Introdução à Probabilidade e Estatística

Apresentação

Estatística Descritiva

Departamento de Matemática
Universidade de Évora

Ano lectivo de 2016/17

Patrícia Filipe

Programa resumido

1. Estatística Descritiva (Revisões)
2. Noções Básicas de Probabilidades (Revisões)
3. Noções de Probabilidade Condicional e de Independência
4. Variáveis Aleatórias Discretas e Contínuas
5. Vectores Aleatórios Discretos
6. Distribuições Discretas e Contínuas mais Importantes
7. Introdução à Amostragem. Distribuições Amostrais
8. Estimação: Pontual e Intervalar
9. Testes de Hipóteses
10. Testes Não-Paramétricos: Ajustamento e Independência
11. Regressão Linear Simples

Horário de Atendimento

Para alguma dúvida, ou questão que queiram colocar, os alunos poderão contactar as docentes presencialmente nos seguintes horários de atendimento ou através dos respectivos e-mails para marcação de outro horário:

Patrícia Filipe (pasf@uevora.pt):

2ª e 4ªfeira, 10h-11h e 6ªfeira das 14h-16h, gab. 235

Ana Isabel Santos (aims@uevora.pt):

2ªfeira 16h-18h e 4ªfeira 14h30m-16h30m, gab. 241

Método de avaliação

Avaliação contínua 2 frequências. A nota final corresponde à média das notas das 2 frequências. O aluno é aprovado se obtiver uma nota final superior ou igual a 9.5 valores, caso contrário dispõe ainda de um exame de recurso. **A nota mínima em cada frequência é de 7 valores.**

Avaliação por exame exame de época normal e/ou de um exame de recurso. O aluno é aprovado se obtiver uma nota no exame superior ou igual a 9.5 valores.

Caso os alunos desistam da avaliação contínua na 1ª ou na 2ª frequência passam para o regime de avaliação por exame.

No regime de avaliação contínua é exigida uma assiduidade mínima de 75% de assistência às aulas.

Método de avaliação

Trabalho prático Facultativo. De grupo com máximo 4 elementos.
Utilização do *SPSS*.
Classificação de *Muito Bom* (p=0.2), *Bom* (p=0.15), *Suficiente* (p=0.1) ou *Mau* (p=0).
NE –nota média das duas frequências ou nota obtida no exame; **NF** a nota final:

- NF = NE+p(20-NE), se NE $\geqslant$ 10;
- NF = NE+pNE, se 8 $\leqslant$ NE$<$10;
- NF = NE, se NE$<$8.

Os trabalhos são sujeitos a discussão oral.

Datas de Avaliação

1ª Frequência 7 de Abril de 2017, 14h, sala a anunciar;

Entrega e discussão do trabalho 6 de Junho de 2017, 15h, sala a anunciar;

2ª Frequência/Exame de época normal 8 de Junho de 2017, 14h, sala a anunciar;

Exame de época de Recurso 22 de Junho de 2017, 14h, sala a anunciar;

Exame de época especial 4 de Julho de 2017, 14h, sala a anunciar;

Exame de época especial extraordinária 13 de Julho de 2017, 14h, sala a anunciar.

Bibliografia principal

1. Afonso, A. e Serpa, C. (2011). *Probabilidades e à Estatística. Aplicações e Soluções em SPSS*. Escolar Editora.
2. Murteira, B., Ribeiro, C. S., Silva, J. A. e C. Pimenta (2010). *Introdução à Estatística*. Escolar Editora.
3. Marôco, J. (2011). *Análise Estatística com o SPSS Statistics*. 5ª edição. Edições Report Number.
4. Pestana, M. H. e Gageiro, J. N. (2014). *Análise de dados para ciências sociais – A complementaridade do SPSS*, 6ª edição. Edições Sílabo.
5. Pestana, D. e Velosa, S. (2006). *Introdução à Probabilidade e à Estatística*. Vol. 1. Fundação Calouste Gulbenkian.
6. Ross, S.M. (2014). *Introduction to Probability Models*. 11ª edição. Academic Press.
7. Zar, J. H. (1999). *Biostatistical Analysis*. 4ª edição. Prentice

Estatística Descritiva

Classificação de Variáveis – natureza das variáveis

Variáveis Qualitativas

Identificam alguma qualidade ou característica

- **Escala nominal** — com uma sequência arbitrária (Sexo, país de origem, cor dos olhos,...)
- **Escala ordinal** — com uma ordenação natural (Estado civil, Apreciação de algo, grau de escolaridade,...)

Variáveis Quantitativas

São medidas numa escala numérica

- **Discretas** — assumem um número finito ou infinito numerável de valores (nota final, número de filhos,...)
- **Contínuas** — assumem valores numa escala contínua (Idade, altura, peso,...)

Dados em bruto, amostra ordenada e dados agrupados

Variável em estudo – População X

$\downarrow$

Amostra em bruto – $(x_1, x_2, \cdots, x_n)$

$\downarrow$

Amostra ordenada – $(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$

$\downarrow$

Dados agrupados – Tabela de Frequências

Tabela de frequências para variáveis qualitativas e quantitativas discretas

X_i'	n_i	f_i	N_i	F_i
x_1'	n_1	f_1	$N_1 (= n_1)$	$F_1 (= f_1)$
x_2'	n_2	f_2	$N_2 (n_1 + n_2)$	$F_2 (f_1 + f_2)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k'	n_k	f_k	$N_k = n$	$F_k = 1$
	n	1		

- X_i' – categorias ou valores distintos que surjem na amostra
- n – dimensão da amostra; $k =$ nº de cat. ou valores distintos
- n_i – Frequências absolutas simples
- $f_i = \frac{n_i}{n}$ – Frequências relativas simples
- $N_i = \sum_{j=1}^{i} n_j$ – Frequências absolutas acumuladas
- $F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j$ – Frequências relativas acumuladas

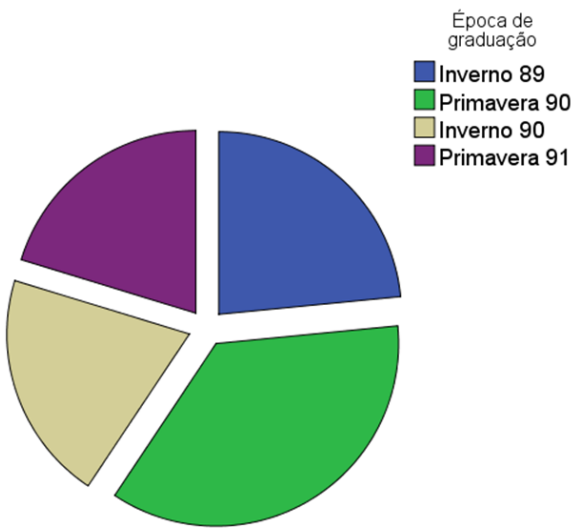
Exemplo: Número de sms enviados numa tarde de estudo

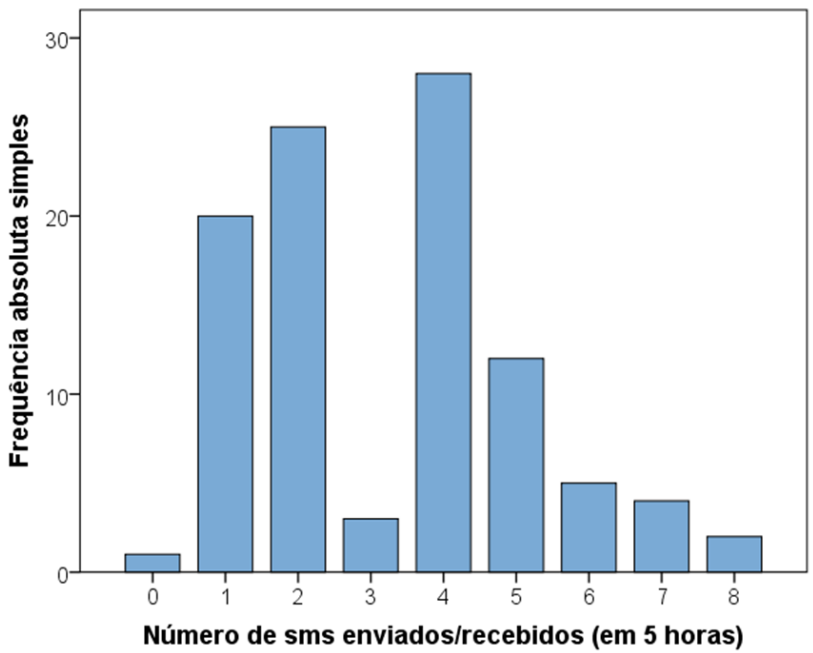
Número de sms enviados/recebidos, por 100 alunos, durante uma tarde de estudo (5 horas).

X_i'	n_i	f_i	N_i	F_i
0	1	0,01	1	0,01
1	20	0,2	21	0,21
2	25	0,25	46	0,46
3	3	0,03	49	0,49
4	28	0,28	77	0,77
5	12	0,12	89	0,89
6	5	0,05	94	0,94
7	4	0,04	98	0,98
8	2	0,02	100	1

Representações Gráficas para dados qualitativos ou quantitativos discretos

Gráfico de Barras e **Gráfico Circular**





Variáveis Quantitativas Contínuas

Construção de Classes

1. Número de classes a construir (**Regra de Sturges**):

$$k = \left[\frac{\ln n}{\ln 2}\right] + 1 \qquad (n = \text{dimensão da amostra})$$

2. Amplitude do conjunto de dados (ou Amplitude Total):
   $\Delta = X_{(n)} - X_{(1)} = \text{Máximo-Mínimo}$
3. Amplitude das classes: $A = \frac{\Delta}{k}$
4. As classes:

$C_1 = [\text{Mínimo}; \text{Mínimo} + A[$
$C_2 = [\text{Mínimo} + A; \text{Mínimo} + 2A[$
$\ldots$
$C_k = [\text{Mínimo} + (k-1)A; \text{Mínimo} + kA]$

Variáveis quantitativas contínuas – Dados Agrupados

C_i	X_i'	n_i	f_i	N_i	F_i
C_1	x_1'	n_1	f_1	N_1	F_1
C_2	x_2'	n_2	f_2	N_2	F_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
C_k	x_k'	n_k	f_k	$N_k = n$	$F_k = 1$
		n	1		

- $C_i = [l_i, L_i[$ – classes ou intervalos de classes
- $X_i' = (l_i + L_i)/2$ – Pontos médios das classes

Exemplo: Notas da 2ª frequência de IPE

Dados **Não** Agrupados – amostra em bruto

11,55	17,55	8,00	8,45	16,25	8,35	10,48	6,45	10,20	10,20
13,55	11,15	13,40	11,73	15,40	12,83	8,55	9,08	10,63	14,18
4,68	17,73	13,30	14,25	13,93	9,15	14,70	15,85	3,75	14,80
11,30	14,75	2,25	8,20	15,70	6,85	13,58	4,00	15,10	8,48
12,30	19,35	13,50	13,05	13,35	15,80	11,00	15,50	11,50	13,08
14,30	10,20	14,40	8,50	7,75	6,50	13,25	12,25	15,50	14,95
11,90	11,90	8,30	14,40	13,30	12,30	12,50	16,05		

Exemplo: Notas da 2ª frequência de IPE

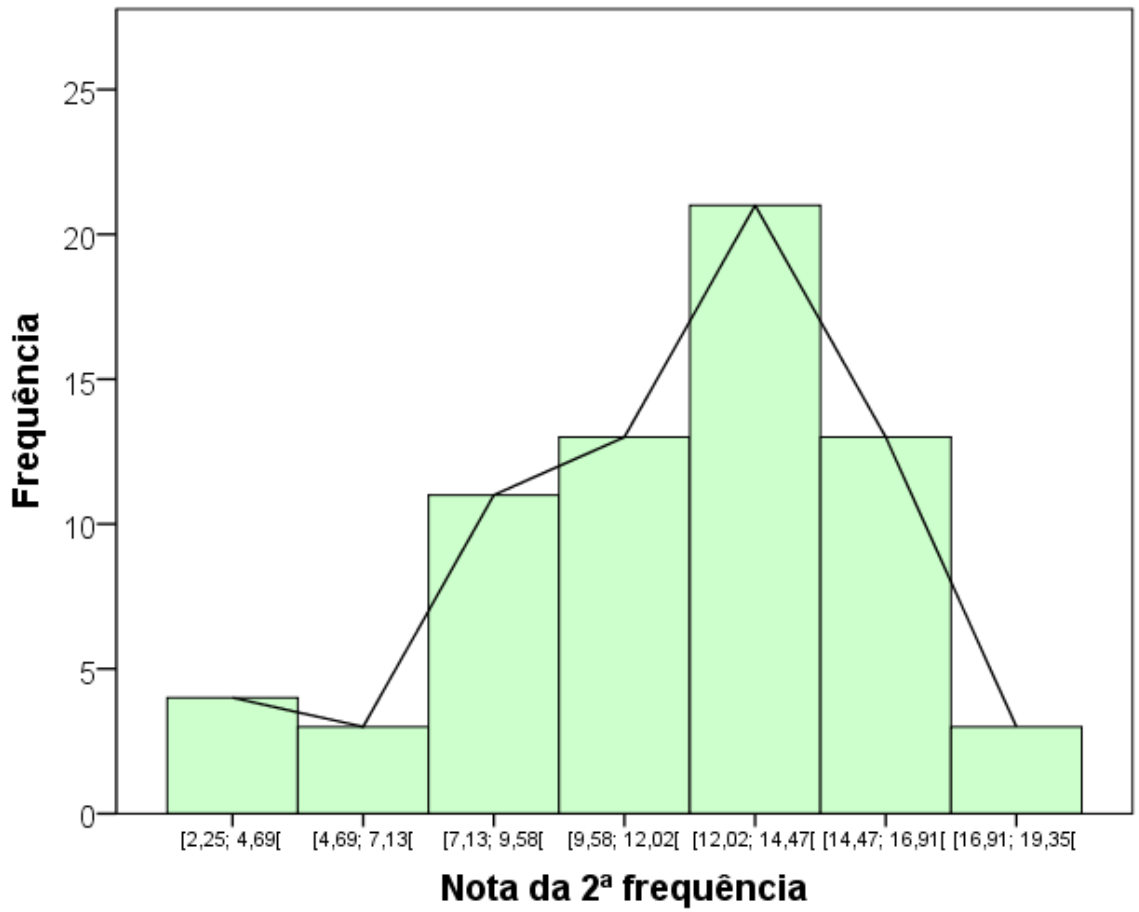
Dados Agrupados

$$n = 68; k = 7; \Delta = 17,1; A = 2,443$$

C_i	X_i'	n_i	$f_i(\%)$	N_i	$F_i(\%)$
$[2,25;\ 4,69[$	$3,47$	4	$5,9$	4	$5,9$
$[4,69;\ 7,14[$	$5,92$	3	$4,4$	7	$10,3$
$[7,14;\ 9,58[$	$8,36$	11	$16,2$	18	$26,5$
$[9,58;\ 12,02[$	$10,80$	13	$19,1$	31	$45,6$
$[12,02;\ 14,47[$	$13,25$	21	$30,9$	52	$76,5$
$[14,47;\ 16,91[$	$15,69$	13	$19,1$	65	$95,6$
$[16,91;\ 19,35]$	$18,13$	3	$4,4$	68	100
		68	100		

Representações Gráficas para dados contínuos

Histograma e **Polígono de frequências**



Medidas de Localização (Tendência Central) – Moda (M_o)

Dados não agrupados:

$M_o = \hat{X} =$ Valor mais frequente

Dados agrupados qualitativos ou discretos:

$M_o = \hat{X} =$ Valor ou categoria com maior frequência simples

Dados agrupados contínuos:

$$M_o = \hat{X} = l_i + A_i \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right)$$

l_i – Limite inferior da classe modal
A_i – Amplitude da classe modal
$\Delta_1 = n_i - n_{i-1}$; $\Delta_2 = n_i - n_{i+1}$
n_i – Frequência absoluta simples da classe modal
n_{i-1} – Frequência absoluta simples da classe anterior à classe modal
n_{i+1} – Frequência absoluta simples da classe posterior à classe modal

Medidas de Localização (Tendência Central) – Mediana (M_e)

Dados não agrupados ou discretos:

$$M_e = \tilde{X} = \begin{cases} \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & n \quad \text{par} \\ X_{(\frac{n+1}{2})}, & n \quad \text{ímpar} \end{cases}$$

Dados agrupados contínuos:

$$M_e = \tilde{X} = l_i + A_i \left(\frac{\frac{n}{2} - N_{i-1}}{n_i} \right)$$

l_i – Limite inferior da classe mediana
A_i – Amplitude da classe mediana
N_{i-1} – Frequência absoluta acumulada da classe anterior à classe mediana
n_i – Frequência absoluta simples da classe mediana

Medidas de Localização (Tendência Central) – Média Amostral

Dados não agrupados:

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

Dados agrupados:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} n_i X'_i = \sum_{i=1}^{k} f_i X'_i$$

Medidas de Localização (Tendência Não Central) – Quantis (Q_p)

Dados não agrupados ou discretos:

$$Q_p = \begin{cases} \frac{X_{(np)} + X_{(np+1)}}{2}, & np \quad \text{inteiro} \\ X_{([np]+1)}, & np \quad \text{não-inteiro} \end{cases}, \; p \in]0,1[$$

Dados agrupados contínuos:

$$Q_p = l_i + A_i \left(\frac{np - N_{i-1}}{n_i} \right), \; p \in]0,1[$$

l_i – Limite inferior da classe do quantil
A_i – Amplitude da classe do quantil
N_{i-1} – Frequência absoluta acumulada da classe anterior à classe do quantil
n_i – Frequência absoluta simples da classe do quantil

Quartis: $p = 1/4 \rightarrow Q_1 \equiv Q_{0.25}$; $p = 2/4 \rightarrow Q_2 \equiv Q_{0.5} \equiv M_e$; $p = 2/4 \rightarrow Q_3 \equiv Q_{0.75}$
Decis: $p = 1/10 \rightarrow D_1 \equiv Q_{0.1}$; $p = 2/10 \rightarrow D_2 \equiv Q_{0.2}$; ... ; $p = 9/10 \rightarrow D_9 \equiv Q_{0.9}$
Percentis: $p = 1/100 \rightarrow P_1 \equiv Q_{0.01}$; $p = 2/100 \rightarrow P_2 \equiv Q_{0.02}$; ... ; $p = 99/100 \rightarrow P_{99} \equiv Q_{0.99}$

Medidas de Dispersão

Amplitude Total: $\Delta = X_{(n)} - X_{(1)} =$ Máximo-Mínimo

Amplitude Inter-Quartílica ou Dispersão Quartal:
$Q = Q_3 - Q_1$

Intervalo de Variação: $Q' = Q_{0.9} - Q_{0.1} \equiv P_{90} - P_{10}$

Variância:

Dados não agrupados

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{n}{n-1}\overline{X}^2$$

Dados agrupados

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}n_i(X_i' - \overline{X})^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{k}n_iX_i'^2 - \frac{n}{n-1}\overline{X}^2$$

Desvio-Padrão: $S = \sqrt{\text{Variância}}$

Medidas de Dispersão

Coeficiente de Dispersão: $CD = \frac{S}{\overline{X}}$

Coeficiente de Variação: $CV = CD \times 100\%$

Para valores do CV inferiores a 50% a média será tanto mais representativa quanto menor o valor deste coeficiente. Valores superiores a 50% indicam uma pequena representatividade da média.

NOTA: Estes coeficientes só se utilizam quando a variável toma valores de um só sinal, todos positivos ou todos negativos.

Momento empírico de ordem m – $M\prime_m$

Dados não agrupados:

$$M'_m = \frac{1}{n}\sum_{i=1}^{n} X_i^m$$

Dados agrupados:

$$M'_m = \frac{1}{n}\sum_{i=1}^{k} n_i X_i'^m$$

$M\prime_1 = \bar{x}$

Momento empírico centrado de ordem m – M_m

Dados não agrupados:

$$M_m = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^m$$

Dados agrupados:

$$M_m = \frac{1}{n}\sum_{i=1}^{k} n_i(X_i' - \overline{X})^m$$

$$M_0 = 1; M_1 = 0$$

Relações existentes entre momentos centrados e momentos:

$M_2 = M_2' - M_1'^2$
$M_3 = M_3' - 3M_1'M_2' + 2M_1'^3$
$M_4 = M_4' - 4M_1'M_3' + 6M_1'^2M_2' - 3M_1'^4$

Medidas de Assimetria (*Skewness*)

A distribuição dos dados pode classificar-se quanto à assimetria como:

Simétrica:

$$\overline{X} = M_e = M_o$$

Assimétrica positiva (ou enviesada à esquerda):

$$\overline{X} > M_e > M_o$$

Assimétrica negativa (ou enviesada à direita):

$$\overline{X} < M_e < M_o$$

Medidas de Assimetria (*Skewness*)

Coeficiente de Assimetria de Fisher:

$$\beta_1 = \frac{M_3}{S^3}$$

Grau de Assimetria de Pearson:

$$G_P = \frac{\overline{X} - M_o}{S}, \quad -3 < G_P < 3$$

Grau de Assimetria de Bowley:

$$G_B = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}, \quad -1 < G_B < 1$$

O tipo de Assimetria é determinado pelo sinal de β_1, G_P ou de G_B.

> 0 – Ass+; < 0 – Ass-; $= 0$ – Simétrica.

Medidas de Achatamento (*Kurtosis*)

As distribuições podem classificar-se quanto achatamento como: Leptocúrticas, Mesocúrticas ou Platicúrticas.

Coeficiente de Achatamento:

$$\beta_2 = \frac{M_4}{S^4} \begin{cases} < 3 & \text{Distribuição Platicúrtica} \\ = 3 & \text{Distribuição Mesocúrtica} \\ > 3 & \text{Distribuição Leptocúrtica} \end{cases}$$

Coeficiente Percentil de Achatamento:

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} \begin{cases} < 0,263 & \text{Distribuição Leptocúrtica} \\ = 0,263 & \text{Distribuição Mesocúrtica} \\ > 0,263 & \text{Distribuição Platicúrtica} \end{cases}$$

Assimetria e Achatamento no SPSS

Assimetria:

$$Sk = Skewness/stdErrorSkewness$$

- assumimos distribuição simétrica se $|Sk| \leqslant 1,96$;
- distribuição assimétrica positiva se $Sk > 1,96$;
- distribuição assimétrica negativa se $Sk < -1,96$.

Achatamento:

$$Kt = Kurtosis/stdErrorKurtosis$$

- assumimos distribuição mesocúrtica $|Kt| \leqslant 1,96$;
- distribuição leptocúrtica se $Kt > 1,96$;
- distribuição platicúrtica se $Kt < -1,96$.

Outliers

Um ***Outlier*** é um valor cuja magnitude se afasta de maneira evidente do centro da distribuição.

X_i é um outlier **Moderado** se ultrapassa uma das Barreiras moderadas:

$$X_i < Q_1 - 1.5\,Q \quad \text{ou} \quad X_i > Q_3 + 1.5\,Q$$

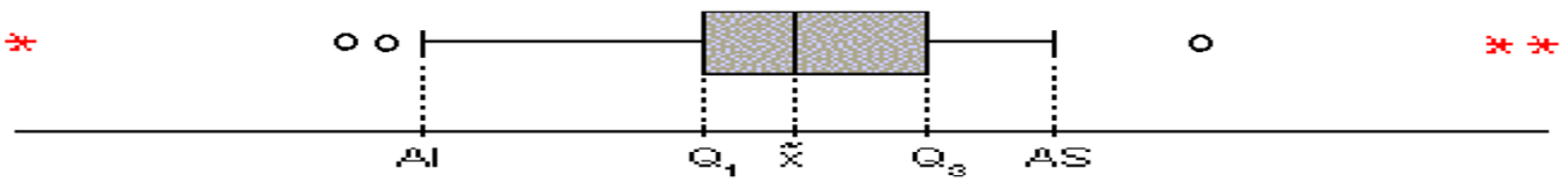
X_i é um outlier **Severo** se ultrapassa uma das Barreiras severas::

$$X_i < Q_1 - 3\,Q \quad \text{ou} \quad X_i > Q_3 + 3\,Q$$

Caixa-com-bigodes, diagrama de extremos e quartis ou *Boxplot*

Apresenta algumas das principais características descritivas de um conjunto de dados, numa imagem compacta. São representadas às escala Q_1, Q_2, Q_3, Q, o menor valor não outlier (AI), maior valor não outlier(AS) e outliers (’o’ e ’$*$’).

Fornece uma boa visualização da variabilidade dos dados e do tipo da assimetria e achatamento da distribuição.



Consultar ficheiro Boxplot.pdf

Exemplo

O tempo (em segundos) do vencedor dos 400 m masculinos, em cada Olimpíada entre 1896 e 2016 (www.olympic.org):

Ano	400m masculinos	Ano	400m masculinos
1896	54.20	1964	45.15
1900	49.40	1968	43.86
1904	49.20	1972	44.66
1908	50.00	1976	44.26
1912	48.20	1980	44.60
1920	49.60	1984	44.27
1924	47.60	1988	43.87
1928	47.80	1992	43.50
1932	46.28	1996	43.49
1936	46.66	2000	43.84
1948	46.20	2004	44.00
1952	45.09	2008	43,75
1956	46,85	2012	43,94
1960	45.07	2016	43,03

Utilizando o SPSS...

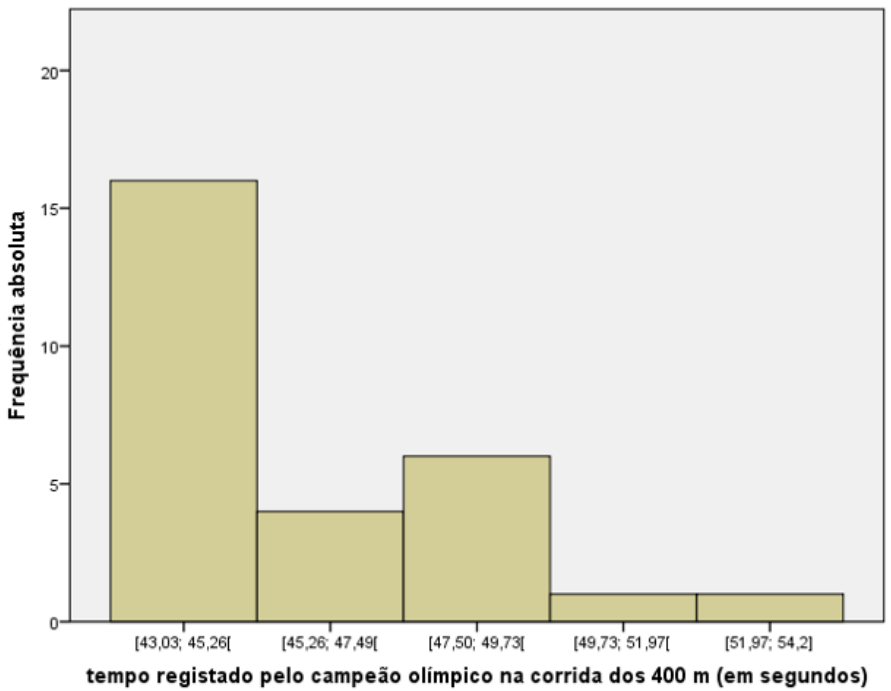
1. Determine o tempo médio (em segundos) dos tempos.
2. Qual a moda e a mediana desta amostra?
3. 25% dos atletas obtiveram um tempo inferior ou igual a x. Determine o valor de x.
4. O que pode dizer quanto à dispersão da amostra?
5. Represente graficamente os dados.
6. Como classifica a distribuição dos dados quanto à assimetria e achatamento?

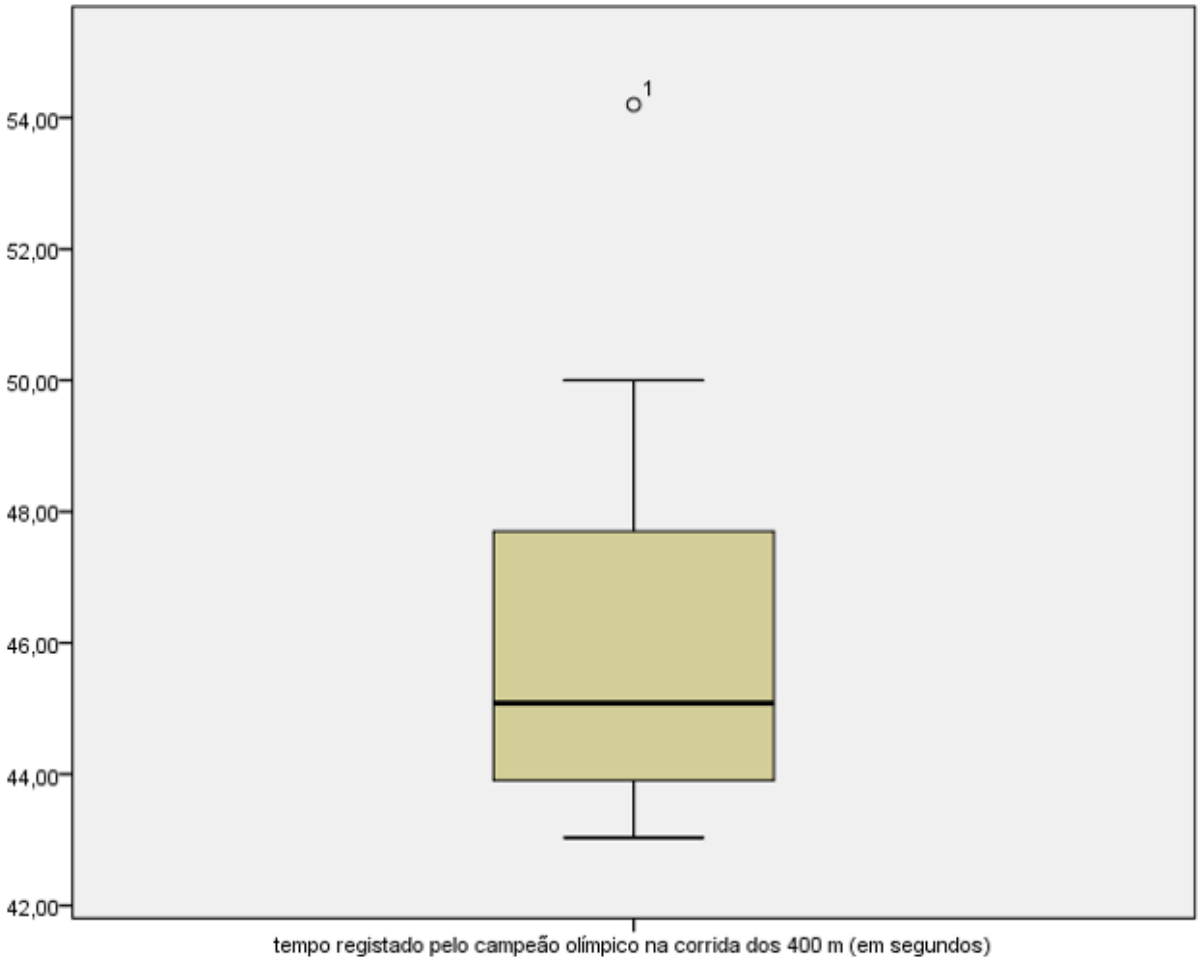
Output do SPSS

		Statistic
N	Valid	28
	Missing	0
Mean		46,0132
Median		45,0800
Mode		43,03[a]
Std. Deviation		2,6567
Variance		7,0580
Skewness		1,3080
Std.Error of Skewness		0,4410
Kurtosis		1,7430
Std.Error of Kurtosis		0,8580
Range		11,1700
Minimum		43,0300
Maximum		54,2000
Percentiles	25	43,8875
	50	45,0800
	75	47,7500

[a] Multiple modes exist. The smallest value is shown

Resposta às questões...

1. Tempo médio dos tempos=46,0132 s
2. Moda? Mediana=45,08 s
3. $x = Q_1 = 43,8875$ s
4. Desvio-padrão=2,66 s; Amplitude dos dados=11,17 s
5. 



- Barreiras moderadas: $43,8875 - 1.5(47,75 - 43,8875) = 38,09$ e $47,75 + 1.5(47,75 - 43,8875) = 53,54$
- Barreiras severas: $43,8875 - 3(47,75 - 43,8875) = 32,3$ e $47,75 + 3(47,75 - 43,8875) = 59,34$

Não existem outliers inferiores. Como o tempo **54,20** ultrapassa a barreira moderada superior mas não chega a ultrapassar a severa superior identificamos como **outlier moderado superior**

6. Assimetria

$Sk = \frac{1.308}{0.441} = 2,966 \rightarrow$ a distribuição dos dados é assimétrica positiva;

Achatamento

$Kt = \frac{1.743}{0.858} = 2.031 \rightarrow$ distribuição dos dados é leptocúrtica.

Covariância e Correlação

Covariância mede o tipo associação linear entre duas amostras de dados quantitativos. Depende das unidades de medida.

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1}\sum_{i=1}^{n} X_i Y_i - \frac{n}{n-1}\overline{X}\,\overline{Y}$$

> 0 associação linear positiva; $= 0$ não existe associação linear; < 0 associação linear negativa

Coeficiente de correlação amostral de Pearson mede o grau de associação linear entre duas amostras de dados quantitativos. Não depende das unidades de medida. Exige a normalidade.

$$R = \frac{S_{xy}}{S_x S_y}, \ -1 \leqslant R \leqslant 1$$

$0 \leqslant |r| < 0,2$ não existe correlação ou é desprezável; $0,2 \leqslant |r| < 0,7$ correlação moderada; $0,7 \leqslant |r| < 0,9$ correlação forte; $|r| \geqslant 0,9$ correlação muito forte.

Correlação positiva ou negativa consoante $r > 0$ ou $r < 0$, respectivamente.

Correlação

Coeficiente de correlação amostral de Spearman mede o grau de associação entre duas amostras de dados quantitativos ou qualitativos ordinais. Método não paramétrico. Não exige a normalidade.

$$R_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \ -1 \leqslant R_S \leqslant 1$$

d_i– diferença entre os valores de ordem dex_i e y_i