Distribuições Bidimensionais

Existem estudos estatísticos que incidem sobre dois caracteres da mesma população, análises bivariadas, para tentar estabelecer se existe algum tipo de relação entre eles, ou se pelo contrário não existe nenhuma relação.

Exemplos:

1. O número de trabalhadores a realizar uma tarefa e o tempo de execução da mesma.
2. A idade (em anos) e a estatura (em cm).

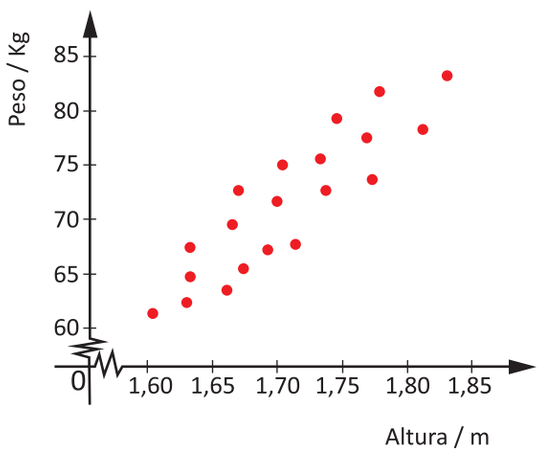
Para estudar relações entre duas caraterísticas conjuntas, de uma mesma população, é necessário recolher uma amostra de dados bivariados que pode ser representada: $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$.

> Uma distribuição em que os dados são bivariados chama-se **distribuição bidimensional**.

Quando existe uma relação próxima à linear entre as duas variáveis diz-se que existe correlação. Significa que a variabilidade de uma é acompanhada, de forma tendencial, pela variabilidade da outra.

Exemplo:

Perguntou-se a cada aluno de uma turma a sua "altura" *x*, em centímetros, e o seu "peso" *y*, em quilogramas. Para cada aluno temos um par ordenado (*x*,*y*).



O conjunto de pontos obtidos no exemplo anterior dá-se o nome de **nuvem de pontos ou diagrama de dispersão**.

Tarefa 70

Num hospital registou-se o peso de seis crianças, em quilogramas:

Criança 1	14
Criança 2	12
Criança 3	17
Criança 4	A
Criança 5	21
Criança 6	B

a) Indica os valores em falta representados por A e B sabendo que a média é 17 e a variância é 26.

b) Supõe que houve um engano no registo das massas e que afinal o peso da primeira criança não é 14 mas sim 20. Das estatísticas amostrais que estudamos até agora, quais são as que serão afetadas por essa alteração?

Tarefa 71

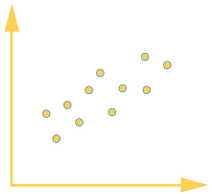
Considera a seguinte distribuição das idades dos elementos dos casais que se encontram numa festa.

Homem	Mulher
21	19
24	21
24	23
26	24
28	24
25	25
32	27
38	35
29	26
28	29

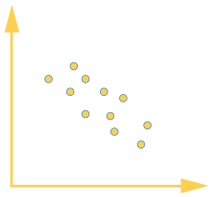
Constrói o diagrama de dispersão dos dados fornecidos.

Dizemos que:

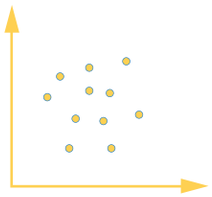
Existe **correlação linear positiva** entre duas variáveis se a nuvem de pontos se ajustar a uma reta com declive positivo, ou seja as variáveis evoluem, no mesmo sentido. (Se uma cresce a outra também cresce, se uma decresce a outra decresce.)



Existe **correlação linear negativa** entre duas variáveis se a nuvem de pontos se ajustar a uma reta com declive negativo, ou seja as variáveis evoluem em sentido contrário. (Se uma cresce a outra decresce e vice-versa.)



Existe **correlação nula** se não há qualquer influência de uma variável na outra.



Numa distribuição bidimensional ao ponto $(\overline{x}, \overline{y})$ **chama-se ponto medio da nuvem de pontos ou centro de gravidade da distribuição**.

Reta de Regressão

Quando realizamos estudos de distribuições bidimensionais, sempre que existe correlação entre as variáveis, queremos prever o valor de uma das variáveis conhecido o valor correspondente da outra.

Pretendemos traçar a reta que melhor se "ajuste" à nuvem de pontos.

A reta que passa no centro de gravidade da distribuição e melhor se ajusta à nuvem de pontos chama-se **reta de regressão**.

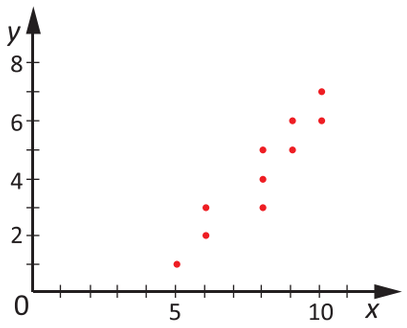
Exemplo:

Numa operação da polícia de transito, que realizaram 10 polícias na cidade de Díli, com o objetivo de verificar se os respetivos condutores tinham carta de condução válida registaram-se as seguinte infrações:

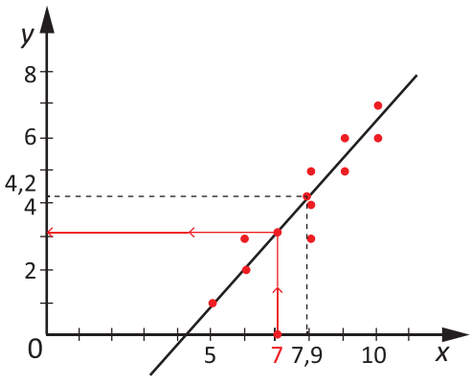
Nº viaturas	5	6	6	8	8	9	9	10	10
Nº multas	1	2	3	3	5	5	6	6	7

Qual é o número de infratores que se prevê sejam multados por um agente que verificou sete viaturas?

Primeiro determinamos o centro de gravidade $\left(\bar{x},\bar{y}\right)=\left(7,9\,;4,2\right)$.



Passando por o centro de gravidade traçamos uma reta que melhor se ajuste à nuvem de pontos:

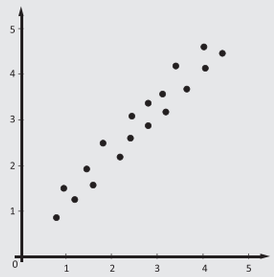


Podemos determinar a equação da reta usando dois pontos o centro de gravidade e um dos pontos da nuvem que pertence a reta.

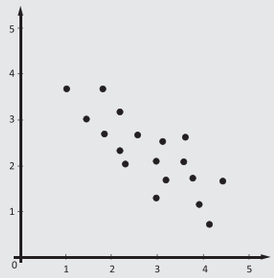
Tarefa 72

Observa os seguintes diagramas de dispersão.

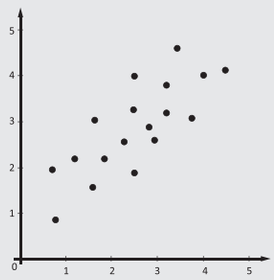
A



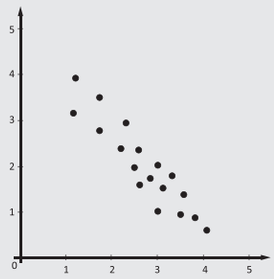
B



C



D



Indica, em cada caso, o tipo de correlação existente.

Tarefa 73

Suspenderam-se objetos de diferentes massas numa mola e registaram-se os correspondentes alongamentos da mola. Na tabela seguinte registaram-se os resultados da experiência.

Massa (em g) x_i	Alongamento (em mm) y_i
10	3
25	7
30	10
45	13
55	15
60	20
70	22
75	23
85	25
100	29

a) Determina o ponto de coordenadas $(\overline{x}, \overline{y})$.

b) Representa a nuvem de pontos e representa graficamente a reta de regressão linear fazendo-a passar pelo ponto $(\overline{x}, \overline{y})$.

Por exemplo $(5,1)$ $(7{,}9\,;4{,}2)$ definimos o vetor diretor da reta $\vec{u} = (2{,}9\,,3{,}2)$. Seguidamente determina-se o declive da reta $m = \frac{3{,}2}{2{,}9} = 1{,}10$ (2 c.d.).

Falta só determinar a ordenada na origem $y = mx + b$.

Como já temos o declive com um dos pontos, podemos encontrar o valor de b:

$$y = 1{,}10x - 4{,}5$$

Com ajuda da reta podemos determinar que o ponto de abcissa 7 tem como ordenada um valor muito próximo de 3. Assim prevemos que sejam aplicadas três multas.

Coeficiente de Correlação

Analisando uma nuvem de pontos ou diagrama de dispersão é possível, intuitivamente, verificar se há correlação e quantificar essa correlação.

Uma das medidas estatísticas que permite estabelecer o grau de correlação existente entre as variáveis relacionadas é o **coeficiente de correlação que se representa por r e toma valores entre -1 e 1**.

Coeficiente de correlação:

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

Se $r < 0$

A correlação é negativa. A variação das variáveis é feita em sentidos opostos, isto é uma aumenta quando a outra diminui.

Se $r > 0$

A correlação é positiva. A variação das variáveis é feita no mesmo sentido, isto é, uma aumenta quando a outra também aumenta.

Se $r = 0$

A correlação é nula.

Quando $r = 1 \quad ou \quad r = -1$

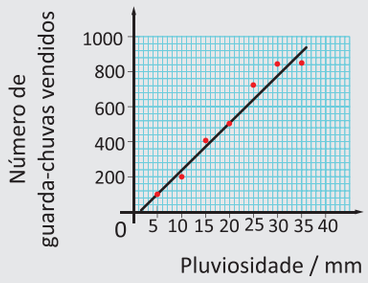
Os pontos no diagrama de dispersão estão sobre uma reta. Essa reta tem declive positivo $r = 1$ e tem declive negativo se $r = -1$.

Na correlação positiva, quanto mais próximo de 1 for o valor, mais forte é a correlação.

Na correlação negativa, quanto mais próximo de -1 for, valor mais forte é a correlação.

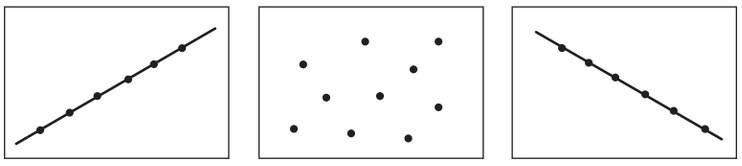
Tarefa 74

Considera o seguinte diagrama de dispersão e a reta de regressão representada no seguinte gráfico.

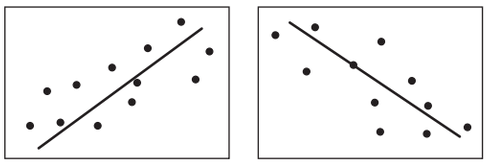


Obtém a equação da reta de regressão.

Exemplos:



$r = 0.6$ $r = 0.0$ $r = -1.0$



$r \simeq 0.6$ $r \simeq -0.6$

Distribuições de Probabilidade

Nas diferentes experiencias aleatórias que estudamos, consideramos, para cada caso, o espaço de resultados e a cada resultado fizemos corresponder a sua ocorrência.

Esta função associada à experiência aleatória diz-se variável aleatória X.

> Dada uma experiencia aleatória, chama-se variável aleatória associada a essa experiência, a toda função X definida no espaço de resultados Ω, com valores em IR, que permite expressar os resultados por números reais.

Considera a experiência que consiste em lançar uma moeda ao ar e registar-se a face que fica voltada para cima.

A variável aleatória associada seria:

X: "Número de vezes que a face verso fica voltada para cima"

Estabelecemos que a ocorrência da face verso corresponde o valor 1 e a não ocorrência corresponde ao valor 0. Temos uma variável aleatória X que a cada elemento do espaço amostral faz corresponder os valores 1 e 0.

Tarefa 75

Faz corresponder a cada uma das distribuições os respetivos coeficientes de correlação:

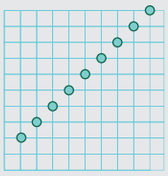
(A) -0,46

(B) 0,72

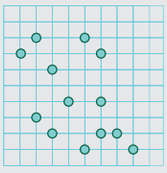
(C) -0,94

(D) 1

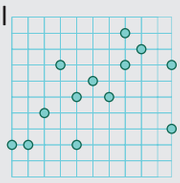
I



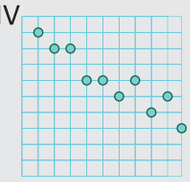
II



III



IV



 (V)  (A) Temos então X(V) =1 X(A) = 0

Quando o conjunto dos valores assumidos por uma variável aleatória X (isto é, o seu contradomínio) for finito ou, sendo infinito, puder ser contado ou enumerado (infinito numerável), X diz-se uma variável **aleatória discreta**.

Caso o conjunto dos valores assumidos pela variável seja um intervalo de números reais (ou união de intervalos), a **variável aleatória** diz-se **contínua**.

Exemplo:

Considere-se a experiência aleatória que consiste em lançar duas moedas perfeitas ao ar, com o objetivo de contar o número de versos que aparecem.

A variável aleatória X:"número de versos observado, num lançamento"

estabelece uma correspondência entre o espaço de resultados

$$\Omega = \{(A,A),(A,V),(V,V),(V,A)\},$$

onde a letra A designa a face "anverso" e V a face "verso", e o conjunto dos números reais {0,1, 2} é o contradomínio da função.

Temos assim,

$$X :\{(A,A),(A,V),(V,V),(V,A)\} \rightarrow \{0,1, 2\}.$$

A variável aleatória X é uma variável discreta.

Dada uma experiência aleatória, sendo X a variável aleatória associada a essa experiência, se a cada valor da variável aleatória X associarmos a respetiva probabilidade obtemos a **distribuição de probabilidade dessa variável**.

Relativamente a variável aleatória X definida no exemplo anterior, podem calcular-se as seguintes probabilidades,

$$P(X=0)=\frac{1}{4} \qquad P(X=1)=\frac{1}{2} \qquad P(X=2)=\frac{1}{4}$$

Então a distribuição de probabilidade será:

x_i	0	1	2
$P(X=x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Dada uma variável aleatória discreta X que assume um número finito de valores distintos $x_1, x_2, \cdots x_n$ então as probabilidades $p_i = P(X = x_i)$.

Devem satisfazer as condições:

- $0 \leq p_i \leq 1$
- $\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots p_n = 1$

Existem determinadas caraterísticas numéricas associadas à distribuição de uma variável aleatória. A estas caraterísticas dá-se o nome de **parâmetros de distribuição**.

Entre os vários parâmetros existentes destacamos os parâmetros de localização e os parâmetros de dispersão.

Os ***parâmetros de localização*** são indicadores de como determinada caraterística de uma população se distribui ao longo do respetivo espaço de resultados. Um exemplo de um parâmetro de localização é o ***valor médio***.

Os ***parâmetros de dispersão*** são medidas da variabilidade de determinadas caraterísticas de uma população. Um exemplo de um parâmetro de dispersão é a ***variância***.

Valor médio de uma variável aleatória

O valor médio de uma variável aleatória X que toma os valores $x_1, x_2, \cdots x_n$ com probabilidades $p_1, p_2, \cdots p_n$ é o número

$$\mu_X = x_1 \times p_1 + x_2 \times p_2 + \cdots + x_k \times p_n = \sum_{i=1}^{n} x_i \times p_i$$

.

Sendo x_i um qualquer resultado entre os possíveis e p_i a probabilidade associada a esse resultado.

O **valor médio** também se chama **valor esperado** ou **esperança matemática**.

Desvio-padrão de uma variável aleatória

O desvio-padrão, σ, de uma variável aleatória X que toma os valores $x_1, x_2, \cdots x_n$ com probabilidades $p_1, p_2, \cdots p_n$ é o número real não negativo.

$$\sigma = \sqrt{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

Tarefa 76

Pensa-se que os casais têm tendência para terem alturas semelhantes. Considere os seguintes pares, que são referentes às alturas (em cm) de 10 casais.

Mulher	Homem
170	183
164	168
167	178
165	173
164	167
165	164
166	165
165	170
162	165
163	164

a) Representa os pontos num diagrama de dispersão.

b) Calcula o valor do coeficiente de correlação.

Nota

A letra grega μ:

lê-se "mi" minúsculo.