ESTATÍSTICA

Michelle Hanne Soares de Andrade

michellehanne.andrade@gmail.com

Estatística Descritiva

Box-Plot (Diagrama de caixa)

- O **boxplot**, ou diagrama de caixa, é um gráfico que capta importantes aspectos de um conjunto de dados através do seu resumo dos cinco números (valor mínimo, primeiro quartil, segundo quartil, terceiro quartil e valor máximo). Bem como, o centro, dispersão, desvio da simetria e identificação das observações que estão longe do centro dos dados (*outliers*).
- **O gráfico e formado por uma caixa construída paralelamente ao eixo da escala dos dados (pode ser horizontal ou vertical).**
- **Esse *box* vai desde o primeiro quartil até o terceiro quartil e nela traça-se uma linha na posição da mediana.**

Exemplo 1 Box-Plot

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Valor	3,0	3,5	4,5	5,0	5,0	5,5	6,5	6,5	6,5	7,5	7,6	7,9	8,0	8,0	9,0	9,5	10,0	15,0

A mediana divide o conjunto em duas partes, cada uma com 9 observações. A mediana será, então, a média dos dois valores centrais:

$$Q2 = \frac{6,5 + 7,5}{2} = 7,0$$

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Valor	3,0	3,5	4,5	5,0	5,0	5,5	6,5	6,5	6,5	7,5	7,6	7,9	8,0	8,0	9,0	9,5	10,0	15,0

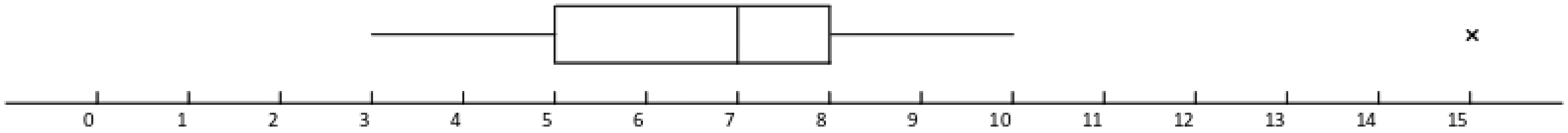
Exemplo 1 Box-Plot

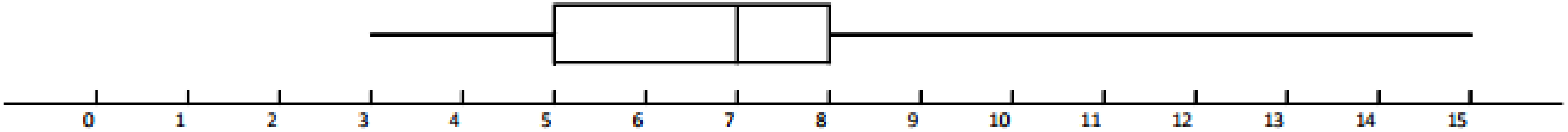
- **O cálculo do primeiro e do terceiro quartis:**
  - Calcular as medianas das duas metades – o primeiro quartil é a mediana da metade inferior e o terceiro quartil é a mediana da metade superior.

$$Q1 = 5{,}0$$

$$Q3 = 8{,}0$$

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Valor	3,0	3,5	4,5	5,0	5,0	5,5	6,5	6,5	6,5	7,5	7,6	7,9	8,0	8,0	9,0	9,5	10,0	15,0





Exemplo 2 Box-Plot

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13
Valor	15	17	18	19	19	20	25,0	26	26	28	30	32	42

Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13
Valor	15	17	18	19	19	20	25,0	26	26	28	30	32	42

$$Q2 = 25{,}0$$

$$Q1 = \frac{18 + 19}{2} = 18{,}5$$

$$Q3 = \frac{28 + 30}{2} = 29$$

Exemplo 3 - Dispositivo Prático

- Suponha o conjunto de números 1, 2, 3, 4 e 5.
- Organize os dados em ordem numérica (Um gráfico de frequência cumulativa facilita o trabalho, mas não é essencial).

  1 - Encontre a mediana**. O número, no centro de dados → 3**

  2 - Encontre os quartis superiores e inferiores, encontrando as medianas dos números maiores do que a mediana (quartil superior) e os números inferiores a mediana (quartil inferior).

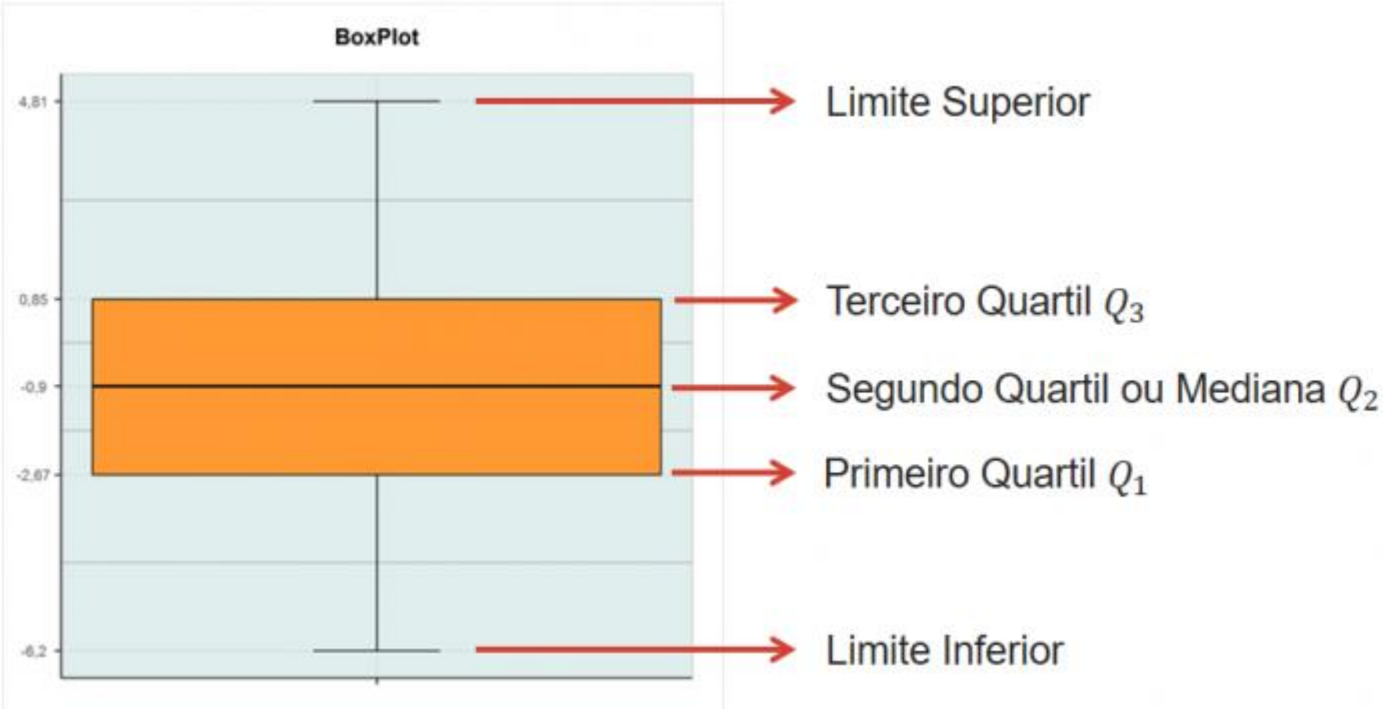
  3 - Marque seus valores discrepantes, ou os extremos. Estes são os pedaços maiores e menores de dados e devem ser marcados com um ponto (ou uma pequena linha vertical) praticamente em direção ao centro da sua caixa. Neste caso, o extremo inferior é 1 e o extremo superior é 5.

Exemplo 3 - Dispositivo Prático

- Desenhe uma linha guia. Esta deve ser longa o suficiente para conter todos os seus dados. Os números devem ser colocados no gráfico em intervalos regulares.
- Marque as suas medianas e quartis. Desenhe uma linha a partir desses pontos na altura que você deseja que sua caixa tenha. Conecte os topos para fazer a caixa.

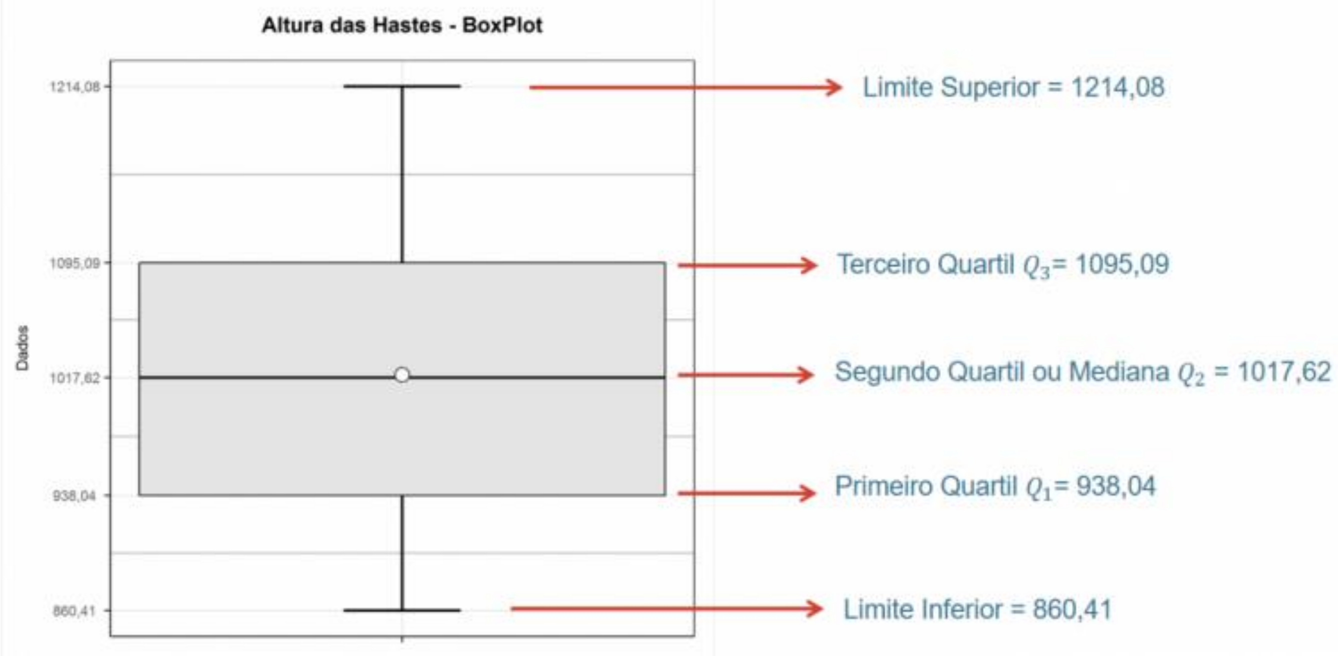
Box-Plot

- Exemplo



Box-Plot – Exemplo 1

- Na Tabela a seguir temos as medidas da altura de 20 hastes. Faça o Boxplot correspondente.



Dados da usinagem			
903,88	1036,92	1098,04	1011,26
1020,70	915,38	1014,53	1097,79
934,52	1214,08	993,45	1120,19
860,41	1039,19	950,38	941,83
936,78	1086,98	1144,94	1066,12

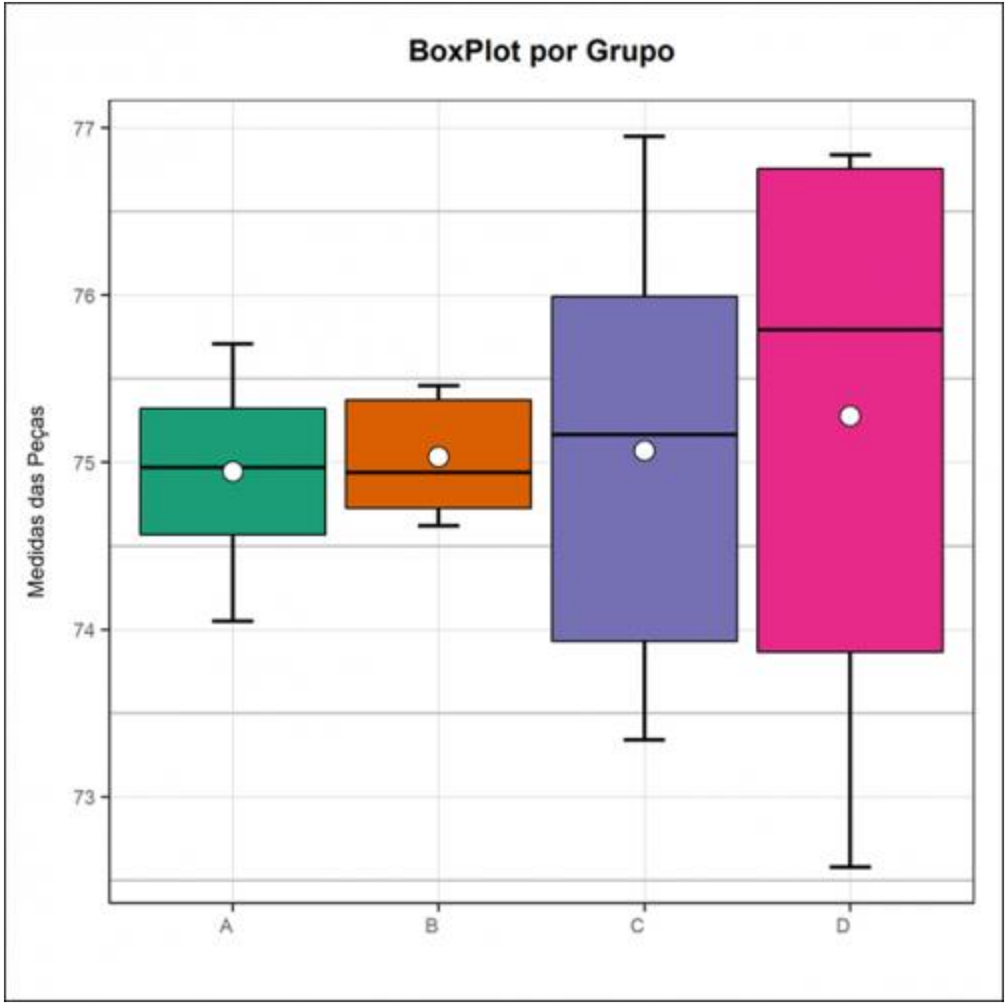
Box-Plot Exemplo 2

- Uma indústria produz uma peça automotiva cujo valor de referência é 75cm. Após verificar lotes com peças fora de especificação, enviaram duas equipes de trabalhadores (A e B) para um treinamento.
- Para verificar a eficiência do treinamento, foram selecionadas 10 peças produzidas pelas equipes A e B e 10 peças produzidas pelas equipes C e D que não participaram do treinamento

A		B		C		D	
75,27	74,93	74,94	74,75	75,93	73,34	75,98	76,75
75,33	74,72	75,25	74,65	76,95	74,04	75,61	76,78
74,58	74,53	75,44	74,94	75,47	75	74,2	74,74
75,01	75,32	74,62	74,92	73,6	76,18	76,44	72,58
75,71	74,05	75,35	75,46	74,85	75,33	76,84	72,86

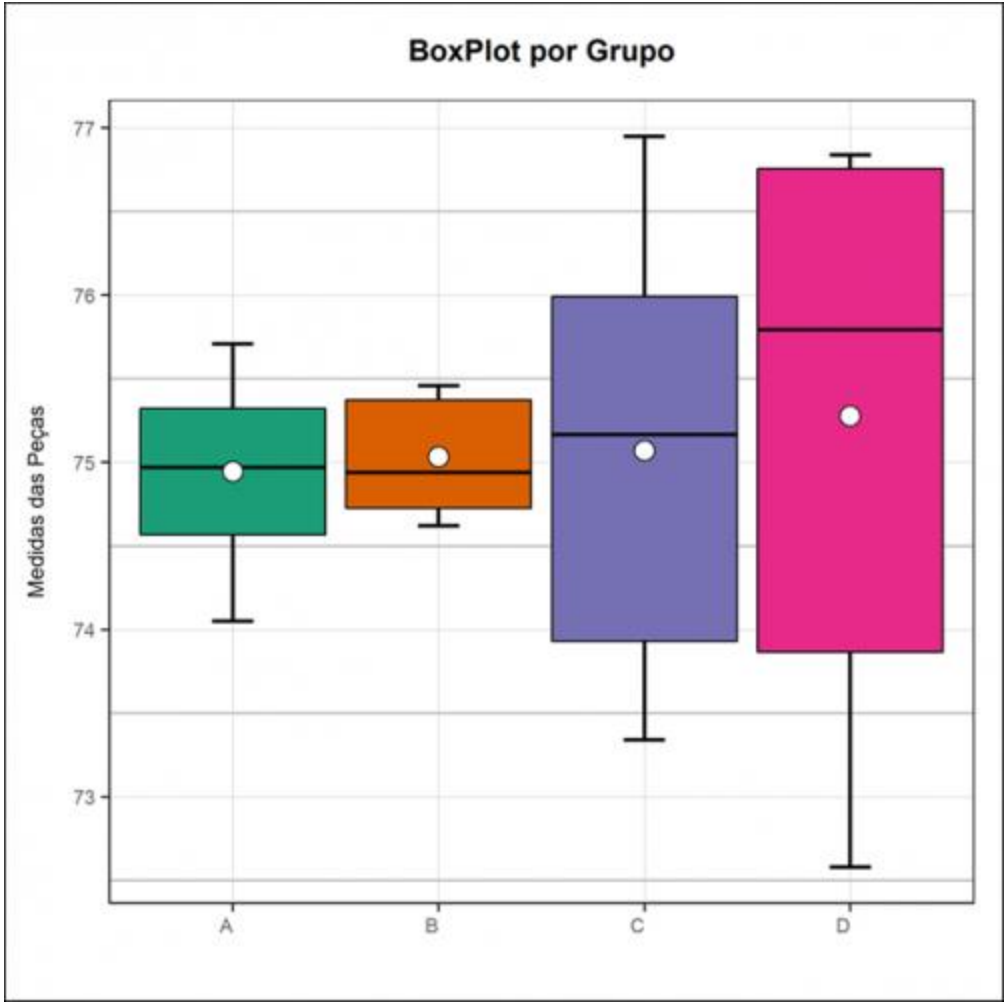
Box-Plot Exemplo 2

- Analisando o gráfico podemos observar que:
  - As equipes A e B produzem peças com menor variabilidade, indicando que o treinamento teve o efeito desejado;
  - A equipe D é a que produz peças com maior variabilidade;
  - A equipe B é a que produz peças com menor variabilidade.



Box-Plot Exemplo 2

- **Considerações:** Como as peças das equipes A e B tem menor variabilidade e com valor médio próximo do valor de referência, vale a pena enviar as demais equipes para o treinamento.



Histograma

- Organizar os dados coletados em ordem crescente;
- Determinar a amplitude total;
- Dividir a amplitude total em um nº adequado de intervalos de preferência com a mesma amplitude;
- Nº mínimo de intervalos 5, número máximo 20;

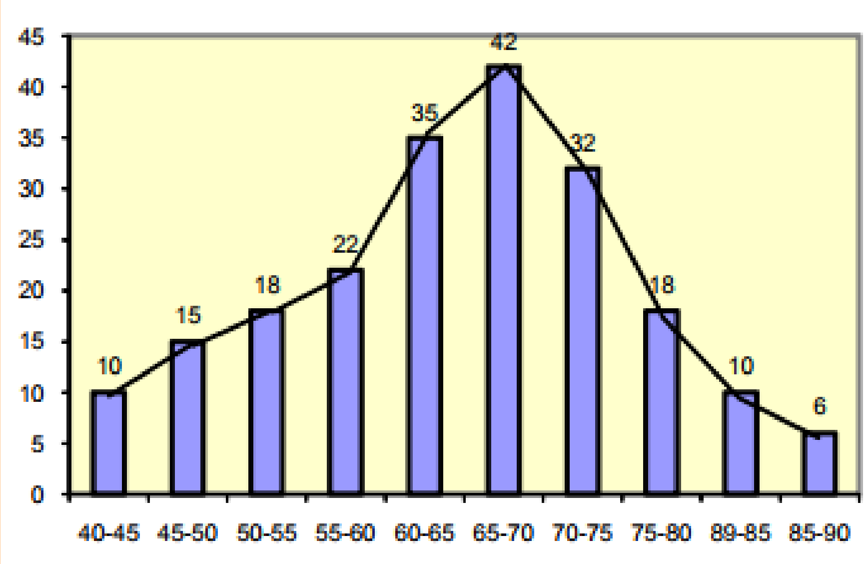
Histograma

- Quando possível os pontos médios dos intervalos devem coincidir com os valores realmente observados
- **Distribuições Simétricas e Assimétricas** - Os histogramas podem apresentar distribuição simétricas ou assimétricas
- **Polígono de Frequências** – Unindo os valores médios dos intervalos de classe, transforma-se o histograma num polígono de frequências. Pode então compará-la com uma curva teórica (Normal).

Histograma Simétrico

HISTOGRAMA E POLÍGONO DE FREQUÊNCIA

Pesos (x_1)	Nº alunos (f_1)
40-45	10
45-50	15
50-55	18
55-60	22
60-65	35
65-70	42
70-75	32
75-80	18
89-85	10
85-90	6
Total	208

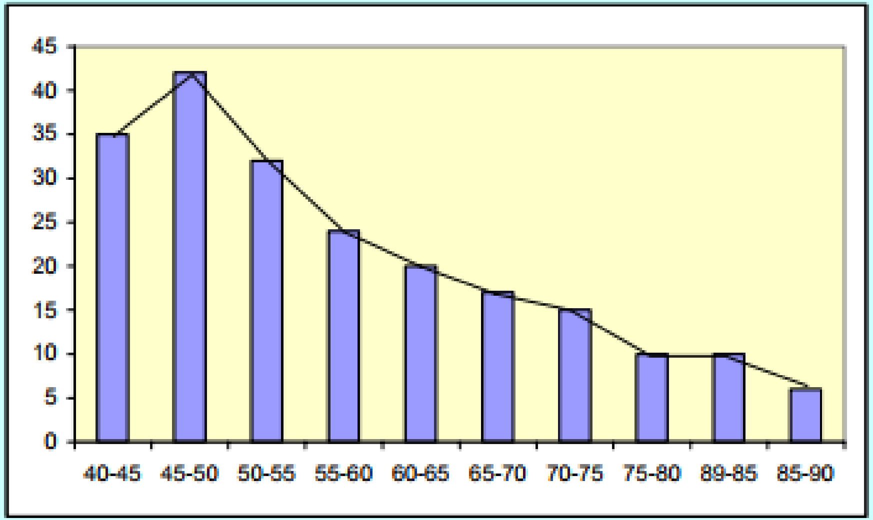


Histograma Assimétrico à Esquerda

HISTOGRAMA E POLÍGONO DE FREQUÊNCIA

Assimétrico à esquerda

Pesos (x_1)	Nº alunos (f_1)
40-45	35
45-50	42
50-55	32
55-60	24
60-65	20
65-70	17
70-75	15
75-80	10
89-85	10
85-90	6
Total	208

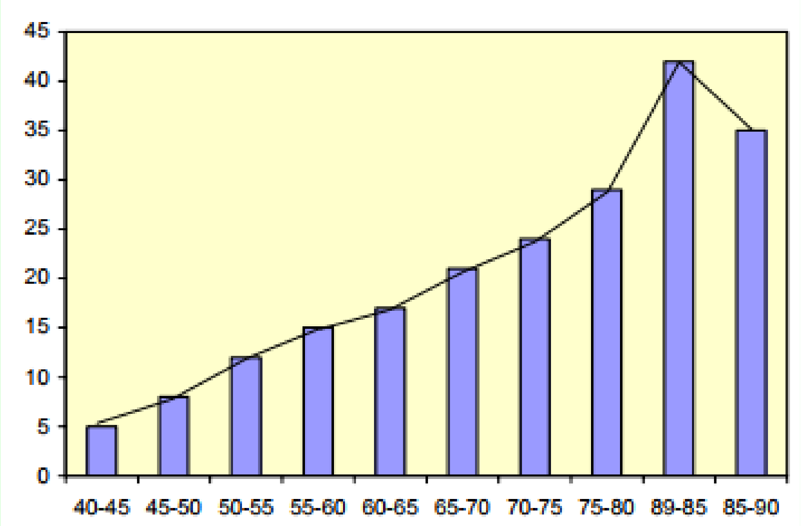


Histograma Assimétrico à Direita

HISTOGRAMA E POLÍGONO DE FREQUÊNCIA

Assimétrico à direita

Pesos (x_i)	Nº alunos (f_i)
40-45	5
45-50	8
50-55	12
55-60	15
60-65	17
65-70	21
70-75	24
75-80	29
89-85	42
85-90	35
Total	173



Medidas de dispersão

- Medidas de dispersão descrevem a variabilidade dos dados em torno de uma tendência central.
- Podem ser interpretadas como uma medida de precisão associada as medidas de posição.
- **Principais medidas: variância, desvio padrão, amplitude, intervalo interquartil.**

Amplitude Total (A)

- É a diferença entre o maior e o menor dos valores da série.
- A utilização da amplitude total como medida de dispersão é muito limitada, pois sendo uma medida que depende apenas dos valores externos, é instavel, não sendo afetada pela dispersão dos valores internos.

$$(Amplitude\ Total)\ de\ dados\ \rightarrow\ AT\ =\ x_{max} - x_{min}$$

Desvio Médio (DM)

- O conceito estatístico de desvio corresponde ao conceito matemático de distância. **A dispersão dos dados em relação à média de uma sequência pode ser avaliada através dos desvios de cada elemento da sequência em relação à média da sequência**. O desvio médio é definido como sendo uma média aritmética dos desvios de cada elemento da série para a média da série, ou seja,

$$\boxed{DM = \frac{\sum f_i . \left| x_i - \bar{x} \right|}{n}}$$

Exemplo Desvio Médio

- Considere as notas 2, 8, 5, 6 obtidas por 4 alunos, numa avaliação de Estatística. Determine o desvio médio.
  - Inicialmente, calcularemos a média:

$$\bar{x} = \frac{2+8+5+6}{4} = 5{,}25$$

  - Agora, calculamos o desvio médio, lembrando que $f_i = 1$, visto que cada um dos quatro valores apareceu uma única vez.

Exemplo Desvio Médio

$$DM = \frac{\sum f_i . |x_i - \bar{x}|}{n} =$$

$$= \frac{|2-5{,}25| + |8-5{,}25| + |5-5{,}25| + |6-5{,}25|}{4} = \frac{|-3{,}25| + |2{,}75| + |-{,}025| + |0{,}75|}{4} =$$

$$= \frac{3{,}25 + 2{,}75 + 0{,}25 + 0{,}75}{4} = \frac{7}{4} = 1{,}75$$

- **Interpretação:** Em média, cada elemento da seqüência está afastado do valor 5,25 por 1,75 unidades.

Variância (S^2 ou σ^2) e Desvio padrão (s ou σ)

- Uma dificuldade em se operar o DM se deve à presença do módulo, para que as diferenças $x_i - \bar{x}$ possam se interpretadas como distâncias.
- Outra forma de se conseguir que as diferenças $x_i - \bar{x}$ se tornem sempre positivas ou nulas é considerar o quadrado destas diferenças, isto é, $(x_i - \bar{x})^2$.
- Se substituirmos, na fórmula do DM a expressão $|x_i - \bar{x}|$ por $(x_i - \bar{x})^2$, obteremos nova medida de dispersão chamada **variância.**

Variância (S^2 ou σ^2) e Desvio padrão (s ou σ)

- A **variância populacional** é representada por $\boldsymbol{\sigma^2}$ (sigma ao quadrado), enquanto que a **variância amostral** é representada por $\boldsymbol{S^2}$.
- A fórmula geral da **variância populacional** e da **variância amostral** são, respectivamente

$$\boxed{\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{n}} \quad \text{e} \quad \boxed{s^2 = \frac{\sum f_i . (x_i - \bar{x})^2}{n - 1}}$$

n-1 para amostras <30

Variância (S^2 ou σ^2) e Desvio padrão (s ou σ)

- O **desvio padrão** é a raiz quadrada da variância, ou seja:

$$\boxed{\sigma = \sqrt{\sigma^2}} \text{ ou } \boxed{s = \sqrt{s^2}\ .}$$

- De modo mais simples, podemos generalizar:

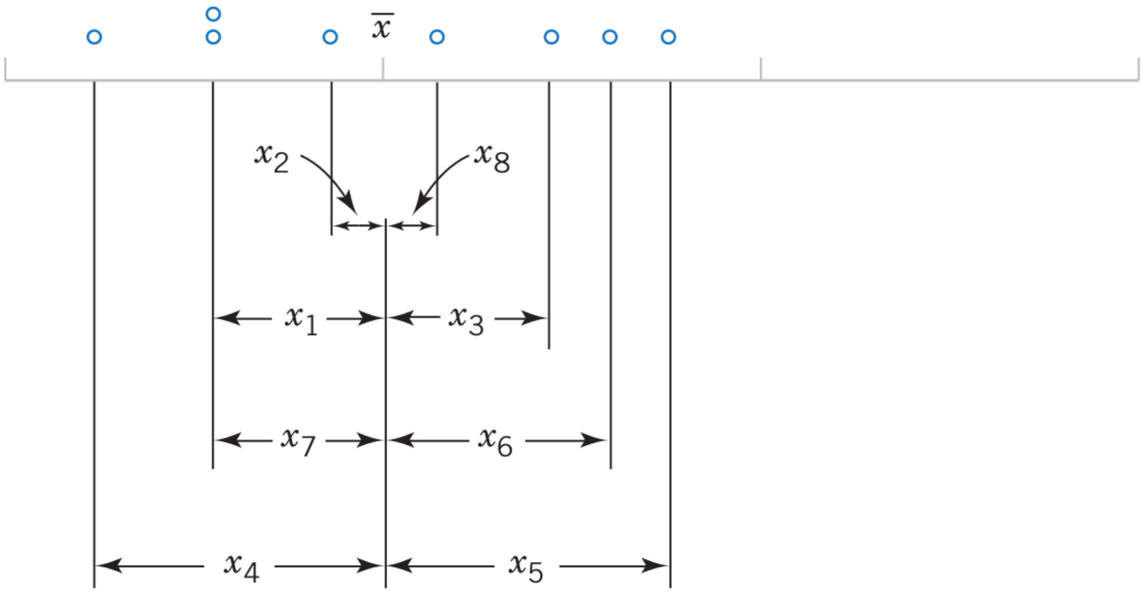
$$DP = \sqrt{Var}\ .$$

Variância (S^2 ou σ^2) e Desvio padrão (s ou σ)

- Quando estamos trabalhando com uma amostra, sem conhecermos o verdadeiro **valor da média ou do desvio padrão**, admitimos que a média da amostra ($\bar{x}$) esteja próxima do valor da média populacional, e que a variância da amostra **(variância amostral)** esteja próxima da variância populacional. A raiz quadrada da variância amostral é chamada **desvio padrão amostral.**

Variância Amostral (S^2)

- A variância amostral, usualmente denotada por S^2, **fornece uma maneira de medir o desvio médio das observações com respeito ao valor central x.**



Variância Amostral (S^2) - Exemplo

- Calcular a variância amostral da sequencia: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1:

i	x_i	$d_i^2 = (x_i - \overline{x})^2$
1	12,6	0,16
2	12,9	0,01
3	13,4	0,16
4	12,3	0,49
5	13,6	0,36
6	13,5	0,25
7	12,6	0,16
8	13,1	0,01
$\sum$	104,0	1,6
$\frac{1}{n}\sum$	$\overline{x} = 13,0$	$S^2 = 0,2$

Variância (S^2) – Cálculo Alternativo

- Em muitos casos pode ser útil trabalhar com a variância em função da média dos quadrados das observações:

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

- No caso a Fórmula alternativa:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right).$$

Variância (S^2) – Cálculo Alternativo

- Calcular a variância amostral da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.

i	x_i	$d_i^2 = (x_i - \overline{x})^2$	x_i^2
1	12,6	0,16	158,76
2	12,9	0,01	166,41
3	13,4	0,16	179,56
4	12,3	0,49	151,29
5	13,6	0,36	184,96
6	13,5	0,25	182,25
7	12,6	0,16	158,76
8	13,1	0,01	171,61
$\sum$	104,0	1,6	1353,6
$\frac{1}{n}\sum$	$\overline{x} = 13,0$	$S^2 = 0,2$	169,2

- Pelo cálculo alternativo:

$$S^2 = 169,2 - 13,0^2 = 0,2.$$