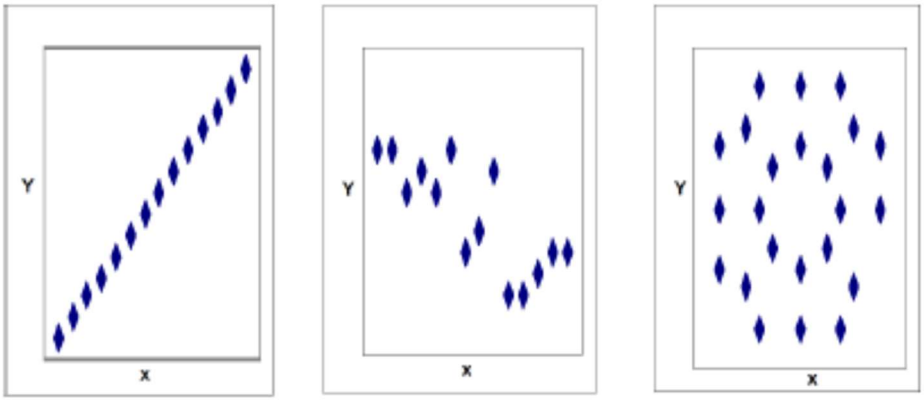
Exercício 2.1 Indique, justificando, qual dos valores abaixo indicados se aproxima mais do coeficiente de correlação dos dados descritos nas seguintes nuvens de pontos,



1. $r_{xy} = 0$.
2. $r_{xy} = 1$.
3. $r_{xy} = -0.5$.
4. $r_{xy} = 2$.

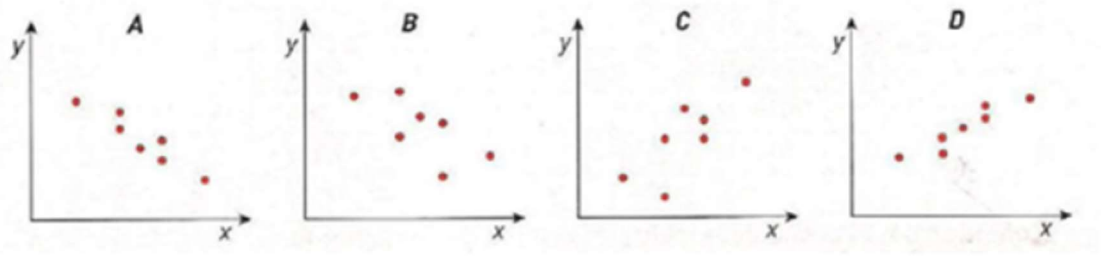
1. $r_{xy} = 0$ corresponde ao 3º gráfico, pois a nuvem de pontos não apresenta qualquer semelhança com uma reta, estando os pontos muitos dispersos. Neste caso, diz-se que existe ausência de relação linear.

2. $r_{xy} = 1$ corresponde ao 1º gráfico, pois apresenta uma relação linear do tipo perfeita (positiva).

3. $r_{xy} = -0.5$ corresponde ao 2º gráfico, pois embora não apresente uma relação do tipo linear (forte), apresenta alguma tendência linear, sendo um possível valor do coeficiente $r_{xy} = -0.5$.

4. $r_{xy} = 2$ não corresponderá a nenhum gráfico pois o coeficiente r_{xy} só toma valores entre -1 e 1, sendo este valor impossível para r_{xy}.

Exercício 2.2 A cada uma das nuvens de pontos A, B, C e D representadas a seguir, faça corresponder um e um só dos seguintes coeficientes de correlação.



1. $r_{xy} = -0.70$.
2. $r_{xy} = 0.96$.
3. $r_{xy} = -0.94$.
4. $r_{xy} = 0.75$.

1. Corresponde à nuvem de pontos B.
2. Corresponde à nuvem de pontos D.
3. Corresponde à nuvem de pontos A.
4. Corresponde à nuvem de pontos C.

Exercício 2.3 Foi medida a várias profundidades (entre 1 e 5 metros), a percentagem de dióxido de urânio numa dada zona geológica, tendo-se obtido os seguintes dados:

$$\sum_{i=1}^{8} x_i = 23 \quad \sum_{i=1}^{8} x_i^2 = 81.5 \quad \sum_{i=1}^{8} x_i y_i = 139.8 \quad \sum_{i=1}^{8} y_i = 39.1 \quad \sum_{i=1}^{8} y_i^2 = 247.05$$

onde X representa a profundidade e Y representa a percentagem de dióxido de urânio. Verifique se a profundidade explica a percentagem de dióxido de urânio e determine, comentando o resultado, a percentagem de dióxido de urânio, previsível a 2 metros de profundidade.

Sendo

X – profundidade e Y – percentagem de dióxido de urânio,

X explica Y se existir um modelo do tipo linear em que se considere X – variável independente e Y – variável dependente.

Para tal, é necessário que exista entre X e Y uma relação do tipo linear.

Para averiguar a existência deste tipo de relação (relação linear) teríamos de construir o diagrama de dispersão e simultaneamente calcular o coeficiente de correlação linear.

Atendendo que não dispomos dos dados, mas apenas dos somatórios, não é possível construir o diagrama de dispersão para averiguar essa tendência linear. Pelo que, apenas poderemos calcular o coeficiente de correlação linear, para averiguar essa tendência linear.

Formas de cálculo do r_{xy}

$$r_{xy} = \frac{\text{covariância}_{xy}}{\sqrt{\text{variância}_x \times \text{variância}_y}} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{s_{xy}}{s_x s_y} =$$

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right) \times \left(\sum_{i=1}^{n} (y_i - \overline{y})^2\right)}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}$$

- $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$ (variância amostral da variável X);
- $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)$ (variância amostral da variável Y);
- $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\right)$ (covariância amostral entre X e Y).

O cálculo do coeficiente de correlação será feito da seguinte forma, mas existem outras possibilidades de cálculo:

$$r_{xy} = \frac{\sum_{i=1}^{8} x_i y_i - 8 \times \bar{x} \times \bar{y}}{\sqrt{\left(\sum_{i=1}^{8} x_i^2 - 8 \times \bar{x}^2\right)\left(\sum_{i=1}^{8} y_i^2 - 8 \times \bar{y}^2\right)}} = *$$

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{23}{8} = 2.875$$

$$\bar{y} = \frac{\sum_{i=1}^{8} y_i}{8} = \frac{39.1}{8} = 4.8875$$

$$* = \frac{139.8 - 8 \times 2.875 \times 4.8875}{\sqrt{(81.5 - 8 \times 2.875^2)(247.05 - 8 \times 4.8875^2)}} = \frac{27.3875}{\sqrt{15.375 \times 55.94875}} = 0.9338$$

Atendendo ao valor de r_{xy} estar próximo de 1, podemos dizer (embora falte o diagrama de dispersão para o comprovar) que existe uma **correlação linear positiva forte** entre X e Y, pelo que X explica Y.

Uma vez mais refiro que confirmação da existência de uma correlação linear positiva forte entre X e Y deveria ser sempre acompanhada pelo diagrama de dispersão, mas neste caso não dispomos dos dados.

Uma vez que pretendemos fazer uma previsão para 2 metros de profundidade e atendendo que X explica Y, ou seja, uma vez que este modelo matemático adequa-se ao conjunto de dados, vamos agora determinar a reta de regressão linear simples ($\hat{y} = a + bx$).

Determinar a reta de regressão: $\widehat{y} = a + bx$

e assim obtém-se a ordenada na origem (a) e o declive da reta (b):

$$\begin{cases} a = \overline{y} - b\overline{x} \\ b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = r_{xy} \times \frac{s_y}{s_x} \end{cases}$$

Assim, determinemos a e b.

$$b = \frac{\sum_{i=1}^{8} x_i y_i - 8 \times \bar{x} \times \bar{y}}{\sum_{i=1}^{8} x_i^2 - 8 \times \bar{x}^2} = \frac{139.8 - 8 \times 2.875 \times 4.8875}{81.5 - 8 \times 2.875^2} = \frac{27.3875}{15.375} = 1.7813$$

$$a = \bar{y} - b\bar{x} = 4.8875 - 1.7813 \times 2.875 = -0.2337$$

Logo o modelo linear que se adequa a este conjunto de dados é

$$\hat{y} = -0.2337 + 1.7813x$$

Assim a previsão de percentagem de dióxido de urânio para uma profundidade de 2 metros é

$$\hat{y}(2) = -0.2337 + 1.7813 \times 2 = 3.3289$$

Esta previsão é de boa qualidade atendendo que o modelo linear se adequa muito bem aos dados pois $r_{xy} = 0.9338$ (muito próximo de 1), apesar de não o termos comprovado através do diagrama de dispersão, por não dispormos dos dados. Além disso, o conjunto de dados que deu origem a este modelo tem os valores de x_i entre 1 e 5 metros (como refere no enunciado), pelo que a previsão para 2 metros é adequada ao conjunto de dados que deu origem a este modelo.

Assim sendo, prevê-se que a percentagem de dióxido de urânio seja de 3.3289 % para 2 metros de profundidade.

Exercício 2.9 O quadro seguinte é o resultado de observações feitas num túnel rodoviário durante períodos de 5 minutos, para o estudo da fluidez do tráfego,

Densidade (veículos / km)	43	55	40	52	39	33	50	33	44	21
Velocidade (km / hora)	27	23	31	24	35	41	27	40	32	51

1. Calcule a variância de cada um dos conjuntos de dados observados. Qual dos conjuntos de dados apresenta maior dispersão? Justifique.
2. Calcule o coeficiente de correlação linear entre as duas variáveis. Que conclusões pode retirar?
3. Determine a equação da reta de regressão, caso se justifique.

1. Seja X - densidade (veículos/km) e Y- velocidade (km/hora).
   Sendo
   - $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\right)$ (variância amostral da variável X);
   - $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2 - n\overline{y}^2\right)$ (variância amostral da variável Y);

   Temos

   Variância de X:

$$s_x^2 = \frac{1}{10-1}\left(\sum_{i=1}^{10}{x_i}^2 - 10\times\bar{x}^2\right) = \frac{1}{9}(17754 - 10\times 41^2) = \frac{944}{9} = 104.8889$$

$$\bar{x} = \frac{\sum_{i=1}^{10}x_i}{10} = \frac{410}{10} = 41$$

   Variância de Y:

$$s_y^2 = \frac{1}{10-1}\left(\sum_{i=1}^{10}{y_i}^2 - 10\times\bar{y}^2\right) = \frac{1}{9}(11655 - 10\times 33.1^2) = \frac{698.9}{9} = 77.6556$$

$$\bar{y} = \frac{\sum_{i=1}^{10}y_i}{19} = \frac{331}{10} = 33.1$$

   Para averiguarmos qual o conjunto de dados com maior dispersão teremos de ir calcular o coeficiente de variação.

$$CV_X = \frac{s_X}{\bar{x}}\times 100\% \quad e \quad CV_Y = \frac{s_Y}{\bar{y}}\times 100\%$$

   Vejamos os coeficientes de variação:

$$CV_X = \frac{s_X}{\bar{x}} \times 100\% = \frac{\sqrt{104.8889}}{41} \times 100\% = 24.98\%$$

$$CV_Y = \frac{s_Y}{\bar{y}} \times 100\% = \frac{\sqrt{77.6556}}{33.1} \times 100\% = 26.62\%$$

Atendendo aos valores dos coeficientes de variação, os dados relativos a Y apresentam maior dispersão.

2. .

Formas de cálculo do r_{xy}

$$r_{xy} = \frac{\text{covariância}_{xy}}{\sqrt{\text{variância}_x \times \text{variância}_y}} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{s_{xy}}{s_x s_y} =$$

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right) \times \left(\sum_{i=1}^{n} (y_i - \overline{y})^2\right)}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}$$

- $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$ (variância amostral da variável X);
- $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right)$ (variância amostral da variável Y);
- $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\right)$ (covariância amostral entre X e Y).

O cálculo do coeficiente de correlação será feito da seguinte forma, mas existiria ainda outras possibilidades:

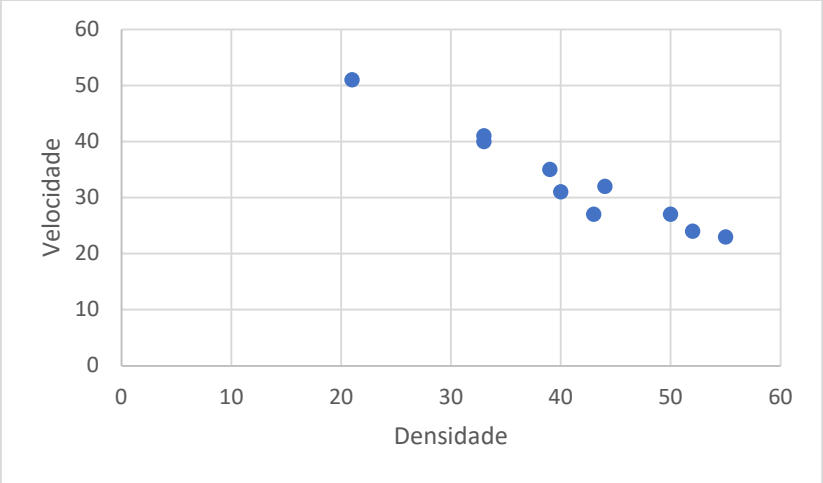
$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \times s_y^2}} = *$$

$$s_{xy} = \frac{1}{10-1}\left(\sum_{i=1}^{10} x_i y_i - 10 \times \bar{x} \times \bar{y}\right) = \frac{1}{9}(12781 - 10 \times 41 \times 33.1) = -\frac{790}{9} = -87.7778$$

$$* = \frac{-87.7778}{\sqrt{104.8889 \times 77.6556}} = -0.9726$$

<u>Para averiguar a existência deste tipo de relação (relação linear) teríamos de construir o diagrama de dispersão e simultaneamente calcular o coeficiente de correlação linear.</u>

Diagrama de dispersão.



Por observação do diagrama de dispersão podemos dizer que existe uma tendência linear entre X e Y e que foi possível comprovar pelo cálculo do r_{xy}, que deu próximo de -1, e portanto podemos dizer que existe uma **correlação linear negativa forte** entre X e Y.

3. Justifica-se determinar a reta de regressão uma vez que o modelo linear que se adequa ao conjunto de dados (visto na alínea anterior).

Determinar a reta de regressão: $\widehat{y} = a + bx$

e assim obtém-se a ordenada na origem (a) e o declive da reta (b):

$$\begin{cases} a = \overline{y} - b\overline{x} \\ b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = r_{xy} \times \frac{s_y}{s_x} \end{cases}$$

Assim, determinemos a e b.

$$b = \frac{s_{xy}}{s_x^2} = \frac{-87.7778}{104.8889} = -0.8369$$

$$a = \bar{y} - b\bar{x} = 33.1 - (-0.8369) \times 41 = 67.4129$$

Logo o modelo linear que se adequa a este conjunto de dados é

$$\hat{y} = 67.4129 - 0.8369x$$