Fatec São José dos Campos Prof. Jessen Vidal

Centro Paula Souza

GOVERNO DO ESTADO SÃO PAULO

II- MEDIDAS DE DISTRIBUIÇÃO DE FREQUÊNCIAS

nanci.oliveira@fatec.sp.gov.br

MEDIDAS DE DISTRIBUIÇÃO DE FREQUÊNCIAS

PARTE 1 - Medidas de Tendência Central

- Média Aritmética
- Mediana
- Moda

PARTE 2 - Medidas de Dispersão ou Variabilidade

- Variância
- Desvio Padrão
- Coeficiente de variação

PARTE 3 - Medidas de Posição

- Quartil, Decil e Percentil
- Representação gráfica dos quartis: Box Plot

PARTE 1 - MEDIDAS DE TENDÊNCIA CENTRAL

- Média Aritmética ($\bar{X}$)
- Mediana (Md)
- Moda (Mo)

MÉDIA ARITMÉTICA

- É a medida estatística mais utilizada.
- **É a soma dos valores de todos os dados dividido pelo número de dados.**

MEDIANA

É o dado que fica na posição central de um conjunto de dados que estão em ordem crescente ou decrescente.

Como encontrar a mediana?

Depois de ordenados os valores por ordem crescente ou decrescente, a **mediana** é:

- o valor que ocupa a posição central, se a quantidade desses valores for **ímpar**;
- a média dos dois valores centrais, se a quantidade desses valores for **par**.

MODA

É valor que mais se repete num conjunto de dados, ou seja, é o dado que ocorre com maior frequência.

No caso de dois dados apresentarem a mesma frequência elevada, os dados são **bimodais**.

Caso não haja dados repetidos, os dados são **amodais**.

EXEMPLO 1:
Média, Moda e Mediana

Gastos em eletricidade:

Esta tabela não é de distribuição de frequências.

Meses	JAN.	FEV.	MAR.	ABR.	MAI.
Gasto (em €)	25€	22€	35€	28€	35€

$$\overline{X} = \frac{25+22+35+28+35}{5} = \frac{145}{5} = 29 \Rightarrow \overline{X} = 29$$

Moda: Mo = 35 €

Média: $\overline{X}$ = 29 €

Rol: 22 25 **28** 35 35

Mediana: Md = 28 €

Número impar de dados

EXEMPLO 2:
Média, Moda e Mediana

Gastos em eletricidade:

Meses	JAN.	FEV.	MAR.	ABR.	MAI.	JUN.
Gastos (em €)	25€	22€	35€	28€	35€	33€

$$\overline{X} = \frac{25+22+35+28+35+33}{6} = \frac{178}{6} = 29{,}67 \Rightarrow \overline{X} = 29{,}67$$

Moda: 35 €

Média: 29,67 €

Rol: 22 25 **28** **33** 35 35

$$Md = \frac{28+33}{2} = \frac{61}{2} = 30{,}5$$

Número par de dados

Mediana: Md = 30,5 €

MÉDIA DE DADOS AGRUPADOS

A **média de uma distribuição de frequências,** para uma amostra é dada por:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} f_i}$$

Onde:

$X_i = Pm_i = variável\ ou\ ponto\ médio\ da\ classe\ i$

$f_i = frequência\ da\ classe\ i$

EXEMPLO 1
MÉDIA PARA DADOS AGRUPADOS

Use a distribuição de frequências abaixo para calcular a média do número de minutos que uma amostra de internautas gastou durante sua navegação mais recente na rede.

X_i	f_i
12,5	6
24,5	10
36,5	13
48,5	8
60,5	5
72,5	6
84,5	2
	$\Sigma f_i = 50$

Como os dados estão agrupados em uma distribuição de frequências, temos que usar a fórmula:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} f_i}$$

Solução

X_i	f_i	$X_i \cdot f_i$
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
	$\Sigma f_i = 50$	$\Sigma X_i \cdot f_i = 2089,0$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{2089}{50} = 41,8$$

Logo, $\bar{x} = 41,8$ *minutos.*

EXEMPLO 2
MÉDIA PARA DADOS AGRUPADOS

Use a distribuição de frequências abaixo para calcular a média dos estaturas, em cm, de 50 atletas.

Estaturas	***f_i***
150 ├─ 155	2
155 ├─ 160	10
160 ├─ 165	12
165 ├─ 170	15
170 ├─ 175	5
175 ├─ 180	6
	$\Sigma f_i = 50$

Como os dados estão agrupados em uma distribuição de frequências, temos que usar a fórmula da média aritmética, porém, inicialmente, temos que encontrar os ***pontos médios de cada classe X_i*** *:*

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} f_i}$$

Solução

Estaturas	X_i	f_i	$X_i \cdot f_i$
150 |— 155	152,5	2	305
155 |— 160	157,5	10	1575
160 |— 165	162,5	12	1950
165 |— 170	167,5	15	2512,5
170 |— 175	172,5	5	862,5
175 |— 180	177,5	6	1065
		$\Sigma f_i = 50$	$\Sigma X_i \cdot f_i = 8270$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{8270}{50} = 165,4$$

Logo, $\bar{x} = 165,4 \; cm$.

PARTE 2 - MEDIDAS DE DISPERSÃO OU VARIABILIDADE

- As medidas de dispersão visam atribuir um valor numérico que expresse a homogeneidade ou não entre os diversos valores obtidos numa pesquisa.
- As medidas de dispersão:
  – Podem indicar se os dados estão próximos ou não de uma medida de tendência central (como a média aritmética);
  – Podem indicar se há uma grande variação ou não entre os dados, ou seja, se os dados são mais homogêneos ou heterogêneos;
  – Permitem a comparação entre dois ou mais grupos de dados com o intuito de verificar qual teve maior ou menor variação.

MEDIDAS DE DISPERSÃO OU VARIABILIDADE

- Amplitude Total (AT)
- Desvio Médio (DM)
- Variância (s^2)
- Desvio Padrão (s)
- Coeficiente de Variação (C.V.)

VARIÂNCIA PARA DADOS AGRUPADOS

Variância de uma população é a média dos quadrados dos desvios.

Os desvios são dados por $\left(X_i - \overline{X}\right)$.

VARIÂNCIA AMOSTRAL:

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \cdot f_i}{n-1}$$

onde $\boldsymbol{n = \Sigma f_i}$ é o número total de dados.

DESVIO PADRÃO PARA DADOS AGRUPADOS

O desvio padrão da amostra de uma distribuição de frequências é dado pela raiz quadrada da variância.

DESVIO PADRÃO AMOSTRAL:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot f_i}{n - 1}}$$

onde ***n = Σfi*** é o número total de dados.

EXEMPLO 1- DESVIO PADRÃO PARA DADOS AGRUPADOS EM DISTRIBUIÇÃO DE FREQUÊNCIAS

Os resultados de uma amostra do número de crianças por família em uma região estão dispostos na tabela abaixo. Determine a média e o desvio padrão.

número de crianças por família

x_i	f_i
0	10
1	19
2	7
3	7
4	2
5	1
6	4
	$\Sigma f_i = 50$

número de famílias

X_i	f_i	$X_i \cdot f_i$	$(X - \bar{X}_i)^2 f_i$
0	10	0	$(0 - 1{,}8)^2 \cdot 10 = (3{,}24) \cdot 10 = 32{,}4$
1	19	19	$(1 - 1{,}8)^2 \cdot 19 = (0{,}64) \cdot 19 = 12{,}16$
2	7	14	$(2 - 1{,}8)^2 \cdot 7 = (0{,}04) \cdot 7 = 0{,}28$
3	7	21	$(3 - 1{,}8)^2 \cdot 7 = (1{,}44) \cdot 7 = 10{,}08$
4	2	8	$(4 - 1{,}8)^2 \cdot 2 = (4{,}84) \cdot 2 = 9{,}68$
5	1	5	$(5 - 1{,}8)^2 \cdot 1 = (10{,}24) \cdot 1 = 10{,}24$
6	4	24	$(6 - 1{,}8)^2 \cdot 4 = (17{,}64) \cdot 4 = 70{,}56$
	$\Sigma f_i = 50$	$\Sigma X_i \cdot f_i = 91$	$\Sigma(X_i - \bar{X})^2 f_i = 145{,}40$

Média: $\bar{X} = \frac{\sum X_i \cdot f_i}{\sum f_i} = \frac{91}{50} = \mathbf{1{,}8}\ \boldsymbol{crianças}$

Desvio padrão: $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot f_i}{n - 1}} = \sqrt{\frac{145{,}4}{49}} = \sqrt{2{,}967} = \mathbf{1{,}7\ crianças}$

EXEMPLO 2- DESVIO PADRÃO PARA DADOS AGRUPADOS EM DISTRIBUIÇÃO DE FREQUÊNCIAS

O resultado de uma sondagem na qual mil adultos foram indagados sobre quanto gastavam anualmente na preparação de uma viagem de férias resultou na distribuição de frequência abaixo. Estime a média e o desvio padrão amostrais do conjunto de dados.

Gastos (US$)	f_i
0 ├─ 100	380
100 ├─ 200	230
200 ├─ 300	210
300 ├─ 400	50
400 ├─ 500	60
500 ├─ 600	70
	$\Sigma f_i = 1000$

Gastos (US$)	f_i	X_i	$X_i \cdot f_i$	$(x_i - \bar{x})^2 f_i$
0 |— 100	380	**50**	19.000	$(50 - 189)^2 \cdot 380 = 7.341.980$
100 |— 200	230	**150**	34.500	$(150 - 189)^2 \cdot 230 = 349.830$
200 |— 300	210	**250**	52.500	$(250 - 189)^2 \cdot 210 = 781.410$
300 |— 400	50	**350**	17.500	$(350 - 189)^2 \cdot 50 = 1.296.050$
400 |— 500	60	**450**	27.000	$(450 - 189)^2 \cdot 60 = 4.087.260$
500 |— 600	70	**550**	38.500	$(550 - 189)^2 \cdot 70 = 9.122.470$
	$\Sigma f_i = 1000$		$\Sigma X_i \cdot f_i = 189.000$	$\Sigma (x_i - \bar{x})^2 f_i = 22.979.000$

Média: $\bar{X} = \frac{\sum X_i \cdot f_i}{\sum f_i} = \frac{189000}{1000} = \mathbf{189}\ dólares$

Desvio padrão:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot f_i}{n - 1}} = \sqrt{\frac{22979000}{999}} = \sqrt{23002{,}002} = \mathbf{151{,}66\ dólares}$$

COEFICIENTE DE VARIAÇÃO: MEDIDA DE DISPERSÃO RELATIVA

- A variância e o desvio padrão somente são comparáveis quando se referem a mesma escala de medida e quando os grupos têm média não muito diferentes.
- **A avaliação da variação de dados de uma pesquisa e a comparação entre grupos de dados é feita através de um índice percentual, denominado coeficiente de variação.**

COEFICIENTE DE VARIAÇÃO

Coeficiente de variação (C.V.) é a razão entre o desvio padrão e a média, multiplicada por 100. Assim, o resultado é dado **em porcentagem.**

$$\boldsymbol{C.V. = \frac{s}{\overline{\overline{X}}} \cdot 100}\ (\%)$$

EXEMPLO
COEFICIENTE DE VARIAÇÃO

Compare a variação das idades dos 2 grupos seguintes.

1º grupo - Idade, em anos, de três crianças: 1, 3, 5

2º grupo- Idade, em anos, de três adultos: 53, 55, 57

***Cálculo da média aritmética dos dois grupos*:**

1º grupo

$$\bar{X} = \frac{1+3+5}{3} = \frac{9}{3} = \mathbf{3}\ \boldsymbol{anos}$$

2º grupo

$$\bar{X} = \frac{53+55+57}{3} = \frac{165}{3} = \mathbf{55}\ \boldsymbol{anos}$$

Cálculo do desvio padrão dos dois grupos:

1º grupo

Xi	fi	$\left(X_i - \overline{X}\right)^2 \cdot fi$
1	1	$(1 - 3)^2 \cdot 1 = 4$
3	1	$(3 - 3)^2 \cdot 1 = 0$
5	1	$(5 - 3)^2 \cdot 1 = 4$
		$\sum_{1=1}^{n} \left(X_i - \overline{X}\right)^2 \cdot f_i = 8$

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \cdot fi}{n-1} = \frac{8}{2} = 4$$

$$s = \sqrt{4} = 2 \text{ anos}$$

2º grupo

Xi	fi	$\left(X_i - \overline{X}\right)^2 \cdot fi$
53	1	$(53 - 55)^2 \cdot 1 = 4$
55	1	$(55 - 55)^2 \cdot 1 = 0$
57	1	$(57 - 55)^2 \cdot 1 = 4$
		$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \cdot f_i = 8$

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \cdot fi}{n-1} = \frac{8}{2} = 4$$

$$s = \sqrt{4} = 2 \text{ anos}$$

A dispersão absoluta dos dados em torno da média é exatamente a mesma: s = 2 anos.

Cálculo do coeficiente de variação (dispersão relativa):

1o grupo:

$$C.V. = \frac{s}{\overline{X}} \cdot 100\ (\%)$$

$$\boxed{\mathbf{C.V. = \frac{2}{3} \cdot 100\% = 66{,}7\%}}$$

2o grupo:

$$C.V. = \frac{s}{\overline{X}} \cdot 100\ (\%)$$

$$\boxed{\mathbf{C.V. = \frac{2}{55} \cdot 100\ \% = 3{,}64\%}}$$

Conclusão:

O 1º grupo teve maior variação nos seus dados, pois seu coeficiente de variação é maior que do 2º grupo. Isso indica que a diferença de 2 anos (que é o desvio padrão) no 1º grupo é bastante significativa na mudança física, já no 2º grupo não é.

C- MEDIDAS DE POSIÇÃO

- Quartis (Q_1, Q_2, Q_3)
- Decis $(D_1, D_2, \ldots, D_9)$
- Percentis $(P_1, P_2, P_3, \ldots, P_{99})$

QUARTIS

- ✓ **Quartis** são números que dividem aproximadamente um conjunto ordenado de dados em **quatro partes iguais**: $\mathbf{Q_1}$, $\mathbf{Q_2}$ e $\mathbf{Q_3}$
- ✓ Cerca de ¼ dos dados é menor ou igual ao 1º quartil.
- ✓ Cerca de ½ (metade) dos dados é menor ou igual ao 2º quartil. O 2º quartil é igual à mediana do conjunto de dados.
- ✓ Cerca de ¾ dos dados é menor ou igual ao 3º quartil.

EXEMPLO - QUARTIS

A pontuação nos testes de 15 empregados envolvidos em um curso de treinamento está disposta a seguir. Obtenha os 3 quartis da pontuação dos testes.

13	9	18	15	14	21	7	10	11	20	5	18	37	16	17

Em 1º lugar, ordene o conjunto de dados e obtenha a mediana, que é igual ao 2º quartil.
Uma vez obtido o 2º quartil, pode-se dividir o conjunto de dados em 2 metades. O 1º e o 2º quartis são as medianas das metades inferior e superior do conjunto de dados.

5	7	9	10	11	13	14	15	16	17	18	18	20	21	37

Q_1 **$Q_2 = Md$** **Q_3**

Assim, cerca de ¼ dos empregados fez dez pontos ou menos, cerca da metade fez 15 pontos ou menos e cerca de ¾ conseguiu 18 pontos ou menos.

DECIS E PERCENTIS

Analogamente aos QUARTIS:

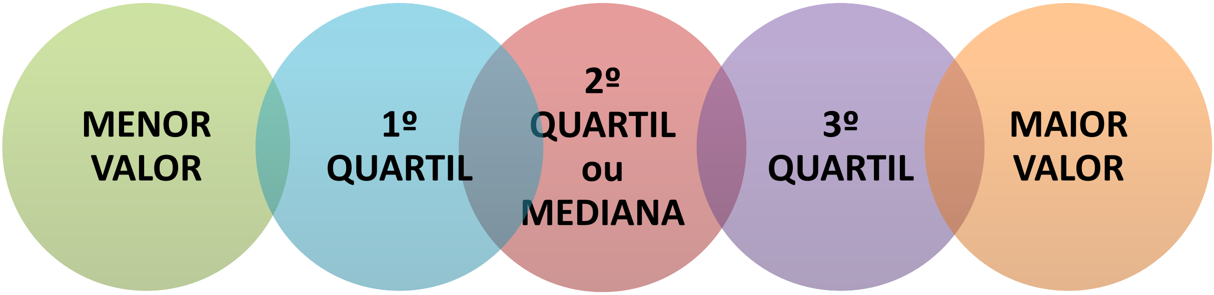
- ✓ **DECIS** dividem o conjunto de dados em 10 partes iguais: $\mathbf{D_1}$, $\mathbf{D_2}$, $\mathbf{D_3}$, ... $\mathbf{D_9}$

- ✓ **PERCENTIS** dividem o conjunto de dados em 100 partes iguais: $\mathbf{P_1}$, $\mathbf{P_2}$, $\mathbf{P_3}$, ... $\mathbf{P_{99}}$

REPRESENTAÇÃO GRÁFICA DOS QUARTIS: BOX PLOT

BOX PLOT é um **gráfico de dispersão** que relaciona os valores de uma variável com os quartis.

Para fazer o box plot é preciso conhecer os seguintes valores (**resumo cinco-números)** do conjunto de dados:



COMO FAZER UM BOX PLOT ?

1. Obtenha o resumo cinco-números do conjunto de dados

2. Construa uma escala horizontal que abranja a amplitude total dos dados.

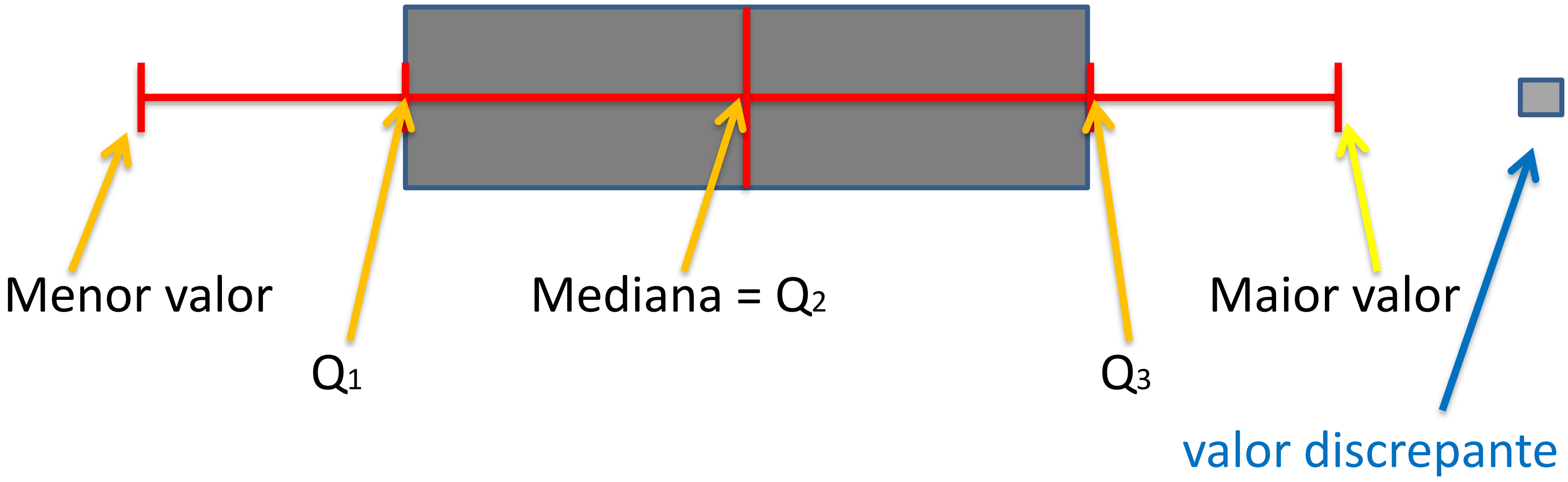
3. Plote os cinco números acima da escala horizontal.

4. Faça uma caixa acima da escala horizontal de Q1 a Q3 e trace uma reta vertical na caixa, passando por Q2.

5. Faça as “tranças” a partir da cabeça para o menor valor e o maior valor.

MODELO DE BOX PLOT

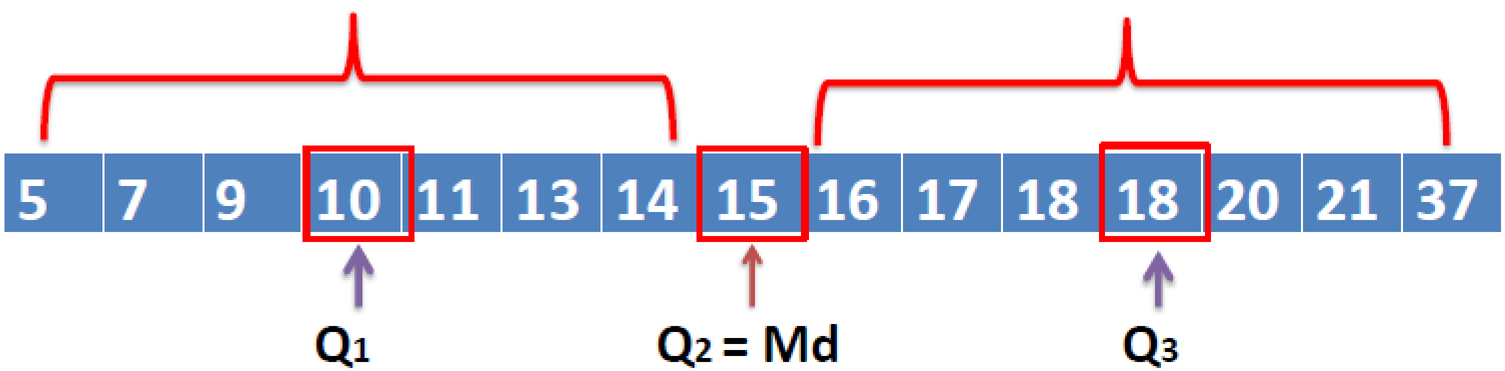
Círculos ou quadradinhos antes do menor valor ou depois do maior valor do box plot indicam valores ou dados distorcidos (discrepantes) em relação aos demais. Existe uma técnica para verificar se existe um ou mais valor discrepante, porém aqui não abordaremos esse cálculo.

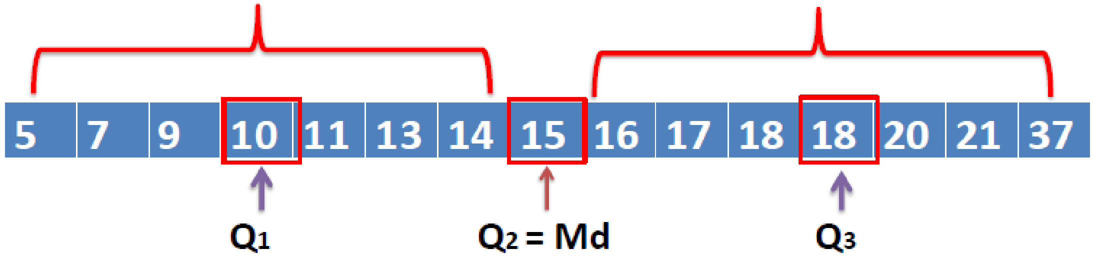


EXEMPLO 1- BOX PLOT

Faça um Box Plot que represente a pontuação dos 15 testes dados... (exemplo dos quartis).

O que você pode concluir a partir do gráfico?

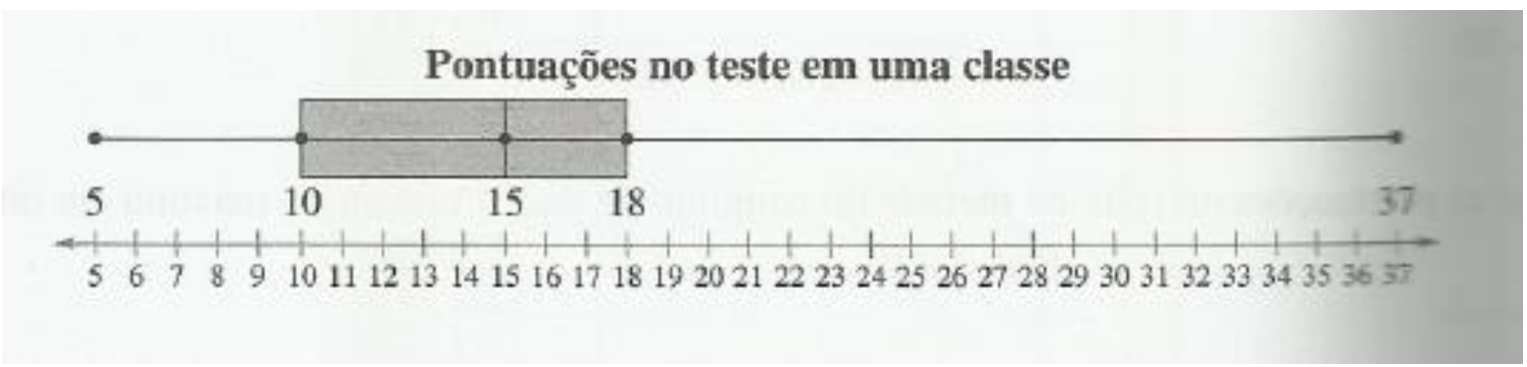




O **resumo cinco-números** das pontuações no teste são:

Menor valor = 5 Q_1 = 10 Q_2 = 15 Q_3 = 18 Maior valor = 37

BOX PLOT:



Uma das conclusões: cerca de metade das pontuações está entre 10 e 18.

EXEMPLO 2- BOX PLOT

Utilizando gráficos box plot feitos a partir dos dados das tabelas abaixo, faça uma comparação do desempenho entre os gêneros masculino e feminino nas quatro modalidades de corrida. Comente.

MULHERES

País	100m (seg)	400m (seg)	3000m (min)	Maratona (min)
Argentina	11,61	54,50	9,79	178,52
Brasil	11,31	52,80	9,77	168,75
Chile	12,00	54,90	9,37	171,38
Colômbia	11,6	53,26	9,46	165,42
Alemanha	11,01	48,16	8,75	148,53
França	11,15	51,73	8,98	155,27
Portugal	11,81	54,30	8,84	151,20
Canadá	11,00	50,06	8,81	149,50
USA	10,79	50,62	8,50	142,72
Kenya	11,73	52,70	9,20	181,05

HOMENS

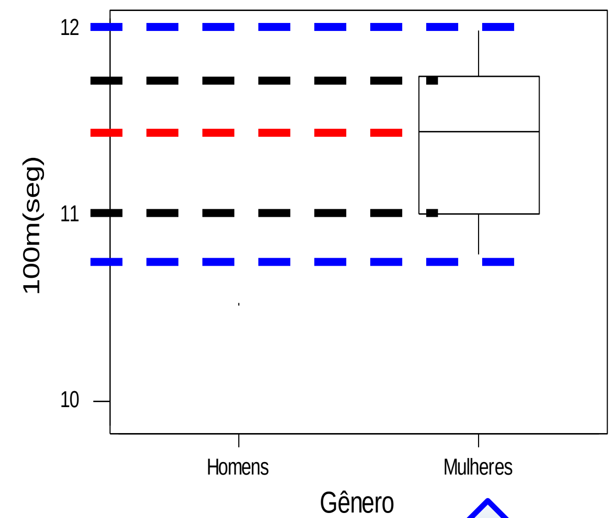
País	100m (seg)	400m (seg)	3000m (min)	Maratona (min)
Argentina	10,39	46,84	14,04	137,72
Brasil	10,22	45,21	13,62	133,13
Chile	10,34	46,20	13,61	134,03
Colômbia	10,43	46,10	13,49	131,35
Alemanha	10,16	44,50	13,21	132,23
França	10,11	45,28	13,34	132,30
Portugal	10,53	46,70	13,13	128,65
Canadá	10,17	45,68	13,55	131,15
USA	9,93	43,86	13,20	128,22
Kenya	10,46	44,92	13,10	129,75

Mulheres – Corrida 100 m

TEMPO (s)	
10,79	Menor valor
11,00	
11,01	2º Quartil = Q_1
11,15	
11,31	
Média dos valores centrais	MEDIANA = Q_2
11,60	
11,61	
11,73	3º Quartil = Q_3
11,81	
12,00	Maior valor

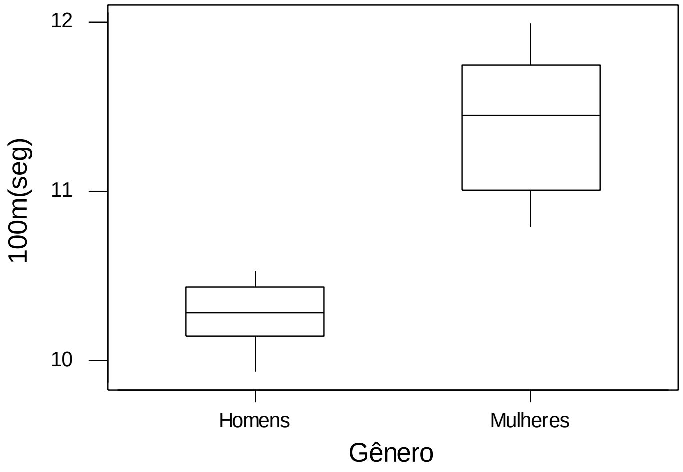
Nº par de dados (n = 10)

$$Q_2 = \frac{11{,}31+11{,}60}{2} = \frac{22{,}91}{2} = 11{,}455$$



Mulheres 100 m

Modalidade 100m (seg)

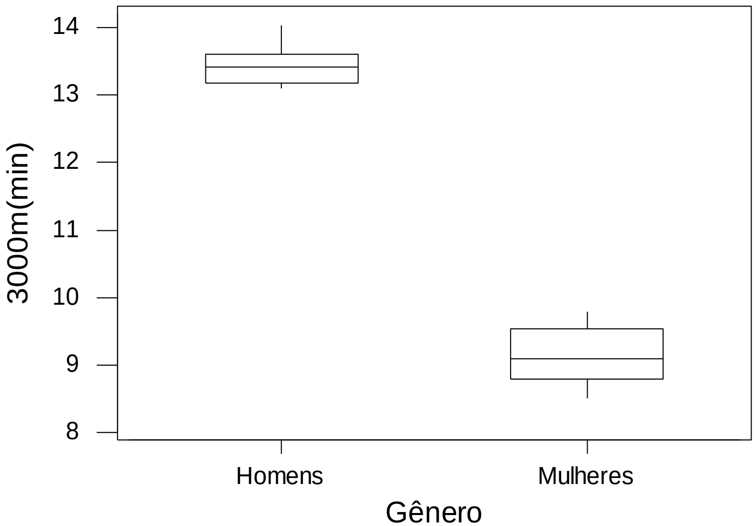


Modalidade 400m (seg)

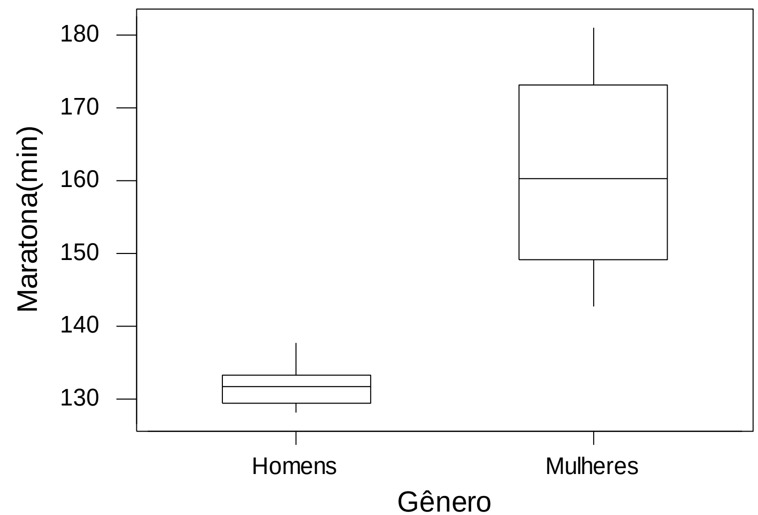
Nos gráficos box plot, pode-se observar que nas modalidades de 100m(seg) e 400m(seg):

- Os homens fazem tempos medianos menores do que das mulheres, o que é um indicativo de melhor desempenho.
- Os tempos das mulheres mostram uma variabilidade maior do que a dos homens.

Modalidade 3000m (seg)



Modalidade Maratona (min)



Nos gráficos boxplot, pode-se observar que na modalidade de 3000m (seg):

- As mulheres conseguem melhores tempos medianos que os homens, o que é um indicativo de melhor desempenho nesta modalidade.

Nos gráficos boxplot, pode-se observar que na modalidade maratona:

- Os homens fazem tempos medianos menores do que das mulheres, o que é um indicativo de melhor desempenho.
- Os tempos das mulheres mostram uma variabilidade maior do que a dos homens.

REFERÊNCIA BIBLIOGRÁFICA

LARSON, Ron; FARBER, Betsy. **Estatística e Métodos Quantitativos**. 2 ed. São Paulo: Pearson Prentice Hall, 2007.