Estatística

Aprendizagens essenciais 11º ano

- Reconhecer o papel relevante desempenhado pela Estatística em todos os campos do conhecimento abordando nomeadamente os conceitos de Recenseamento e Sondagem (população e amostra);
- Organizar e interpretar dados de natureza quantitativa e qualitativa, variáveis discretas e contínuas;
- Interpretar medidas de localização de uma amostra: moda, média, mediana, quartis e percentis; medidas de dispersão: amplitude interquartil, variância, desvio padrão;
- Abordar gráfica e intuitivamente distribuições bidimensionais, nomeadamente o diagrama de dispersão, o coeficiente de correlação e reta de regressão.

Simbologia:

$\underset{\sim}{x}$ amostra da variável x

$\underset{\sim}{x} = (\underbrace{x_1, x_2, x_3, x_4, \ldots x_n}), \text{ com } n \in IN$

Valores da amostra (podem estar ordenados)

Exemplo :

$$\underset{\sim}{x} = (\underset{x_1}{3}, \underset{x_2}{2}, \underset{x_3}{2}, \underset{x_4}{3}, \underset{x_5}{5}, \underset{x_6}{1}, \underset{x_7}{1}, \underset{x_8}{2})$$

Valores da amostra ordenados

$$\underset{\sim}{x} = (\underset{x_{(1)}}{1}, \underset{x_{(2)}}{1}, \underset{x_{(3)}}{2}, \underset{x_{(4)}}{2}, \underset{x_{(5)}}{2}, \underset{x_{(6)}}{3}, \underset{x_{(7)}}{3}, \underset{x_{(8)}}{5})$$

Como alguns valores da amostra estão repetidos então podemos utilizar outra notação onde só apresentamos os valores distintos:

$$\tilde{x} = (1, 2, 3, 5) = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4) \text{ , com } n_1 = 2;\ n_2 = 3;\ n_3 = 2;\ n_4 = 1$$

frequência absoluta de cada dado distinto

Média

A média é dada por: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Somatório:

Seja (a_n) uma sucessão definida por $a_n = 2n - 1$

$$\sum_{i=1}^{5} a_i = a_1 + a_2 + a_3 + a_4 + a_5 = 1 + 3 + 5 + 7 + 9 = 25$$

Média para dados agrupados

No caso de dados agrupados, a média da amostra é dada por:

$$\bar{x} = \frac{\sum_{j=1}^{m} \tilde{x}_j n_j}{n}$$

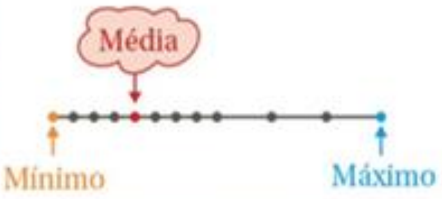
n_j representa a frequência absoluta da variável $\tilde{x}_j$.

No caso do nosso exemplo :

$\tilde{x} = (1, 2, 3, 5) = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4)$, com $n_1 = 2$; $n_2 = 3$; $n_3 = 2$; $n_4 = 1$

$$\bar{x} = \frac{\sum_{i=1}^{4} \tilde{x}_i n_i}{8} = \frac{1 \times 2 + 2 \times 3 + 3 \times 2 + 5}{8} = \frac{19}{8} = 2{,}375$$

Propriedades da média



Propriedade

A **média** situa-se entre o mínimo e o máximo da **amostra** e não pode ser igual ao mínimo sem ser também igual ao máximo, o que acontece **se e somente se** a amostra for constante.

Propriedade

Adicionando-se a cada um dos valores x_i uma constante h $(h \neq 0)$ a nova média será igual à média original adicionada de $\boldsymbol{h}$.

Propriedade

Multiplicando x_i por uma constante a $(a \neq 0)$ a nova média será igual ao produto da média original por $\boldsymbol{a}$.

No caso do nosso exemplo :

$$\tilde{x} = (1, 2, 3, 5) = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4) \text{ , com } n_1=2;\ n_2=3;\ n_3=2;\ n_4=1 \ ; \quad \bar{x} = \frac{19}{8} = 2{,}375$$

$$\tilde{y} = 3\tilde{x} = (3, 6, 9, 15) \qquad \bar{y} = \frac{3\times 2 + 6\times 3 + 9\times 2 + 15}{8} = \frac{57}{8} = 3\times\frac{19}{8} = 3\times\bar{x}$$

Propriedade

Dados $n \in \mathbb{N}$, uma amostra $\underset{\sim}{x}=(x_1, x_2, \ldots, x_n)$ e os números reais não nulos h e a, sendo $\underset{\sim}{y}=(ax_1+h, ax_2+h, \ldots, ax_n+h)$, tem-se que $\bar{y}=a\bar{x}+h$.

Desvio em relação à média

Dados $n \in \mathbb{N}$ e uma amostra $\underset{\sim}{x}=(x_1, x_2, \ldots, x_n)$, com $i \in \{1, \ldots, n\}$, designa-se por **desvio de x_i em relação à média**:

$$d_i = x_i - \bar{x}$$

Propriedade

A soma dos desvios em relação à média dos valores observados é nula.

No caso do nosso exemplo :

x_i	3	2	2	3	5	1	1	2	Soma
$d_i=x_i - \bar{x}$	0,625	-0,375	-0,375	0,625	2,625	-1,375	-1,375	-0,375	0

Telensino

1. Dos 100 bonsais existentes num horto selecionaram-se dez, registando-se as respetivas alturas em centímetros. A amostra obtida é a seguinte:

$\underset{\sim}{x} = (11{,}1\ ;\ 12{,}5\ ;\ 7{,}8\ ;\ 11{,}1\ ;\ 12{,}0\ ;7{,}8\ ;\ 11{,}1\ ;\ 12{,}0\ ;\ 7{,}8\ ;\ 10{,}0)$

1.1 Indique $x_{(1)}$ e $x_{(5)}$.

1.2 Explicite $\tilde{x}$.

1.3 Determine a média da amostra utilizando a fórmula para os dados agrupados.

Máximo 10
Porto Editora

Telensino

1.1 $x_{(1)} = 7{,}8$; $x_{(5)} = 11{,}1$

$\underset{\sim}{x}$=(7,8 ; 7,8; 7,8; 10,0 ; 11,1 ;11,1; 11,1; 12,0 ; 12,0; 12,5)

1.2 $\tilde{x}$ =(7,8 ; 10,0 ; 11,1 ; 12,0 ; 12,5)

$\underset{\sim}{y}$=(7,8 ; 7,8; 10,0 ; 11,1 ;11,1; 11,1; 12,0 ; 12,0; 12,5; **30**)

1.3 $$\bar{x} = \frac{\sum_{i=1}^{5} \tilde{x}_i n_i}{10} = \frac{7{,}8\times 3 + 10{,}0\times 1 + 11{,}1\times 3 + 12{,}0\times 2 + 12{,}5\times 1}{10} = \frac{103{,}2}{10} = 10{,}32$$

1.4 $$\bar{y} = \frac{\sum_{i=1}^{6} \tilde{y}_i n_i}{10} = \frac{7{,}8\times 2 + 10{,}0\times 1 + 11{,}1\times 3 + 12{,}0\times 2 + 12{,}5\times 1 + 30\times 1}{10} = \boxed{12{,}54}$$

A **média** é uma característica amostral "com pouca resistência"

Do ponto de vista da Física, a média de uma amostra pode ser interpretada como o ***centro de gravidade de um segmento de reta***, contido numa reta numérica, no qual se colocou, para cada valor x_i da amostra, um ponto material de abcissa x_i de massa unitária.

A média é o "ponto de equilíbrio" da distribuição, dando-nos o "centro de gravidade" da amostra.

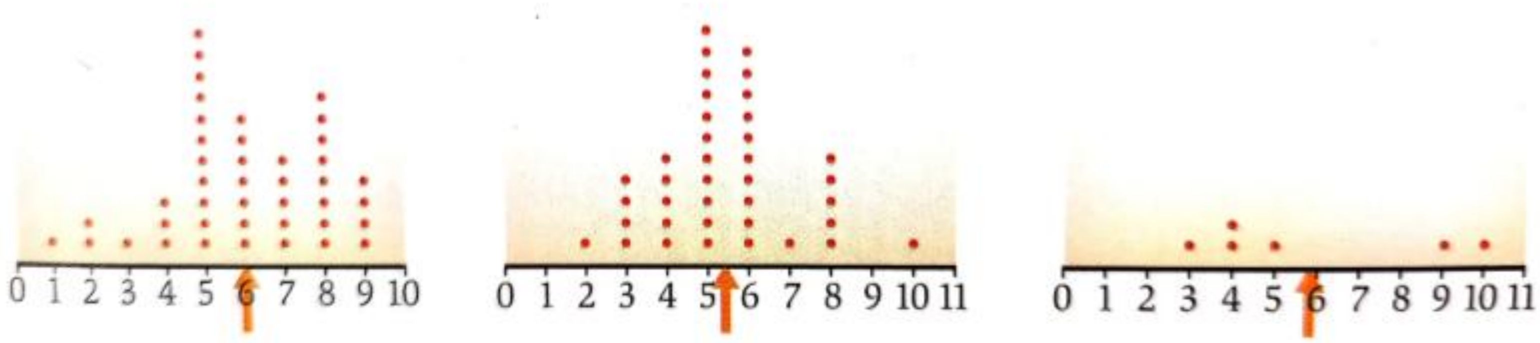


Imagem do manual Matemática 10
Areal

A média é uma <u>medida de localização.</u>
A moda e a mediana já estudada no 3º ciclo também são medidas de localização.

Variância e Desvio Padrão de uma amostra

Telensino

Medidas de dispersão mais comuns :

- a amplitude interquartis (já estudada no 3º ciclo)
- a variância
- o desvio padrão

No nosso exemplo:

$$\tilde{x} = (1, 2, 3{,}5)\ , \text{com } n_1=2;\ n_2=3;\ n_3=2;\ n_4=1$$

x_i	3	2	2	3	5	1	1	2	Soma
$d_i = x_i - \bar{x}$	0,625	-0,375	-0,375	0,625	2,625	-1,375	-1,375	-0,375	0
$(x_i - \bar{x})^2$	0,3906	0,1406	0,1406	0,3906	6,8906	1,8906	1,8906	0,1406	11,8748

Soma dos quadrados dos desvios

A **soma dos quadrados dos desvios dos** x_i **em relação à média** designa-se por SS_x e é dada por:

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Soma dos quadrados dos desvios para dados agrupados

Para uma amostra $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$, $n \in \mathbb{N}$, $SS_x = \sum_{j=1}^{m} (\tilde{x}_j - \bar{x})^2 n_j$, onde $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m$ representam os m valores distintos da amostra $\tilde{x}$ e n_j a frequência absoluta de $\tilde{x}_j$.

A variância e o desvio padrão são <u>medidas de dispersão</u>, indicam-nos a dispersão dos valores em torno da média.

A variância mede-se em $(unidades\ de\ x)^2$ e o desvio padrão em $unidades\ de\ x$, daí utilizar o desvio padrão em vez da variância.

Variância e desvio-padrão

Dados $n \in \mathbb{N}$ $(n > 1)$ e uma amostra $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$, $s_x^2 = \dfrac{SS_x}{n-1}$ designa-se por **variância** e $s_x = \sqrt{\dfrac{SS_x}{n-1}}$ por **desvio-padrão** da amostra $\underset{\sim}{x}$.

No nosso exemplo: $\tilde{x} = (1, 2, 3, 5)$, com $n_1=2;\ n_2=3;\ n_3=2;\ n_4=1$

x_i	3	2	2	3	5	1	1	2	Soma
$d_i = x_i - \bar{x}$	0,625	-0,375	-0,375	0,625	2,625	-1,375	-1,375	-0,375	0
$(x_i - \bar{x})^2$	0,3906	0,1406	0,1406	0,3906	6,8906	1,8906	1,8906	0,1406	11,8748

$$s_x = \sqrt{\frac{SS_x}{n-1}} = \sqrt{\frac{11{,}8748}{7}} = 1{,}3024$$

Na calculadora gráfica:

Rad Norm1 d/c Real
1-Variable
$\bar{x}$ =2.375
Σx =19
Σx^2 =57
σx =1.21834929
sx =1.30247018
n =8

Relações bidimensionais

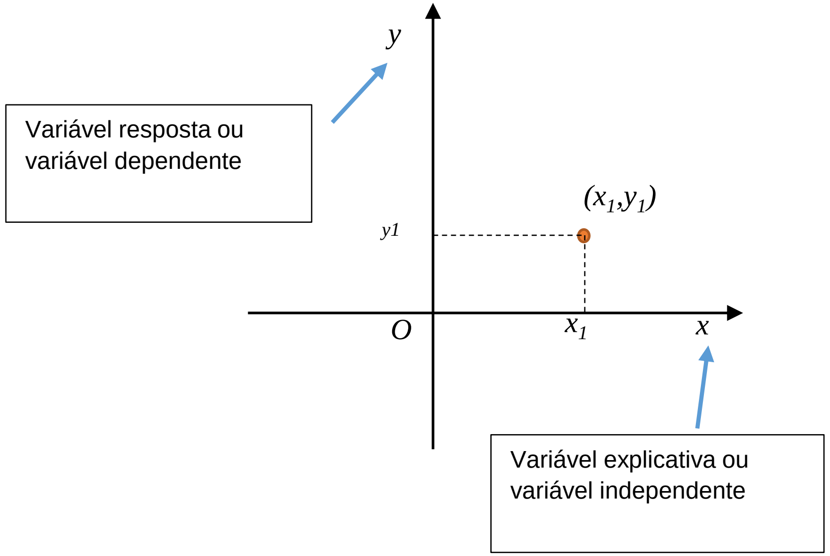
Amostra bivariada

Amostra bivariada

Dadas duas variáveis estatísticas quantitativas x e y em determinada população e uma amostra A de dimensão $n \in \mathbb{N}$ dessa população cujos elementos estão numerados de 1 a n, a sequência:

$$((x_1 , y_1) , (x_2 , y_2) , \dots , (x_n , y_n))$$

designa-se por **amostra bivariada** das variáveis estatísticas x e y ou **amostra de dados bivariados** e representa-se por $(x \underset{\sim}{,} y)$.
O número natural n designa-se por dimensão da amostra bivariada.



Nuvem de pontos

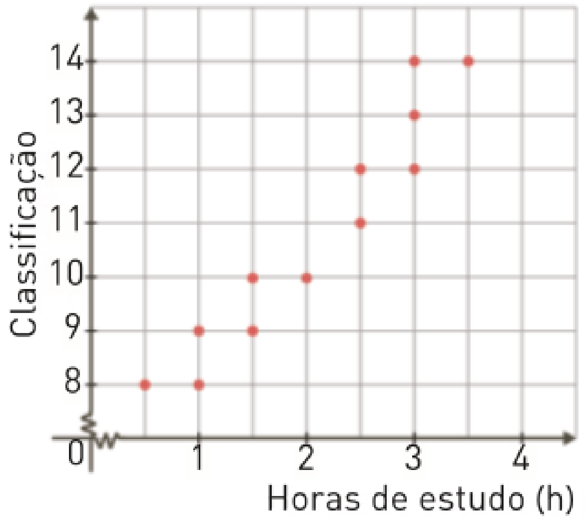
Fixado um referencial ortonormado num plano e uma amostra de dados bivariados quantitativos da forma:

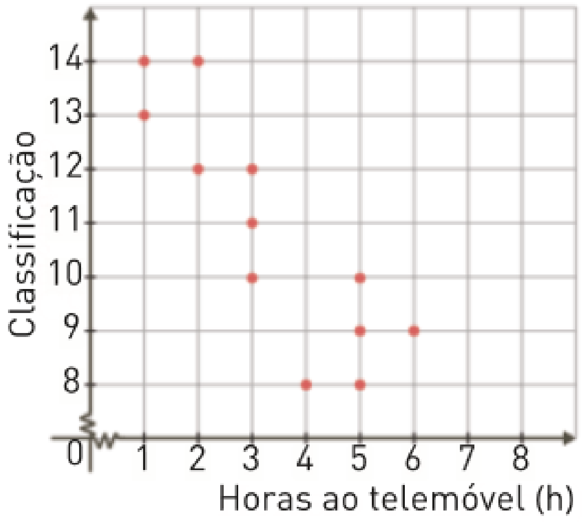
$$\left(x \underset{\sim}{,} y\right) = \left((x_1\,,\,y_1)\,,\,(x_2\,,\,y_2),\ldots,\,(x_n\,,\,y_n)\right),\ n \in \mathbb{N}$$

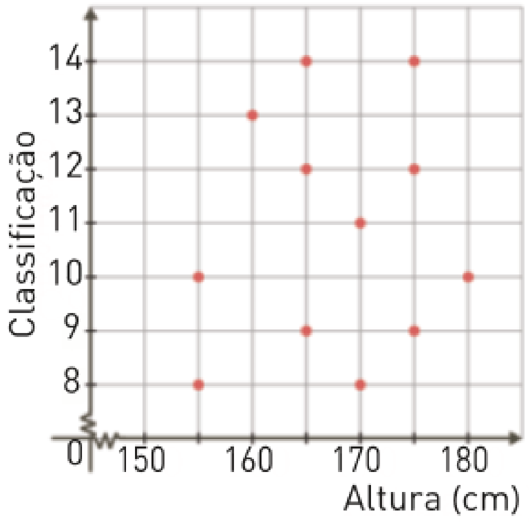
designa-se por **nuvem de pontos** o conjunto definido por:

$$\left\{P_1\,(x_1\,,\,y_1)\,,\ P_2\,(x_2\,,\,y_2),\ldots,\ P_n\,(x_n\,,\,y_n)\right\},\ n \in \mathbb{N}$$

Nota: também se chama diagrama de dispersão







1. Na tabela junta estão registados os dados referentes ao número de vezes, por ano, que os funcionários de um determinado armazém chegavam atrasados e a distância, em quilómetros, a que este viviam do armazém.

Distância (Km)	Número de atrasos
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

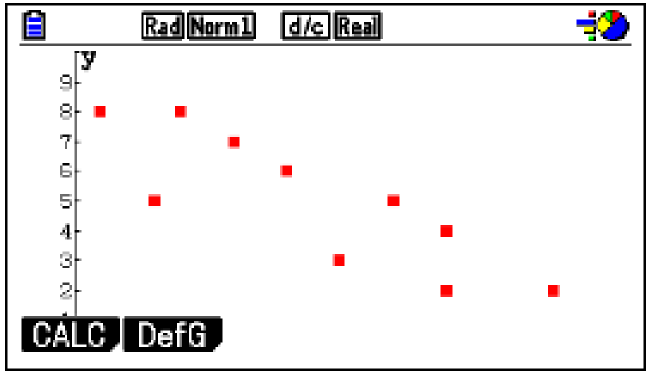
1.1 Represente, num referencial ortonormado, a nuvem de pontos que representa esta amostra e refira se é razoável a existência de uma relação linear entre as duas variáveis.

1.2 Obtenha a representação da reta de regressão

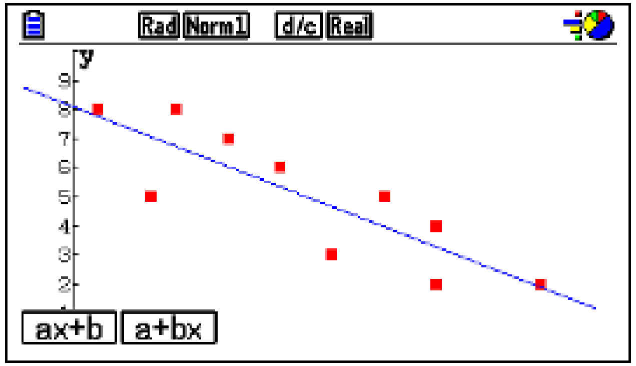
1.3 Utilizando a equação obtida na alínea anterior determine O número de atrasos esperados para um funcionário que vive a 5 quilómetros do armazém.

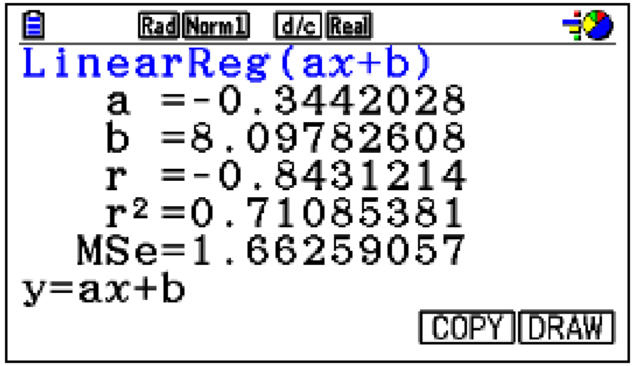
Dimensões 11
Santilhana

1.1



1.2 Obtenha a representação da reta de regressão





1.3 Utilizando a equação obtida na alínea anterior determine o número de atrasos esperados para um funcionário que vive a 5 quilómetros do armazém.

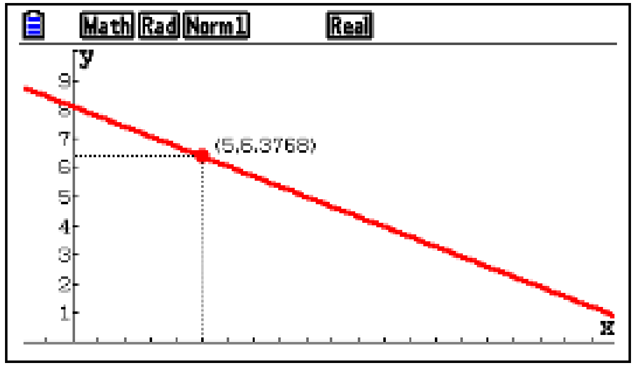
Analiticamente:

$$y = -0{,}34x + 8{,}10$$

$$x = 5 \Rightarrow y = -0{,}34 \times 5 + 8{,}10 = 6{,}4$$

R: Espera-se que o funcionário chegue atrasado 6 vezes durante um ano

Na calculadora gráfica:



Relação linear

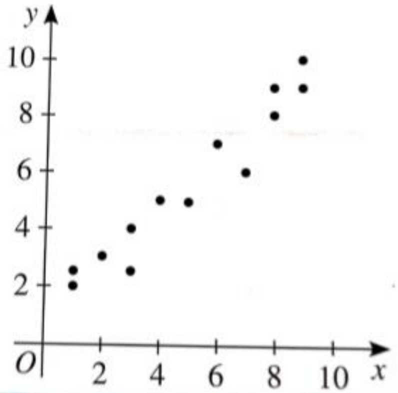
Coeficiente de correlação linear

Dado um número natural n e uma amostra de dados bivariados quantitativos $(x \underset{\sim}{,} y)$, designa-se por **coeficiente de correlação linear** e representa-se por r:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{SS_x \, SS_y}}$$

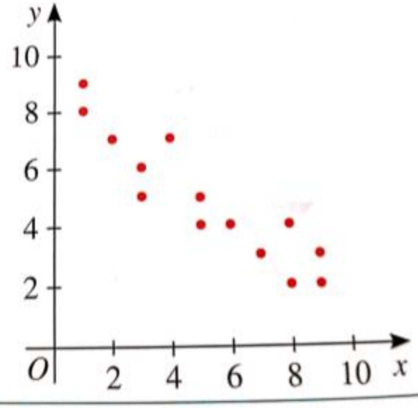
A reter:

- $|r| \leq 1 \Leftrightarrow -1 \leq r \leq 1$
- Se $r >$ o a associação linear entre as variáveis estatísticas é positiva.
- Se $r <$ o a associação linear entre as variáveis estatísticas é negativa.
- A associação linear entre as variáveis estatísticas é tanto mais forte quanto mais perto de 1 estiver $|r|$.



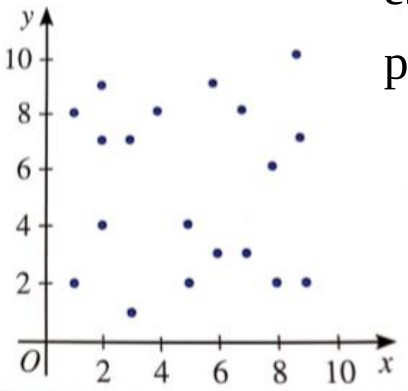
Associação linear positiva forte

$r > 0$



Associação linear negativa forte

$r < 0$



Variáveis independentes

4. O Sr. Silva aquece a sua casa com gás natural. A quantidade de gás utilizada depende da temperatura exterior e o Sr. Silva pretende fazer um estudo dos gastos durante os 9 meses em que se observam menores temperaturas, para poder estabelecer uma previsão para os gastos em função da temperatura exterior. Na tabela junta estão registadas as temperaturas médias observadas em cada um dos meses (em graus Celsius) e o respetivo volume de gás despendido pelo Sr. Silva (em metros cúbicos).

mês	out	nov	dez	jan	fev	mar	abr	mai	jun
temperatura	16,1	12,4	10,3	8,9	10,1	12,8	13,2	15,9	16,4
Volume de gás	0,01	0.10	0,24	0,26	0,19	0,09	0,05	0,03	0,01

4.1. Qual deve ser a variável explicativa e a variável resposta?

4.2. Utilize uma folha de cálculo ou uma calculadora gráfica para responder às seguintes questões:

4.2.1. Represente os dados num referencial ortogonal e diga se é razoável a existência de uma relação linear entre estas duas variáveis.

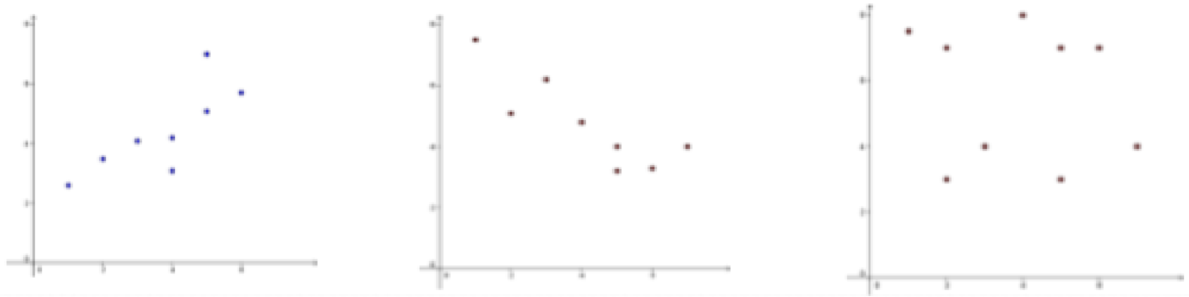
4.2.2. Determine a média dos valores de cada uma das amostras representadas. Apresente os resultados com arredondamento às décimas.

4.2.3. Determine o declive da reta dos mínimos quadrados que se ajusta a esta nuvem de pontos. Apresente o resultado com arredondamento às décimas.

4.2.4. Determine a equação reduzida da reta dos mínimos quadrados.

4.2.5. Utilizando a equação obtida em 4.2.4. determine qual o consumo esperado para um mês em que a temperatura média seja de 7°C.

5. Nos gráficos estão representadas três nuvens de pontos. Faça corresponder a cada gráfico um dos coeficientes de correlação indicados $r_1 = -0{,}25$, $r_2 = 0{,}76$ e $r_3 = -0{,}84$ e justifique.



Exercício do caderno de apoio às metas 11.º ano

Exercícios que podem resolver nos vossos manuais

- Determinação da média, variância e desvio padrão
- Análise da associação linear de duas variáveis estatísticas
- Determinação, através da calculadora gráfica, da equação da reta de regressão

As imagens utilizadas e não identificadas são do manual Máximo 10- Porto Editora