)

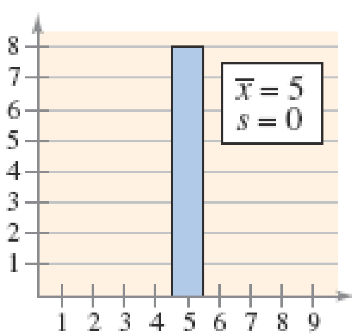
Medidas de Variabilidade

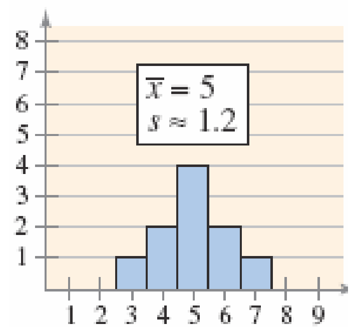
- Amplitude total
  - Diferença entre o maior valor e o menor valor.
- Variância
  - Populacional $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$
  - Amostral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Desvio padrão
  - Populacional $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2}$
  - Amostral $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- Coeficiente de variação $\frac{S}{\overline{x}} \times 100$

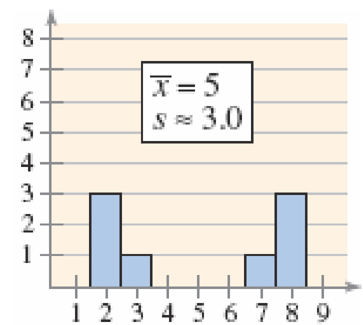
Interpretando o desvio padrão











- Quanto mais espalhados estiverem os dados maior será o desvio padrão

Desvio padrão de dados agrupados

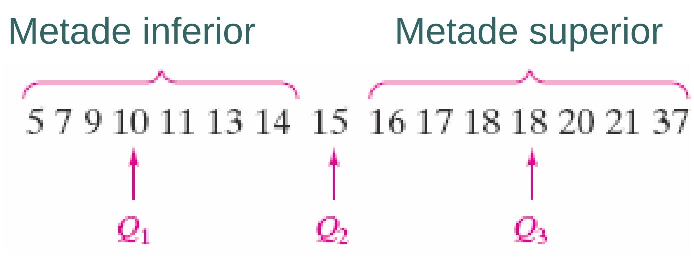
Distribuição de número de crianças
em 50 domicílios

x	f	xf
0	10	0
1	19	19
2	7	14
3	7	21
4	2	8
5	1	5
6	4	24
	$\Sigma = 50$	$\Sigma = 91$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 f_i} = 1{,}7$$

Medidas de posição

- Os três quartis Q_1, Q_2 e Q_3 dividem ao conjunto de dados em quatro partes iguais: 25% ficam dentro ou abaixo de Q_1, 50% ficam dentro ou abaixo de Q_2 e 75% ficam dentro ou abaixo de Q_3.



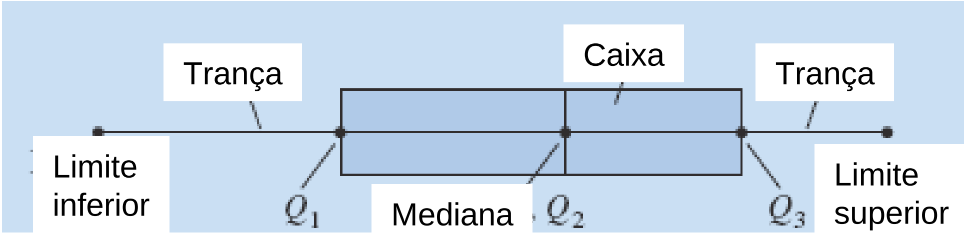
- Amplitude interquartílica: é diferença entre Q_3 e Q_1.
  - Fornece uma idéia de quanto 50% centrais (médios) dos dados variam.

Decis e percentis

- Decis: divide o conjunto de dados em dez partes iguais.
- Percentis: divide o conjunto de dados em cem partes iguais.
  - São freqüentemente usados na educação e nos campos relacionados a saúde para indicar como um indivíduo se compara com outros em um determinado grupo. Pontuações em testes e medidas de crescimento infantil são freqüentemente expressos em percentis.

Box PLot

- Um gráfico que permite identificar os pontos aberrantes em uma amostra e realça características importantes.



- Etapas:
  - 1. Obtenha Q_1, Q_2, Q_3 Q_3-Q_1. Calcule os limites inferior: LI=Q1 – 1,5×(Q3-Q1) e LS= Q1+1,5×(Q3-Q1). Os dados fora do intervalo [LI,IS] são considerados fora da curva.
  - 2. Construa uma escala total que abrange todos os dados.
  - 3. Plote os cincos números acima da escala horizontal.
  - 4. Faça uma caixa acima de Q_1 a Q_3 e trace uma reta vertical passando por Q_2.
  - 5. Faça as tranças