Introdução à Probabilidade e Estatística

Universidade de Évora

Departamento de Matemática

Ano lectivo 2018/19

Ana Isabel Santos

Programa

1. Estatística Descritiva (revisões).
2. Introdução às Probabilidades.
3. Variáveis aleatórias discretas e continuas.
4. Vectores aleatórios discretos
5. Principais Distribuições de Probabilidade.
6. Introdução à Amostragem.
7. Estimação Pontual e Intervalos de Confiança.
8. Testes de Hipóteses.
9. Testes não paramétricos: Independência e Ajustamento.
10. Regressão linear simples.

Horário de atendimento

Para alguma dúvida ou questão que queiram colocar, os alunos poderão contactar a docente presencialmente nos seguintes horários de atendimento:

Prof.ª Ana Isabel Santos, (aims@uevora.pt):

2.ª-feira das 14h – 17h

e

5.ª-feira das 18h15min. – 19h15min.,

gab. 241-CLAV.

Estatística Descritiva

Aula 1

Estatística Descritiva

A **Estatística** é uma área do conhecimento que incluí os instrumentos necessários para estudar determinados fenómenos da sociedade. Por exemplo, estudos sobre a natalidade, o emprego, a criminalidade, a opinião sobre … , etc.

- **Estatística Descritiva:** conjunto de técnicas que permitem recolher, organizar e apresentar dados estatísticos.

- **Inferência Estatística:** conjunto de técnicas que permitem caracterizar uma certa população com base na informação amostral.

População vs Amostra

Processo de amostragem

População

Parâmetros
(desconhecidos)

Amostra

Estatísticas
(conhecidas)

Inferência Estatística

Classificação das variáveis

Variáveis Qualitativas: dados com características não numéricas, identificam uma qualidade ou característica.

- **Escala nominal:** dados divididos por categorias sem ordem.

  Exemplo: sexo, distrito de residência, cor dos olhos, nacionalidade.

- **Escala ordinal:** dados divididos por categorias com sequência.

  Exemplo: opinião sobre algo, estado civil, grau de escolaridade.

Variáveis Quantitativas: dados com características numéricas.

- **Discretas:** tomam um n.º finito ou infinito numerável de valores.

  Exemplo: n.º de filhos, n.º de pessoas que vão ao hospital por hora, nota final.

- **Contínuas:** tomam um n.º infinito não numerável de valores.

  Exemplo: idade, altura, peso, temperatura do ar, salário.

Dados

Variável em estudo - População *X*

Amostra em bruto - $(x_1,\ x_2,\ \ldots,\ x_n)$

Amostra ordenada - $(x_{(1)},\ x_{(2)},\ \ldots,\ x_{(n)})$

Dados agrupados - *Tabela de Frequências*

Tabela de frequências para variáveis qualitativas e quantitativas discretas

Dados Agrupados

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	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k'	n_k	f_k	$N_K = n$	$F_k = 1$
	n	1		

- X_i' – **categoria;** ***k*** **– n.º de categorias;** n – **dimensão da amostra;**
- n_i – **frequência absoluta simples;**
- $N_i = \sum_{j=1}^{i} n_j$ – **frequência absoluta acumuladas;**
- $f_i = \frac{n_i}{n}$ – **frequência relativas simples;**
- $F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j$ – **frequência relativas acumuladas.**

Exemplo 1: Regiões onde se localização hotéis com SPA em Portugal.

X – regiões onde existem hotéis com SPA ;

k = 4

X_i'	n_i	f_i	N_i	F_i
Região norte	258	0.2346	258	0.2346
Região centro	395	0.3591	653	0.5937
Região Sul	224	0.2036	877	0.7973
Ilhas	223	0.2027	1100	1
	1100	1		

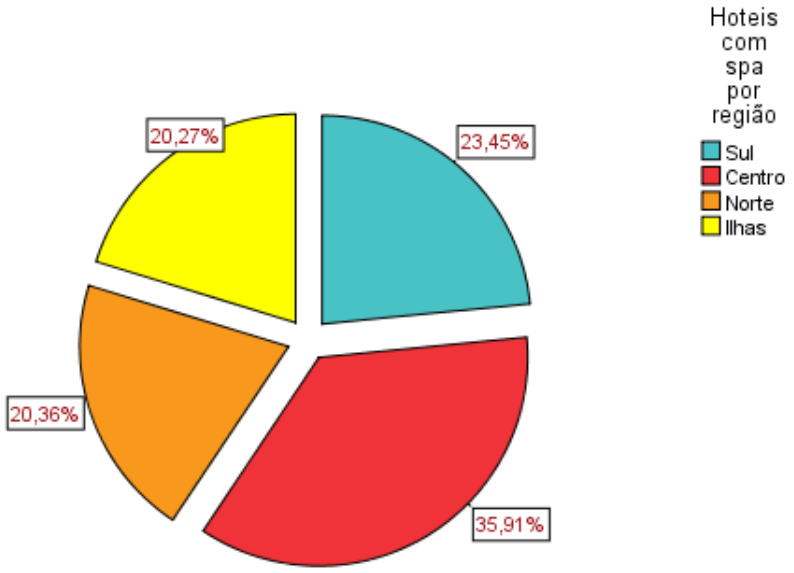
Exemplo 2: Número de livros lidos nas férias de Verão.

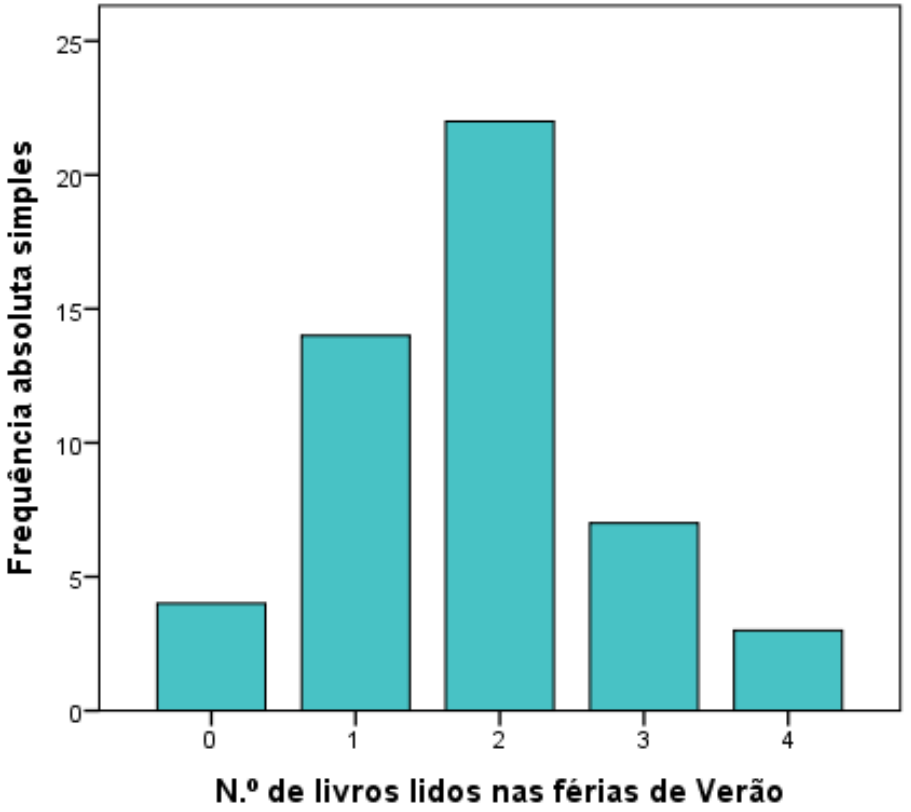
X – N.º de livros lidos nas férias de Verão, por 50 alunos de IPE.

X_i'	n_i	f_i	N_i	F_i
0	4	0.08	4	0.08
1	14	0.28	18	0.36
2	22	0.44	40	0.80
3	7	0.14	47	0.94
4	3	0.06	50	1
	50	1		

Representação gráfica de dados qualitativos e quantitativos discretos

Gráfico Circular e Gráfico de Barras





Para dados de natureza discreta, com um n.º pequeno de valores.

Exercício 1.1:

Considere as notas finais dos alunos de uma determinada unidade curricular:

9 14 12 8 14 12 16 16 8 14 11 12 12 11 11 18 14 18 15 15

a) Os dados em estudo são de tipo qualitativo ou quantitativo?

b) Construa a tabela de frequências associada a estes dados.

c) Represente graficamente a informação disponibilizada.

d) Determine e interprete a média, a moda e a mediana.

e) Calcule a variância e o desvio padrão.

f) Calcule e interprete o valor do percentil 48 e do 8º decil.

g) Indique a amplitude da amostra e a amplitude interquartil.

h) Determine e interprete o coeficiente de variação.

i) Estude a distribuição dos dados quanto à assimetria e ao achatamento.

j) Apresente os dados numa caixa de bigodes.

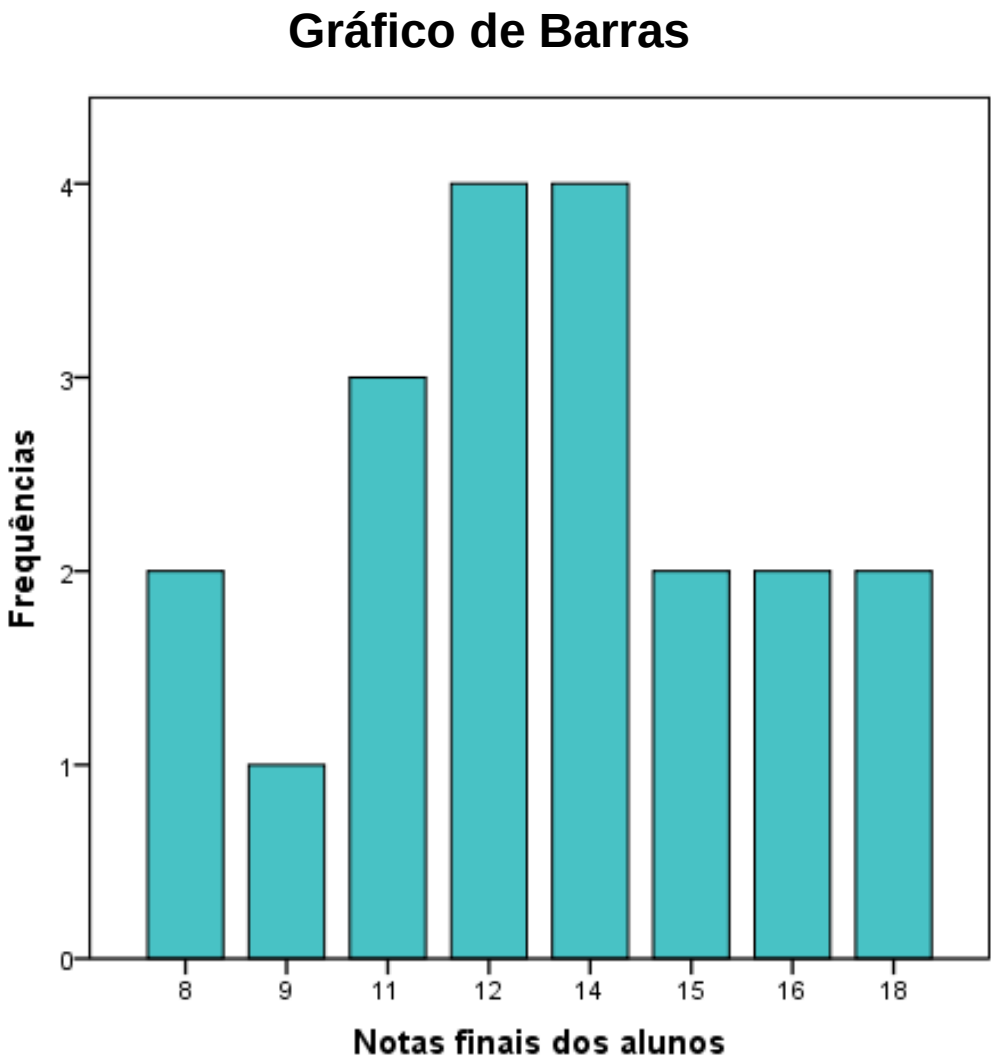
Exercício 1.1:

X – Notas finais dos alunos de uma determinada UC (variável quantitativa discreta)

Tabela de Frequências

X_i'	n_i	f_i	N_i	F_i
8	2	0.10	2	0.10
9	1	0.05	3	0.15
11	3	0.15	6	0.30
12	4	0.20	10	0.50
14	4	0.20	14	0.70
15	2	0.10	16	0.80
16	2	0.10	18	0.90
18	2	0.10	20	1
	20	1		

Exercício 1.1:



Medidas de localização (tendência central)

Média (mean) - $\overline{X}$

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 central)

Mediana (median) - M_e

Valor na amostra ordenada que tem 50% de valores inferiores ou iguais a ele e os restantes 50% superiores ou iguais.

Dados não agrupados ou discretos:

$$M_e = \tilde{X} = \begin{cases} \dfrac{X_{\left[\frac{n}{2}\right]} + X_{\left[\frac{n}{2}+1\right]}}{2} & se \quad n \ par, \\ X_{\left[\frac{n+1}{2}\right]} & se \quad n \ ímpar. \end{cases}$$

Medidas de localização (tendência central)

Moda (mode) - M_0

Dados não agrupados:

$\widehat{X} = M_0$ - Valor mais frequente

Dados agrupados qualitativos ou discretos:

$\widehat{X} = M_0$ - Valor ou categoria com maior frequência simples.

Medidas de localização (tendência não central)

Quantis - Q_p - divide a amostra em duas partes.

Dados não agrupados ou discretos:

$$Q_p = \begin{cases} \dfrac{X_{(np)} + X_{(np+1)}}{2} & se \quad np \ \ inteiro, \\ X_{([np]+1)} & se \quad np \ \ não \ \ inteiro, \end{cases} \qquad \text{para } p \in]0,1[.$$

$$\text{Quantis} \begin{cases} \text{Quartis} = Q_i, & p = \dfrac{i}{4}, \quad i = 1,2,3; \\ \text{Decis} = D_i, & p = \dfrac{i}{10}, \quad i = 1,2,\cdots,9; \\ \text{Percentis} = P_i, & p = \dfrac{i}{100} \quad i = 1,2,\cdots,99; \end{cases}$$

Observações:

$$Q_1 = P_{25}, \qquad Q_2 = \tilde{X} = D_5 = P_{50}, \qquad Q_3 = P_{75}, \qquad D_i = P_{i\times 10}.$$

Medidas de Dispersão

Amplitude total (range): $a = máx - \min$

Amplitude interquartil: $IQ = Q_3 - Q_1$

Intervalo de variação: $Q' = P_{90} - P_{10}$

Variância (variance):

Dados não agrupados

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

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

Dados agrupados

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

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

Desvio-padrão (Std. Deviation): $S = \sqrt{S^2}$

Medidas de Dispersão

Coeficiente de dispersão:

$$CD = \frac{S}{\overline{X}}, \quad 0 \leq CD \leq 1;$$

Coeficiente de variação:

$$CV = \frac{S}{\overline{X}} \times 100\%, \quad 0 \leq CV \leq 100;$$

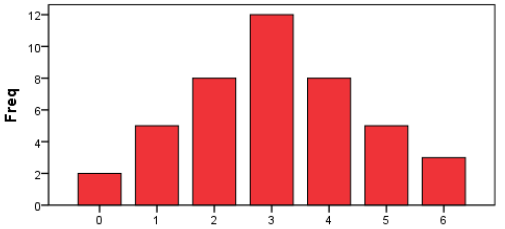
Para valores de *CV* inferiores a 50%, a média será tanto mais representativa dos valores da amostra quanto menor for o valor deste coeficiente. Valores de *CV* superiores a 50% indicam uma baixa representatividade da média.

Medidas de Assimetria (Skewness)

A distribuição dos dados classifica-se, quanto à assimetria, em:

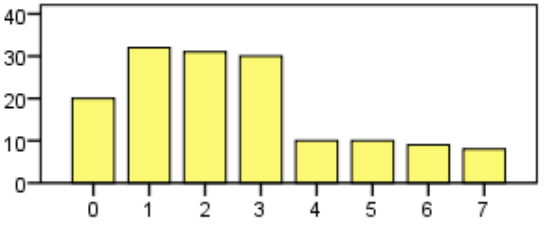
- Simétrica

$$\overline{X} = M_e = M_0$$



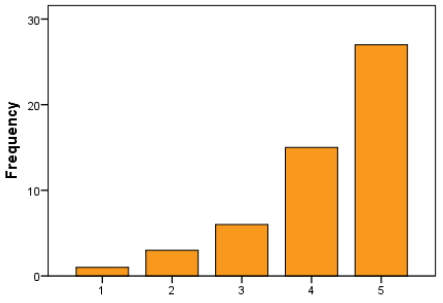
- Assimétrica Positiva ou enviesada à esquerda

$$\overline{X} > M_e > M_0$$



- Assimétrica Negativa ou enviesada à direita

$$\overline{X} < M_e < M_0$$



Medidas de Assimetria (Skewness)

- Grau de Assimetria de Pearson:

$$G_P = \frac{\overline{X} - M_0}{S}, \qquad -3 < G_P < 3$$

- Grau de Assimetria de Bowley:

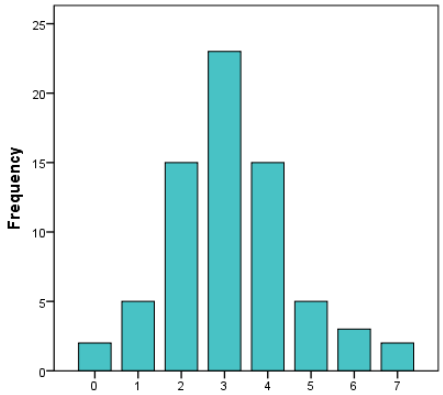
$$G_B = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}, \qquad -1 < G_B < 1$$

O tipo de assimetria é determinado pelo sinal de G_P ou de G_B

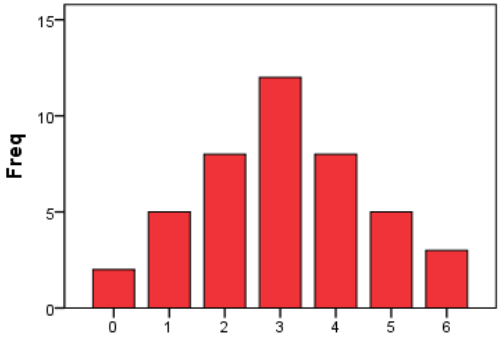
> 0 – **Assimétrica "+"**; = 0 – **Simétrica**; < 0 – **Assimetrica "–"**

Medidas de Achatamento (Kurtosis)

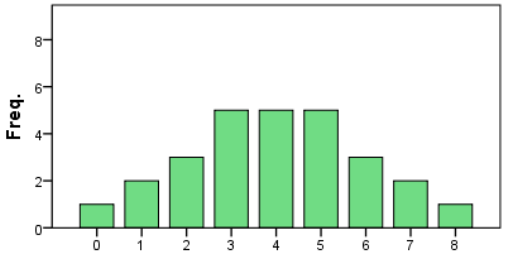
A distribuição dos dados classifica-se, quanto ao achatamento, em:



Leptocúrtica



Mesocúrtica



Platicúrtica

Coeficiente Percentil de Achatamento:

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} \begin{cases} < 0.263 & \text{Distribuição Leptocúrtica} \\ = 0.263 & \text{Distribuição Mesocúrtica} \\ > 0.263 & \text{Distribuição Platicúrtica} \end{cases}$$

Medidas de Assimetria e Achatamento no SPSS

Coeficiente de assimetria do SPSS:

$$g_{SPSS} = \frac{Skewness}{Std.\ error\ of\ Skewness}.$$

- Se $g_{SPSS} < -1,96$, então assume-se que a distribuição é **Assimétrica Negativa**;
- Se $|g_{SPSS}| < 1,96$, então assume-se que a distribuição é **Simétrica**;
- Se $g_{SPSS} > 1,96$, então assume-se que a distribuição é **Assimétrica Positiva**.

Coeficiente de achatamento do SPSS:

$$K_{SPSS} = \frac{Kurtosis}{Std.\ error\ of\ Kurtosis}.$$

- Se $K_{SPSS} < -1,96$, então assume-se que a distribuição é **Platicúrtica**;
- Se $|K_{SPSS}| < 1,96$, então assume-se que a distribuição é **Mesocúrtica**;
- Se $K_{SPSS} > 1,96$, então assume-se que a distribuição é **Leptocúrtica**.

Caixa de bigodes, diagrama de extremos e quartis ou Boxplot

Caixa de Bigodes: apresenta algumas das principais características descritivas de um certo conjunto de dados, numa imagem compacta. Encontram-se representados Q_1, Q_2, Q_3, IQ, o menor valor não outlier, o maior valor não outlier e os outliers.

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

Nota: Nas boxplot do SPSS os outliers moderados são assinalados com um "O" e os outliers severos com um "$*$".

Outliers

Um **Outlier** é um valor que se afasta de modo evidente do centro da distribuição.

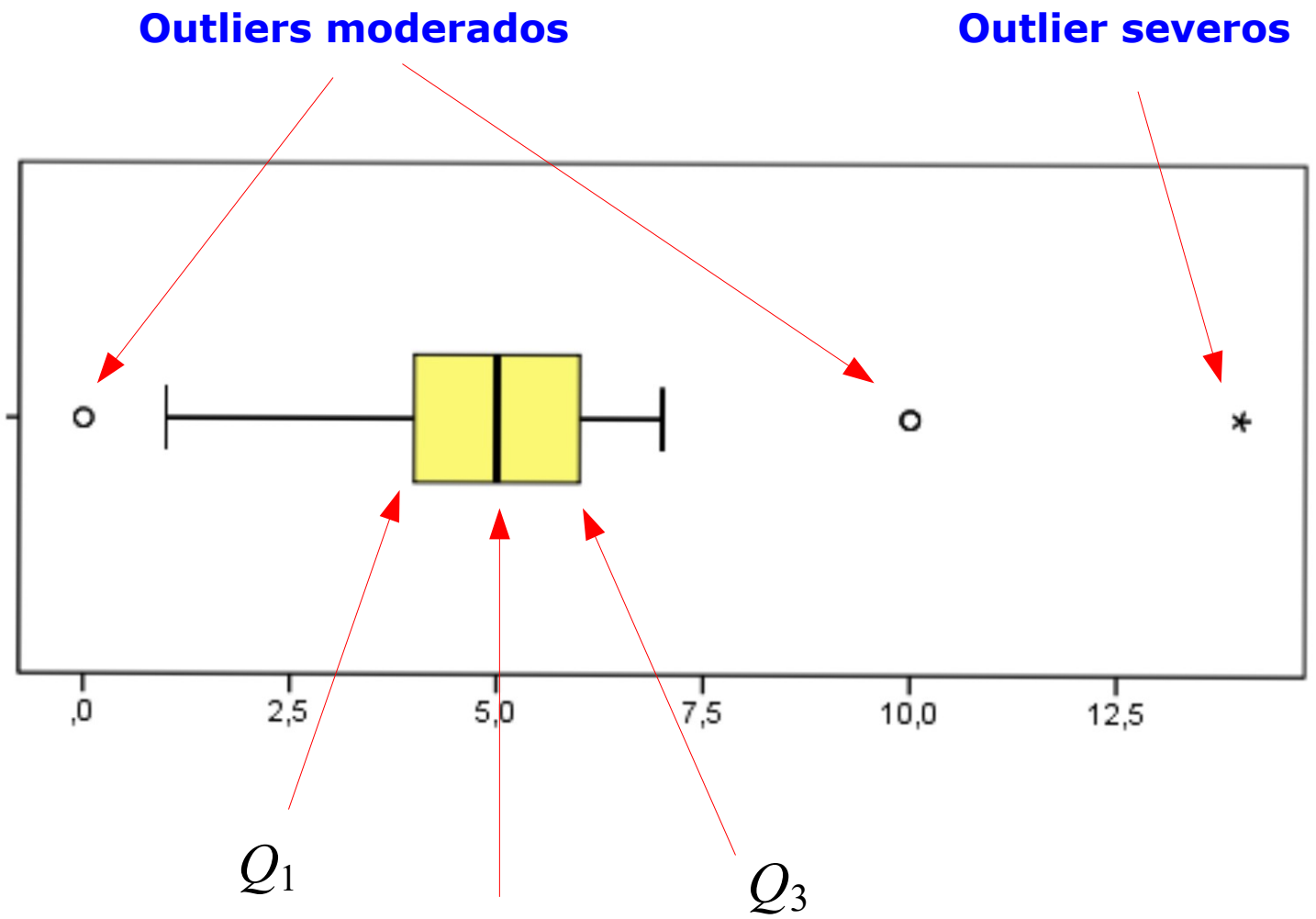
X_i é um **outlier moderado** se ultrapassa uma das barreiras **moderadas:**

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

X_i é um **outlier severo** se ultrapassa uma das barreiras **severas:**

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

Caixa de bigodes, diagrama de extremos e quartis ou Boxplot



Exercício 1.1: *Outputs do SPSS*

Statistics		
Notas finais dos alunos		
N	Valid	20
	Missing	0
Mean		13,00
Median		13,00
Mode		12[a]
Std. Deviation		2,920
Skewness		-,042
Std. Error of Skewness		,512
Kurtosis		-,552
Std. Error of Kurtosis		,992
Percentiles	25	11,00
	50	13,00
	75	15,00

a. Multiple modes exist.
The smallest value is shown

Notas finais dos alunos					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	2	10,0	10,0	10,0
	9	1	5,0	5,0	15,0
	11	3	15,0	15,0	30,0
	12	4	20,0	20,0	50,0
	14	4	20,0	20,0	70,0
	15	2	10,0	10,0	80,0
	16	2	10,0	10,0	90,0
	18	2	10,0	10,0	100,0
	Total	20	100,0	100,0	

Exercício 1.1: *Resolução no SPSS*

- **No OUTPUT do SPSS**

GRAPHS

LEGACY DIALOGS

BOX-PLOT

SIMPLE

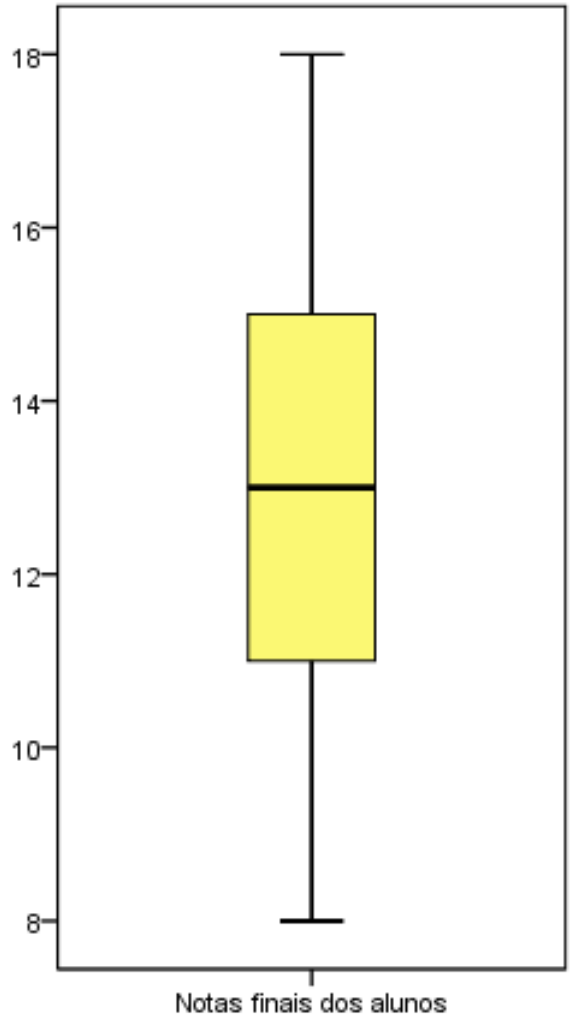
DATA IN CHART ARE

SUMMARIES OF SEPARATE VARIABLES

DEFINE

selecionar a var (notas_fin) para boxes represent

OK



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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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;
- $X_i' = \frac{l_i + L_i}{2}$ - pontos médios das classes.

Exemplo 3: Notas da 2.ª freq. de EACHS (2012/13).

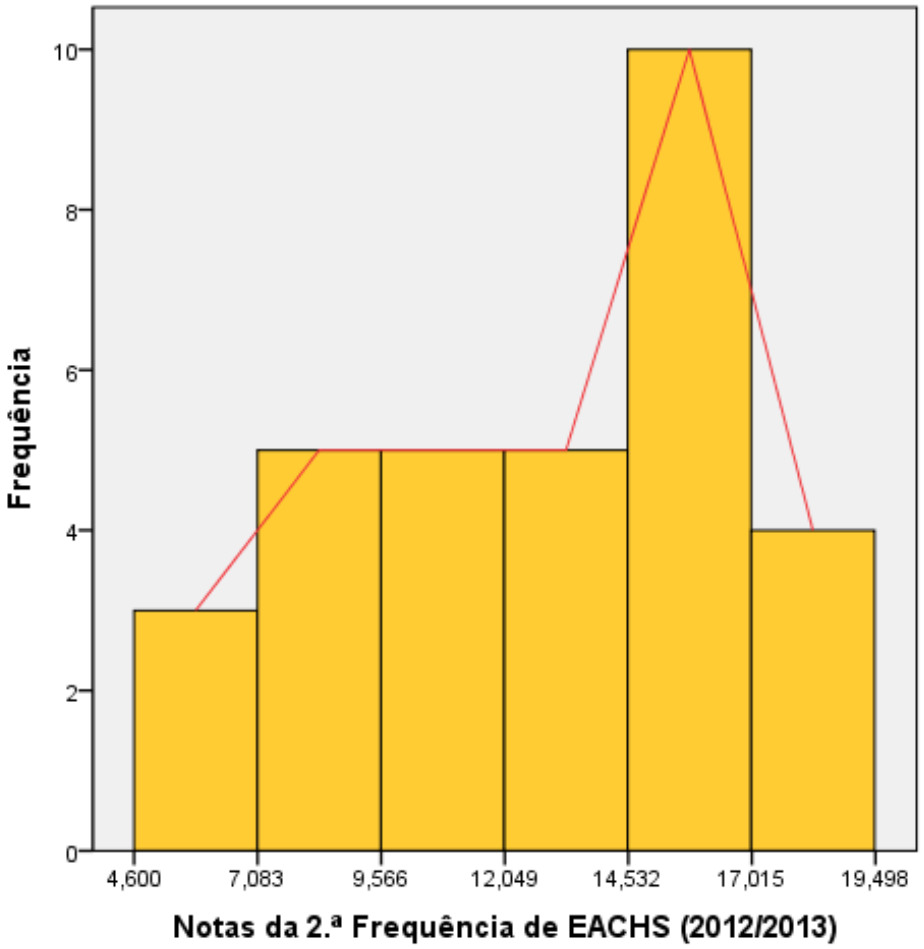
Dados agrupados

$$n = 32, \quad k = 6$$

C_i	X_i'	n_i	$f_i(\%)$	N_i	$F_i(\%)$
$[4.6, 7.083[$	5.84	3	9.375	3	9.375
$[7.083, 9,566[$	8.32	5	15.625	8	25
$[9.566, 12.049[$	10.81	5	15.625	13	40.625
$[12.049, 14.532[$	13.29	5	15.625	18	56.25
$[14.532, 17.015[$	15.77	10	31.25	28	87.5
$[17.015, 19.5[$	18.26	4	12.5	32	100
		32	1		

Representação gráfica de dados quantitativos contínuos

Histograma e Polígono de frequências



Para dados de natureza contínua ou quando o n.º de valores distintos é muito elevado.

Exercício 1.3:

Recolheu-se a seguinte informação diária referente à humidade relativa (em %) e à temperatura máxima (em ºC) na determinada estação meteorológica no mês de agosto de 2014:

Statistics

Áreas		Humidade	Temperatura
N	Valid	31	31
	Missing	0	0
Mean		53,224	23,252
Median		55,273	21,997
Variance		169,278	32,464
Skewness		–0,867	0,948
Std. Error of Skewness		0,421	0,421
Kurtosis		0,172	–0,4108
Std. Error of Kurtosis		0,821	0,821
Minimum		22,187	15,895
Maximum		71,215	35,335
Percentiles	10	33,137	17,763
	25	45,246	18,804
	75	62,148	24,613
	90	69,092	32,748

Covariância

Covariância: mede o grau de associação linear entre duas variáveis quantitativas obtidas do mesmo indivíduo ou unidade.

Covariância amostral:

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

Covariância populacional:

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

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

Correlação

Coeficiente de Correlação de Pearson: mede o grau de associação linear entre duas variáveis quantitativas e assume a normalidade dos dados. Não depende das unidades de medida.

$$r = \frac{Cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}, \quad -1 \leq r \leq 1$$

Se $0 \leq |r| < 0.2$, não existe associação linear entre as variáveis ou é desprezível;

Se $0.2 \leq |r| < 0.7$, existe associação linear moderada;

Se $0.7 \leq |r| < 0.9$, existe associação linear forte;

Se $|r| \geq 0.9$, existe associação linear muito forte.

Exemplo 4: Altura e peso de 10 alunos de Eng. Informática.

Mediram-se e pesaram-se 10 alunos do curso de Eng. Informática.

Na tabela abaixo apresentam-se as alturas (em m) e os pesos (em kg).

Aluno	1	2	3	4	5	6	7	8	9	10
Altura (m)	1,74	1,83	1,68	1,89	1,72	1,72	1,75	1,78	1,83	1,71
Peso (kg)	68	80	62	89	68	70	71	70	78	65

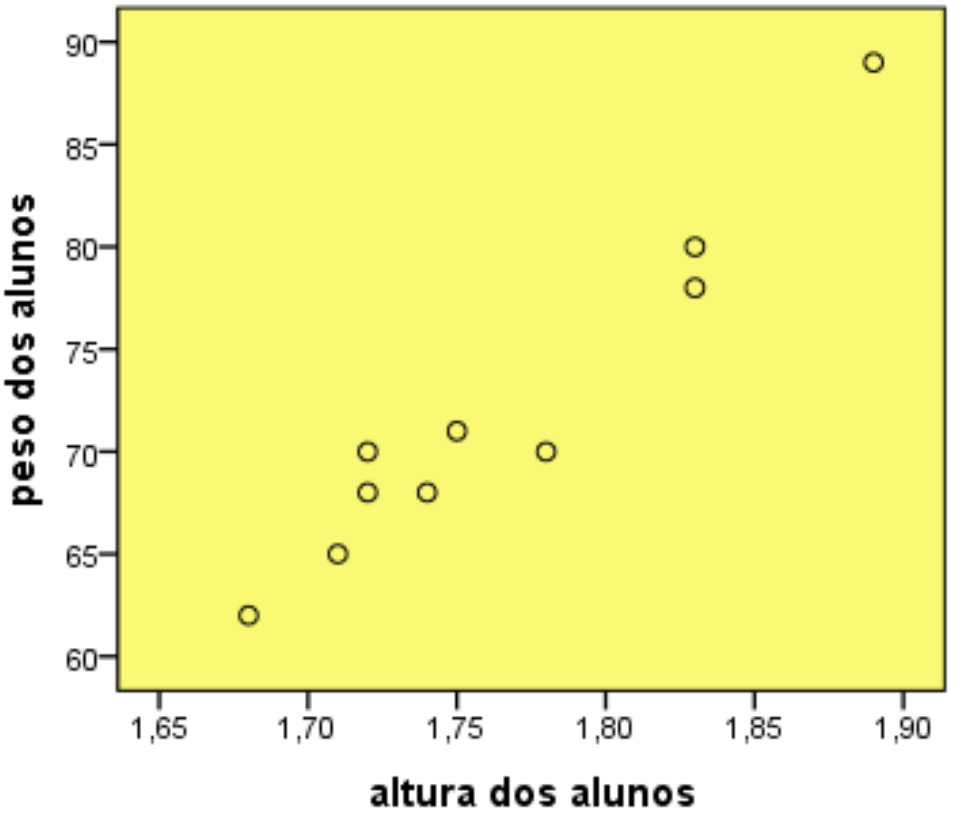
a) Represente graficamente a altura e o peso.

b) Calcule a altura e o peso médios e as respectivas variâncias.

c) Calcule o coeficiente de correlação de Pearson entra a altura e o peso.

Exemplo 4:

a)


peso dos alunos
90
85
80
75
70
65
60
1,65
1,70
1,75
1,80
1,85
1,90
altura dos alunos

Exemplo 4:

b) Médias:

X - altura em metros

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} X_i = \frac{1,74 + 1,83 + 1,68 + 1,89 + 1,72 + 1,72 + 1,75 + 1,78 + 1,83 + 1,71}{10}$$

$$= 1,765.$$

Logo, a altura média dos alunos de Eng. Informática é 1,765 *m*.

Y - peso em kg

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} Y_i = \frac{68 + 80 + 62 + 89 + 68 + 70 + 71 + 70 + 78 + 65}{10} = 72,1$$

Logo, o peso médio dos alunos de Eng. Informática é 72,1 *kg.*

Exemplo 4:

Variâncias amostrais:

X - altura

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) = \frac{1,74^2 + 1,83^2 + \cdots + 1,71^2 - 10 \times 1,765^2}{9} = 0,004.$$

Y - peso

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right) = \frac{68^2 + 80^2 + \cdots + 65^2 - 10 \times 72,1^2}{9} = 64,322.$$

Output do SPSS:

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
altura dos alunos	10	1,7650	,06621	,004
peso dos alunos	10	72,10	8,020	64,322
Valid N (listwise)	10			

Exemplo 4:

c) ***Covariância amostral:***

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

$$= \frac{1,74 \times 68 + 1,83 \times 80 + \cdots + 1,71 \times 65 - 10 \times 1,765 \times 72,1}{9} = 0,514$$

Coeficiente de correlação de Pearson:

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{0,514}{\sqrt{0,004}\,\sqrt{64,322}} = 0,968.$$

Portanto, existe uma relação linear positiva muito forte entre o peso e a altura dos alunos, donde a maiores alturas estão associados pesos mais elevados e a menores alturas estão associados pesos mais baixos.

Note-se que foi assumida a normalidade das distribuições para garantir a aplicabilidade deste coeficiente.

Exercício 4:

No quadro seguinte indicam-se os preços (X) de pacotes de arroz (em euros) praticado durante 12 meses consecutivos e o número de pacotes vendidos (Y) em cada mês:

X	1.10	0.90	0.80	0.76	0.74	0.71	0.70	0.65	0.63	0.60	0.55	0.50
Y	55	70	90	100	90	105	80	110	125	115	130	131

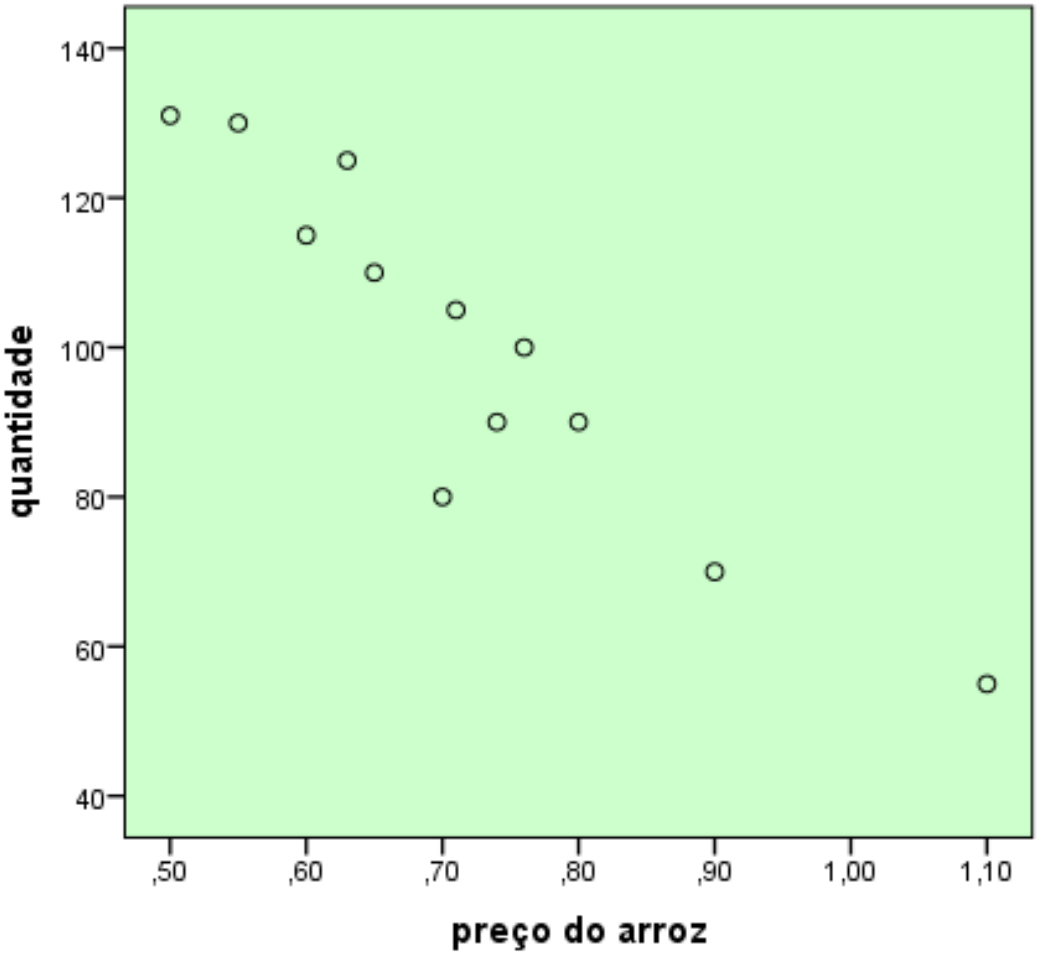
a) Represente graficamente a informação.

b) Comente a seguinte afirmação: "Parece existir relação linear entre as duas variáveis".

c) Calcule o valor do coeficiente de correlação de Pearson.

d) Responda novamente à alínea b).

Exercício 4: *Resolução no SPSS*



Exercício 4: *Resolução no SPSS*

Correlations

		preço do arroz	quantidade
preço do arroz	Pearson Correlation	1	-,926**
	Sig. (2-tailed)		,000
	N	12	12
quantidade	Pearson Correlation	-,926**	1
	Sig. (2-tailed)	,000	
	N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

Exercício 4: *Resolução no SPSS*

Correlations

		preço do arroz	quantidade
preço do arroz	Pearson Correlation	1	-,926**
	Sig. (2-tailed)		,000
	N	12	12
quantidade	Pearson Correlation	-,926**	1
	Sig. (2-tailed)	,000	
	N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

Como r = -0.926 < -0.9, podemos concluir que existe associação linear negativa muito forte entre o preço do arroz e a quantidade vendida, isto é, quanto mais elevado é o preço do arroz menor é a quantidade vendida.