INSTITUTO SUPERIOR DE AGRONOMIA

INTRODUÇÃO
à
ESTATÍSTICA e à PROBABILIDADE

Ano Lectivo - 2014/2015

Manuela Neves

- 2014 -

Nota Introdutória

Estas folhas destinam-se a apoiar as aulas da Unidade Curricular *Estatística*, adaptada a Bolonha, leccionada no 2ºano de quase todas as licenciaturas do Instituto Superior de Agronomia.

A matéria teórica aqui exposta constitui um complemento ao material de apoio que os alunos utilizam nas aulas teóricas e práticas. Podem ser aqui encontradas as deduções e demonstrações que, por razões de tempo, não é possível apresentar nas aulas.

Algumas partes que poderão parecer mais densas não puderam ser evitadas, pois pretende-se que o tratamento matemático seja feito com o rigor necessário.

Tratando-se de apontamentos não podem nem devem substituir a leitura de obras indicadas nas **Referências Bibliográficas**.

Manuela Neves

Setembro de 2014

INTRODUÇÃO

Historicamente, o termo "estatística" deriva da palavra latina "status" que significa "estado". De facto, a estatística surgiu na segunda metade do século XVII, segundo parece como um auxiliar da arte de governação, associada a problemas de economia, demografia, problemas políticos, etc.. Actualmente ocupa um papel cada vez mais importante nas mais variadas disciplinas: química, biologia, economia, medicina, psicologia, meteorologia, agricultura, ciências sociais e políticas e em muitos outros campos da ciência e engenharia.

A estatística dedica-se fundamentalmente ao estudo da teoria e à aplicação de métodos de coleccionar, analisar dados e ainda obter conclusões e tomar decisões válidas, a partir desses dados. É em situações de incerteza na predição de resultados e na obtenção de conclusões, que a estatística está presente.

Podemos então dizer que **a Estatística** é um conjunto de conceitos e métodos utilizados na recolha e interpretação de dados respeitantes a uma determinada área de investigação, permitindo ainda descrever e predizer situações em que a variabilidade e a incerteza estão presentes.

A Estatística pode ser dividida em dois grandes grupos:

- **A estatística descritiva**, cujo objectivo é sumariar e descrever os aspectos relevantes num conjunto de dados;
- **A inferência estatística**, que se preocupa em tirar conclusões a partir de um conjunto de observações (amostra) pela interpretação dos resultados obtidos pela estatística descritiva. Ela permite fundamentalmente tomar decisões quanto ao(s) valor(es) de características importantes da população ou populações, de que foi retirada a amostra.

  Para isto é necessário o recurso à **teoria da probabilidade** na qual a inferência estatística se baseia fortemente.

Estes apontamentos são uma **Introdução** à

- **Estatística Descritiva**
- **Teoria da Probabilidade** e
- **Inferência Estatística**

ESTATISTICA DESCRITIVA

A **estatística descritiva** tem como objectivo apresentar os dados observados sob a forma de tabelas e gráficos, que tornem mais fácil uma primeira análise desses dados e ainda a obtenção de valores numéricos que os caracterizem globalmente.

Dois conceitos básicos em estatística são o conceito de **população** ou **universo** e **amostra**.

População é o conjunto de elementos com alguma característica em comum a qual se pretende estudar. Esses elementos podem ser pessoas, animais, plantas, explorações agrícolas, resultados experimentais, etc. Aos elementos da população chamamos **unidades estatísticas**. À característica em comum, que toma valores diferentes de elemento para elemento, chamamos **variável**.

Uma população pode ser finita ou infinita. Uma população finita pode ter um número muito elevado de elementos, por exemplo, a população de todos os parafusos produzidos por uma fábrica num dado dia é finita, embora de dimensão muito elevada, enquanto a população de todos os locais do território português (para estudo da altitude, por exemplo) é infinita. Nos casos anteriores a observação de todos os elementos da população ou é muito difícil ou é mesmo impossível.

Sendo assim o estudo é feito sobre alguns elementos (unidades estatísticas) retirados da população, constituindo aquilo a que se chama uma **amostra** e que são efectivamente observados. Aos valores observados para a variável de interesse chamamos **dados**. Os dados são os objectos de estudo da Estatística e a partir deles pretendemos fazer inferências sobre características numéricas da população a que se chama **parâmetros**.

Atenda-se a que os **dados** podem ser de natureza **qualitativa** - representam a informação que identifica uma qualidade ou categoria, que não é possível ser medida. Por exemplo, dados referentes às cores das faces de um dado, cor dos olhos, sexo de uma pessoa, naipes de um baralho de cartas, etc., ou de natureza **quantitativa** - referentes a informação suscepível de ser medida. Destes há a considerar o caso de dados de natureza **discreta**, ou contagens, por exemplo o n º de cabeças de gado por exploração, o n º de chamadas telefónicas recebidas durante um certo período de tempo num escritório, o nº de árvores por herdade, etc., e dados de natureza **contínua** ou medições, como, por exemplo, peso e altura dos portugueses num certo intervalo de idades, altura de uma árvore, extensão de uma propriedade agrícola, etc.

O estudo de observações referentes apenas a uma característica é objectivo da estatística descritiva a uma dimensão e, da descrição e do estudo de observações de duas variáveis trata a estatística descritiva a duas dimensões, com a análise das possíveis relações existentes entre essas variáveis. A generalização ao caso de várias variáveis é do domínio da estatística descritiva multidimensional.

ESTATISTICA DESCRITIVA A UMA DIMENSÃO

Objectivos

A estatística descritiva a uma dimensão tem como objectivo sumariar e descrever os aspectos mais importantes de um conjunto de dados resultantes da observação de uma só variável de interesse na população. Utiliza métodos adequados para:

- condensar os dados em tabelas;
- representá-los graficamente;
- calcular características amostrais de localização e variabilidade.

Os aspectos importantes para descrever um conjunto de dados são:

- a apresentação de gráficos e tabelas;
- o exame da forma geral do gráfico para tentar descobrir aspectos particulares, como por exemplo simetria e achatamento;
- o exame do gráfico para tentar descobrir observações atípicas, *outliers*;
- o cálculo de medidas numéricas para
  – um valor representativo da localização dos dados,
  – um valor representativo da dispersão dos dados,
  – um valor representativo da forma de distribuição dos dados.

Descrição dos dados por gráficos e tabelas

De entre os métodos gráficos usados para representar um conjunto de dados, dois dos principais são **o diagrama de barras** e **o histograma**.

O diagrama de barras

Suponhamos que temos o seguinte conjunto de dados relativos às classificações obtidas por 20 alunos numa dada disciplina:

12	13	15	17	4	8	10	11	9	10
8	7	12	10	11	11	14	7	9	13

Verificando-se que as classificações obtidas pelos alunos se situam entre 4 e 17 podemos organizar a seguinte **tabela de frequências**:

Notas obtidas	Frequências absolutas n_i	Frequências relativas f_i
4	1	0.05
7	2	0.10
8	2	0.10
9	2	0.10
10	3	0.15
11	3	0.15
12	2	0.10
13	2	0.10
14	1	0.05
15	1	0.05
17	1	0.05

O exame desta tabela mostra-nos que as classificações mais frequentes são 10 e 11. Verifica-se ainda que há uma percentagem maior de notas positivas do que negativas, sendo ainda as notas mais raras 4, 14, 15 e 17. Uma **tabela de frequências** permite portanto uma análise rápida e sumária dos dados.

Designemos por n o número de observações recolhidas, i.e. a **dimensão da amostra**. A **frequência absoluta**, que habitualmente se representa por n_i, é o número de vezes que o elemento i é observado e a **frequência relativa** da observação i, que designaremos por f_i, é definida como:

$$\text{frequência relativa}=\frac{\text{frequência absoluta}}{\text{dimensão da amostra}} \Leftrightarrow f_i = \frac{n_i}{n}.$$

O procedimento gráfico usado no caso de dados de **natureza discreta**, quando o número de valores distintos é pequeno, é o **diagrama de barras**. Consiste em desenhar um sistema de eixos coordenados, marcar no eixo dos xx os diferentes valores observados e sobre cada um desenhar uma barra vertical de altura igual à frequência absoluta ou à frequência relativa, ver Figura 1.

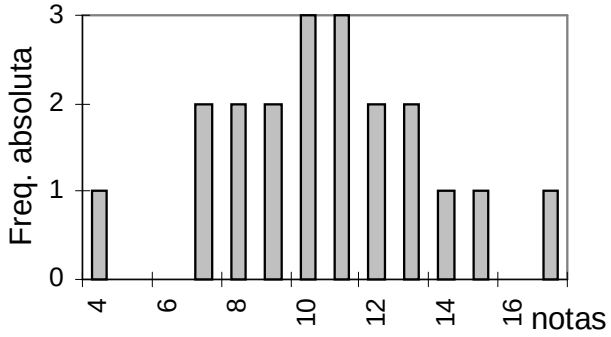


Figura 1: Diagrama de barras das frequências absolutas.

O histograma

Quando o número de observações distintas é elevado ou os dados são de **natureza contínua** (recorde-se que são dados de natureza contínua os que se referem a pesos, alturas, tempos, durações, velocidades, temperaturas, etc, enquanto dados referentes a contagens, são de natureza discreta), dever-se-á fazer a condensação dos dados, agrupando as observações próximas por forma a evidenciar as características subjacentes aos dados. Esta sumarização inicia-se construindo uma **tabela de frequências**, cuja representação gráfica é feita agora por meio de **histogramas** e **polígonos de frequências**.

Dada uma série de n observações, vejamos os passos que é necessário seguir para elaborar uma **tabela de frequências** nestas circunstâncias:

- Determinar o máximo e o mínimo valor do conjunto das observações, $\max(x_i)$ e $\min(x_i)$, respectivamente.

  A $\max(x_i) - \min(x_i)$ chama-se **amplitude total**.

- Escolher um número de subintervalos (regra geral, com a mesma amplitude), cuja reunião (sem sobreposição) cubra a amplitude total. A estes intervalos é costume chamar **classes** e os seus extremos **limites de classes**. Iremos considerar as classes abertas à esquerda e fechadas à direita [1], i.e., intervalos da forma]].

- Para cada classe i calcula-se a **frequência absoluta**, n_i que designa o número de observações que pertencem a essa classe, e a **frequência relativa**, que designaremos por f_i:

$$\text{frequência relativa da classe } i = \frac{\text{frequência absoluta da classe } i}{\text{número total de observações da classe } i}.$$

A escolha do número e posição das classes é um problema de experiência, sendo, regra geral de 5 a 15 o número de classes que se deve considerar.

Na prática, existem regras empíricas para fazer esta escolha, sendo a mais usada a **regra de Sturges**: – toma-se como número de classes o inteiro

$$m \simeq 1 + (\log_2 n) = 1 + \frac{\log_{10} n}{\log_{10} 2}$$

(alguns autores aconselham o maior inteiro inferior ou igual àquela quantidade).

A amplitude h de cada classe, obtém-se agora fazendo o quociente $(\max(x_i) - \min(x_i))/m$.

[1] Alguns autores consideram intervalos da forma [[e outros intervalos [ ], neste caso com escolha conveniente dos limites das classes por forma a não haver sobreposições, regra geral adiciona-se 1/2 aos valores observados.

Para a construção das classes pode iniciar-se o processo considerando a classe $]\overline{x}-h/2, \overline{x}+h/2]$ [2]. A partir desta, formar-se-ão as classes subtraindo h e somando h ao extremo inferior e superior, respectivamente, para ir determinando as classes inferiores e superiores àquela classe. Para que todo o suporte da amostra fique coberto são necessárias $m+1$ classes.

Esta técnica apresenta bons resultados no caso de distribuições simétricas ou aproximadamente simétricas. Caso tal não se verifique, dever-se-á considerar outro modo de elaborar as tabelas.

Um outro procedimento consiste em começar a construção das classes pelo mínimo (ou pelo máximo). A primeira classe deverá ser então escolhida por forma a conter $\min(x_i)$ (ou $\max(x_i)$) e a última a formar-se deverá conter o $\max(x_i)$ (ou $\min(x_i)$).

Construída a tabela de frequências, os dados podem ser agora representados num **histograma**.

Construção do histograma de frequências relativas

Num eixo horizontal marcam-se as classes definidas e, sobre elas, desenham-se rectângulos verticais tendo como base h e altura dada pelo quociente entre a frequência relativa e a amplitude da classe. A área de cada rectângulo é igual à frequência relativa, representando então a proporção das observações que ocorrem na classe correspondente. Como é imediato verificar, a área total do histograma vem então igual a 1.

Unindo por segmentos de recta os pontos médios dos topos dos rectângulos de um histograma obtemos o **polígono de frequências relativas**.

Exemplo 2.

Os dados seguintes referem-se ao peso (em kg) de 57 animais de idade e história genética semelhantes, no final de uma experiência de nutrição animal, durante a qual lhes foi administrada uma mesma dieta em condições controladas.

68	63	42	27	30	36	51	38	25	44	65	43	25	74	49	43
45	12	57	51	12	32	22	79	21	16	24	69	47	23	32	42
46	30	43	49	12	28	36	42	38	19	28	50	23	24	25	27
27	28	27	49	22	31	31	28	23							

Tem-se então $min(x_i) = 12$ $max(x_i) = 79$ $\overline{x} = 36.72$

Dado que o valor obtido pela regra de Sturges é 6.83, iremos considerar $m = 6$, o que daria como um valor a usar para a amplitude das classes $h = 11$.

Construamos então a seguinte **tabela de frequências** (onde F_i designa a frequência relativa acumulada):

[2] $\overline{x}$, média do conjunto das observações, é uma característica numérica de um conjunto de dados cujas propriedades são apresentadas na página 9; é assim definida $\overline{x} = \sum x_i/n$.

Classes		$x_i^{'}$	n_i	f_i	F_i
]10	21]	15.5	6	.105	.105
]21	32]	26.5	24	.421	.526
]32	43]	37.5	10	.175	.702
]43	54]	48.5	10	.175	.877
]54	65]	59.5	3	.053	.930
]65	76]	70.5	3	.053	.982
]76	87]	81.5	1	.018	1
Total			57	1.000	

Na Figura 2. representa-se o histograma de frequências absolutas.

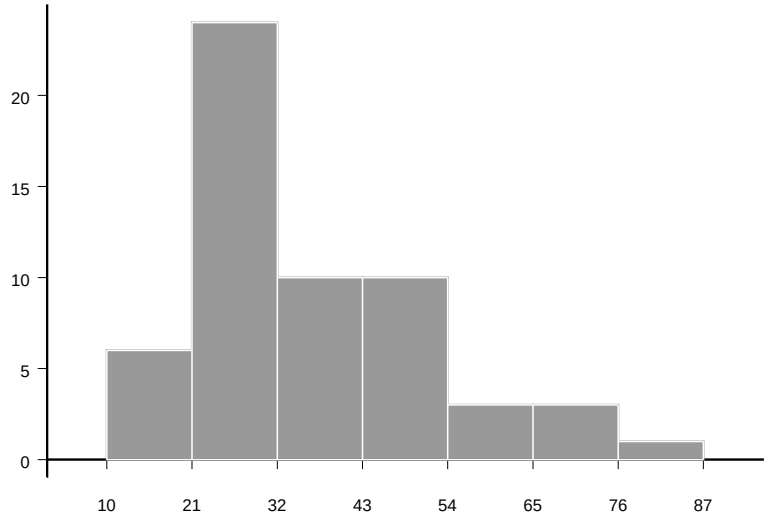


Figura 2: Histograma de frequências absolutas.

Características numéricas de um conjunto de dados

Os métodos gráficos referidos permitem-nos visualizar o modelo subjacente a um conjunto de dados. Para podermos ter uma descrição mais objectiva, necessitamos de medidas quantitativas referentes a

- localização dos dados;
- grau de variação ou dispersão dos dados;
- forma de distribuição dos dados.

Indicadores de localização

Consideremos um conjunto de n observações, $x_1, x_2, ..., x_n$.

Chama-se **medida de localização** a toda a grandeza numérica cujo valor referencie a posição de um conjunto de dados. As medidas de localização mais usadas são a **média**, **a mediana** e ainda **os quartis** e **a moda**.

A **média aritmética**, **média empírica** ou simplesmente **média** é o ponto de "equilíbrio" de um conjunto de dados. Representa-se por $\overline{x}$ e define-se como

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

Chama-se **desvio** de uma observação relativamente à média a $x_i - \overline{x}$.

Exercício 1. Verificar que a soma dos desvios relativamente à média é nula, i.e., $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.

Propriedades da média

1. Dadas as observações $x_1, x_2, ..., x_n$ com média $\overline{x}$, consideremos uma mudança de origem nos dados, i.e., $x_i^{'} = x_i + a,\ i = 1, ..., n$.
Os novos dados têm como média $\overline{x'} = \overline{x} + a$.
Dem: De facto $\overline{x'} = \dfrac{\sum_{i=1}^{n} x_i^{'}}{n} = \dfrac{\sum_{i=1}^{n}(x_i + a)}{n} = \dfrac{\sum_{i=1}^{n}(x_i) + na}{n} = \overline{x} + a$.

2. Efectuando uma mudança de escala nos dados , i.e., $x_i' = b\ x_i\ (b \neq 0)\ i = 1, ..., n$, temos $\overline{x'} = b\ \overline{x}$, de dedução imediata.

Nota: as duas propriedades anteriores podem ser resumidas numa única.

3. Dadas as observações $x_1, x_2, ..., x_n$ com média $\overline{x}$, se $x_i' = a + bx_i,\ i = 1, ..., n$.
tem-se $\overline{x'} = a + b\ \overline{x}$.

4. Sejam $x_1, ..., x_n$ uma série de n observações de média $\overline{x}$ e, $y_1, ..., y_m$ outra série de m observações de média $\overline{y}$. A média do conjunto das $n + m$ observações é dada por

$$\frac{n\ \overline{x} + m\ \overline{y}}{n + m}.$$

Dem: Designando por z_i as $n + m$ observações, tem-se então

$$\overline{z} = \frac{\sum_{i=1}^{m+n} z_i}{n + m} = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i}{n + m} = \frac{n\ \overline{x} + m\ \overline{y}}{n + m}.$$

Outros tipos de médias como, por exemplo, **a média geométrica** e **a média harmónica** não serão consideradas aqui, podendo encontrar-se referências sobre eles na bibliografia indicada.

A média aritmética, apesar de fácil e rápida de calcular, apresenta a desvantagem de ser muito sensível a valores muito pequenos ou muito grandes no conjunto dos dados. Os valores existentes numa amostra que se distinguem muito dos restantes por serem demasiado grandes ou demasiado pequenos, são valores que se apresentam como candidatos a *outliers*. Mais adiante daremos uma regra empírica que permitirá classificar um valor como *outlier*. Uma medida **robusta** relativamente ao valor das observações extremas, no sentido de não ser afectada por esse valor, é a mediana.

A mediana de um conjunto de n observações é o valor do meio, depois de dispostos os dados por ordem crescente de grandeza. Trata-se portanto de uma medida de posição; é costume representar-se por $\tilde{x}$ ou ainda me.

Na escolha do valor do meio há que ter em conta o seguinte:

– se n é **ímpar** há um único valor no meio;

– se n é **par** existem dois valores no meio, sendo a mediana dada pela média aritmética desses dois valores.

Tendo n observações $x_1, ..., x_n$ designe-se por $x_{(1)}, ..., x_{(n)}$, as observações depois de ordenadas, i.e., $x_{(1)} \leq ... \leq x_{(n)}$. A **mediana** é então definida como

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ ímpar} \\ \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & n \text{ par} \end{cases} \tag{2}$$

A interpretação geométrica da mediana para dados agrupados em classes é muito simples: é o valor do eixo das abcissas tal que a recta vertical que passa nesse ponto divide a área do histograma em duas regiões com áreas iguais.

Se a mediana é o valor que divide um conjunto ordenado de dados em duas partes iguais, podemos generalizar este conceito, considerando a amostra ordenada dividida em quatro partes iguais. Aos pontos da divisão chamamos **quartis** e representaremos por $Q_1,\ Q_2,\ Q_3$, o primeiro, segundo e terceiro quartis, respectivamente.

Sendo assim, por exemplo, o primeiro quartil, Q_1, será o valor tal que pelo menos 25% das observações são menores ou iguais a ele e pelo menos 75% das observações são maiores ou iguais.

Repare-se que Q_2 coincide com a mediana.

De forma semelhante se podem definir os **decis**, valores que dividem o conjunto das observações em 10 partes iguais e os **centis** ou **percentis**, como sendo os valores resultantes da divisão da amostra ordenada em 100 partes iguais.

A todas estas medidas, quartis, decis e percentis dá-se a designação genérica de **quantis**.

Dado um número $0 \leq \theta \leq 1$, chama-se **quantil de ordem** θ ao valor do conjunto das observações depois de ordenadas, tal que, pelo menos $\theta \times 100\%$ delas são inferiores ou iguais a esse valor e pelo menos $(1-\theta) \times 100\%$ das observações são maiores ou iguais a esse valor.

Consideraremos a seguinte fórmula de cálculo do quantil de ordem θ, Q^*_θ:

$$Q^*_\theta = \begin{cases} \frac{x_{(n\ \theta)} + x_{(n\ \theta+1)}}{2} & n\ \theta \text{ inteiro} \\ x_{([n\ \theta]+1)} & n\ \theta \text{ não inteiro} \end{cases} \tag{3}$$

onde $[n\ \theta]$ designa o maior inteiro contido em $n\ \theta$. Tem-se, por exemplo, $[3.25] = 3$ e $[8.95] = 8$.

O primeiro e terceiro quartis permitem definir uma regra empírica para identificar um valor atípico como *outlier*.

Assim, chama-se **barreira inferior** que designaremos por $\boldsymbol{B_I}$, a

$$\boldsymbol{B_I = Q_1 - 1.5(Q_3 - Q_1)}$$

e **barreira superior** que designaremos por $\boldsymbol{B_S}$, a

$$\boldsymbol{B_S = Q_3 + 1.5(Q_3 - Q_1)}$$

Um valor observado x_i diz-se que é **um outlier** se

$$x_i < B_I \qquad \text{ou} \qquad x_i > B_S.$$

As características numéricas calculadas após a ordenação dos valores da amostra chamam-se parâmetros de ordem.

Uma outra medida de localização, embora menos usual é a **moda**, *mo*, definida, no caso discreto, como o valor que ocorre com mais frequência, ou como o intervalo de classe com maior frequência se os dados são de natureza contínua.

Um conjunto de observações pode não ter moda ou apresentar mais do que uma moda. Uma distribuição com uma única moda diz-se **unimodal**.

Esta medida é particularmente útil quando temos dados de natureza qualitativa, para os quais não é possível calcular a média ou mesmo a mediana (por não ser possível estabelecer uma ordenação entre eles, para a determinação deste indicador).

Indicadores de dispersão

Uma média ou qualquer outra medida de localização, não são suficientes para dar uma ideia clara da distribuição das observações. De facto, podemos considerar dois conjuntos de dados diferentes mas tendo, por exemplo, a mesma média e mediana. Vejamos:

$$1,\ \ 2,\ \ 5,\ \ 8 \qquad \overline{x} = 4 \qquad \tilde{x} = 3.5$$

$$-2,\ \ 3,\ \ 4,\ \ 11 \qquad \overline{x} = 4 \qquad \tilde{x} = 3.5 \quad .$$

O primeiro conjunto apresenta maior concentração dos dados do que o segundo. É portanto necessária uma medida que nos dê alguma informação sobre a dispersão das observações.

Vejamos então quais os indicadores de dispersão mais usados:

Amplitude Total é a amplitude do intervalo de variação dos dados, assim definida

$$A_{tot} = max(x_i) - min(x_i). \tag{4}$$

É uma medida que se baseia apenas na maior e na menor observação, ignorando a informação presente nas observações intermédias, sendo por isso muito sensível aos extremos.

Uma outra medida análoga, mas mais informativa e menos afectada pelos valores extremos é a

Amplitude inter-quartil definida como

$$AIQ = Q_3 - Q_1. \tag{5}$$

Nas distribuições simétricas o intervalo $(\tilde{x} - ASQ, \tilde{x} + ASQ)$ contém 50% das observações, onde $ASQ = (Q_3 - Q_1)/2$ se designa por **amplitude semi-quartil**.

Mas também ASQ ignora a informação contida na zona central e nas zonas extremas das observações.

Interessa então considerar medidas que tenham em conta a posição de todos os valores observados, relativamente a um ponto de referência. Sendo a média a medida de localização mais usada, regra geral toma-se esta para referenciar a dispersão.

Usar como indicador $\sum(x_i - \overline{x})$ é evidente que não serve, pois como vimos atrás este valor é sempre nulo. Uma medida de dispersão que pareceria então lógica era **o desvio médio**, definido como

$$d = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|. \tag{6}$$

É claro que, quanto menos dispersos estiverem os valores observados relativamente à média, menor será o desvio médio. Apesar de simples de calcular, o desvio médio não é muito usado, porque a existência de módulos torna o seu tratamento matemático pouco acessível.

Uma medida definida com um critério análogo mas baseada na soma dos quadrados dos desvios é a

Variância, que habitualmente se representa por s_x^2 ou mais simplesmente s^2 e se define como

$$\boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{7}$$

Observação: O uso de $(n-1)$ em vez de n como parecia lógico, será justificado mais tarde na Inferência Estatística, sendo no entanto indiferente o uso de um ou outro quando se trate de amostras de grande dimensão.

Uma outra fórmula de cálculo da variância pode ser obtida fazendo o desenvolvimento do quadrado da diferença, resultando então

$$\boldsymbol{s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}}. \tag{8}$$

(fórmula esta que tem interesse prático principalmente quando os valores x_i não são muito grandes, mas o uso da qual requer cuidados especiais para a hipótese de ocorrerem no numerador dois números muito próximos, o que poderá conduzir à perda de dígitos.)

A raiz quadrada da variância fornece-nos uma medida de concepção análoga à do desvio médio, que se representa por s e se designa por **desvio padrão**.

Propriedades da variância

1. A variância é não negativa, i.e., $s^2 \geq 0$, o que é imediato a partir da definição.

2. Sejam $x_1, ..., x_n$, n observações com variância s_x^2 e $y_i = a + bx_i, \qquad i = 1, ..., n.$ Tem-se então como variância das novas observações,

$$s_y^2 = b^2\ s_x^2.$$

Dem:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(a + bx_i - a - b\overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} b^2(x_i - \overline{x})^2}{n-1} = b^2\ s_x^2.$$

Esta propriedade mostra-nos que a variância não é afectada por uma mudança de origem, mas é afectada por uma mudança de escala.

Para o **desvio padrão** tem-se $s_y = |b| s_x$.

Exercício 2. Provar que a soma dos quadrados dos desvios para a média é menor que a soma dos quadrado dos desvios para qualquer outro valor, ou seja

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2 \qquad \forall a \in \mathbb{R}.$$

As medidas de dispersão acabadas de estudar dizem-se **absolutas** porque dependem das unidades adoptadas, i.e., qualquer alteração da unidade provoca uma modificação no valor do indicador calculado. É de todo o interesse a existência de medidas independentes das unidades, permitindo assim o estudo comparativo de dois ou mais conjuntos de dados. São as medidas de dispersão **relativas**.

Uma medida de dispersão relativa , usada apenas quando a variável toma valores de um sinal, i.e., todos positivos ou todos negativos, é **o coeficiente de variação** definido como

$$C.V. = \frac{s}{\overline{x}} \times 100\%. \tag{9}$$

Esta medida é independente das unidades consideradas, permitindo por isso comparar distribuições cujas unidades podem ser diferentes ou que difiram consideravelmente em grandeza. No entanto só pode ser usado quando a variável toma valores só positivos ou só negativos.

Outro processo para comparar conjuntos de dados consiste em trabalhar com as variáveis **estandardizadas** ou **reduzidas**, i.e., são as variáveis da forma

$$z_i = \frac{x_i - \overline{x}}{s_x}.$$

Como facilmente se verifica (recorrendo a propriedades da média e da variância, o que deixamos como exercício), as variáveis reduzidas têm média nula e variância unitária.

Os valores z_i são obviamente quantidades independentes das unidades usadas e, portanto, as distribuições referentes a essas variáveis directamente comparáveis.

Um modo muito fácil de interpretar a localização, dispersão e afastamento da simetria de um conjunto de dados efectuando em simultâneo a sua síntese pode ser feito sob uma forma gráfica muito sugestiva – **o diagrama de extremos e quartis** ou a **caixa-de-bigodes**.

O diagrama de extremos e quartis consiste em marcar num eixo os extremos (máximo e mínimo), a mediana, o 1 º e 3 º quartis. Desenha-se depois um gráfico como o da Figura 3, correspondente aos dados do Exemplo 2.

Este procedimento tem ainda a vantagem de permitir a comparação rápida entre conjuntos de dados, como se pode ver no Exemplo 3, Figura 4.

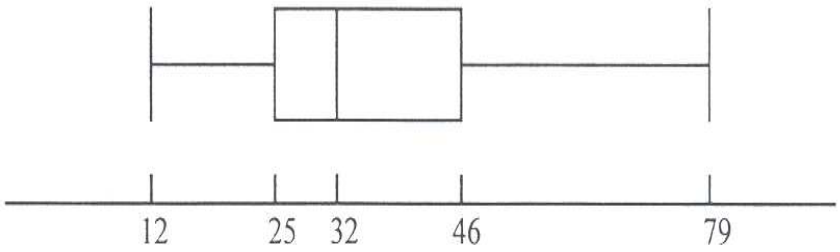


Figura 3: Diagrama de extremos e quartis.

Exemplo 3.(Murteira e Black, 1983)

Baseados em dados das Estatísticas Agrícolas (INE, 1979), têm-se os seguintes valores dos extremos e quartis, das taxas de arborização nos concelhos dos distritos de Aveiro, Beja, Bragança e Faro.

	Aveiro (n=19)	Beja (n=14)	Bragança (n=12)	Faro (n=16)
min	15.9	7.7	3.5	0.7
max	60.6	60.3	28.9	44.0
me	47.8	30.3	7.5	10.1
Q_1	29.1	23.1	6.3	1.75
Q_3	56.3	31.2	12.95	14.55

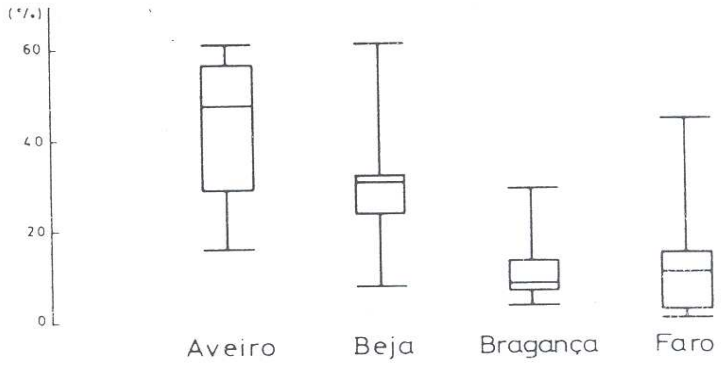


Figura 4: Diagrama de extremos e quartis para as taxas de arborização nos concelhos dos distritos de Aveiro, Beja, Bragança e Faro.

Uma análise rápida do diagrama da Figura 4 permite-nos a visualização da intensidade da arborização nos quatro distritos. Vejamos algumas observações: a amplitude total é maior em Beja e menor em Bragança; o valor central mais elevado é em Aveiro; as observações centrais (50%) estão muito mais concentradas em Beja e Bragança, sendo grande a concentração acima da mediana em Beja e abaixo da mediana em Bragança, etc.

Quando num conjunto de dados se detectar a presença de *outliers*, o diagrama de extremos e quartis deverá ser modificado de modo a incluir esta informação. Assim, devem calcular-se as **barreiras inferior e superior**, ver página 11, e marcar no esquema gráfico os chamados

valor adjacente inferior – que é o *menor* valor do conjunto dos dados (podendo ser o *mínimo*) maior ou igual à barreira inferior; e

valor adjacente superior – que é o *maior* valor do conjunto dos dados (podendo ser o *máximo*) menor ou igual à barreira superior.

A representação do diagrama, ver Fig. 5, apresenta agora mais informação do que anteriormente, será diferente da que foi atrás referida e designá-la-emos genericamente por **caixa de bigodes**.

Considerando novamente os dados do exemplo 2, o valor 79 é superior à barreira superior (para o exemplo referido a barreira superior é 77.5). Sendo assim, a representação para a caixa de bigodes pode ver-se na Figura 5.

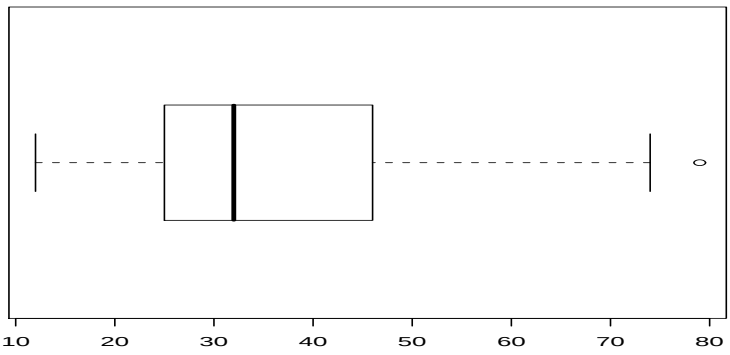


Figura 5: Caixa de bigodes.

Medidas descritivas para dados agrupados

Quando o número de valores observados é elevado e fazemos a sua condensação agrupando-os em classes, ou os dados nos são já fornecidos em tabelas com intervalos

de classe ou grupos com as frequências, teremos que considerar fórmulas de cálculo das características numéricas adequadas àquela representação das observações.

Suponhamos então que n observações foram agrupadas em c ($c < n$) classes, sendo $x'_1, x'_2, ..., x'_c$ os pontos médios de cada uma das classes e $n_1, n_2, ..., n_c$ as frequências absolutas de cada classe, i.e., $\sum_{i=1}^{c} n_i = n$.

As fórmulas de cálculo dos indicadores que temos estado a considerar são agora:

$$\textbf{Média agrupada} = \overline{x} \simeq \frac{\sum_{i=1}^{c} n_i \; x'_i}{n}; \tag{10}$$

$$\textbf{Variância agrupada} \simeq \frac{\sum_{i=1}^{c} (x'_i - \overline{x})^2 \; n_i}{n} = \frac{\sum_{i=1}^{c} x'^2_i \; n_i}{n} - \overline{x}^2 = \sum_{i=1}^{c} x'^2_i \; f_i - \overline{x}^2. \tag{11}$$

A média agrupada é uma medida que poderá servir como um indicador da "qualidade" de um agrupamento. Assim um bom agrupamento deverá ter a média agrupada muito próxima da média obtida considerando todas as observações.

Para o cálculo da **mediana** assim como de qualquer **quantil** de ordem θ, o algoritmo de cálculo aproximado destes indicadores é o seguinte:

- Identifica-se a primeira classe cuja frequência relativa acumulada seja superior ou igual a θ. Designemos por k essa classe e seja F_k a frequência relativa acumulada correspondente.
- O quantil de ordem θ é assim calculado:

$$Q^*_\theta \simeq x^{min}_k + \left(x^{max}_k - x^{min}_k\right) \frac{\theta - F_{k-1}}{f_k}, \tag{12}$$

onde F_{k-1} designa a frequência relativa acumulada da classe anterior à classe k, f_k a frequência relativa da classe k e x^{max}_k e x^{min}_k o limite superior e inferior da classe k, respectivamente. Se $k = 1$ toma-se $F_{k-1} = 0$.

O agrupamento dos dados permite-nos o cálculo de uma medida de localização importante que é a **moda amostral**. Uma vez determinada a **classe modal** – classe com maior frequência – existem várias fórmulas empíricas para determinar a moda (a mais simples consiste em tomar o ponto médio da classe modal), todas elas dando valores aproximados, sendo a mais conhecida a fórmula de King:

$$mo \simeq x^{min}_k + \left(x^{max}_k - x^{min}_k\right) \frac{f_{k+1}}{f_{k-1} + f_{k+1}} \tag{13}$$

onde f_{k-1} e f_{k+1} designam, respectivamente, a frequência da classe anterior e posterior à classe modal.

Amplitude total no caso de dados agrupados é dada por

$$A_{tot} \simeq x_c^{max} - x_1^{min}. \tag{14}$$

Observações finais

- O agrupamento dos dados permite ter uma perspectiva melhor das características amostrais subjacentes à amostra, desde que a amplitude das classes não tenha sido mal escolhido.
- O uso de métodos gráficos permite uma análise rápida e global das características dos dados, embora não permita fazer afirmações objectivas sobre eles.
- O ideal é combinar métodos gráficos e métodos analíticos.

ESTATISTICA DESCRITIVA A DUAS DIMENSÕES

Até aqui estudámos formas de descrever um conjunto de dados referentes à observação de uma variável em cada unidade estatística. Porém, nas mais variadas áreas de investigação, há interesse em estudar observações simultâneas de duas ou mais variáveis em cada unidade estatística, com o objectivo de procurar eventuais relações entre essas variáveis.

Neste curso iremos apresentar apenas o caso de duas dimensões. O objectivo da estatística descritiva a duas dimensões é o de estudar em simultâneo duas séries de observações, pondo em evidência "relações" existentes entre elas. O termo "relação" pode ter dois significados:

– a existência de uma conexão bem definida entre duas variáveis (ex. o perímetro e raio de uma circunferência estão relacionados por meio de uma expressão matemática), ou

– a existência de uma relação mais ténue e indefinida, como por exemplo a relação entre a altura e o peso de uma pessoa.

Como é evidente as duas relações anteriores não são relações do mesmo tipo, enquanto a primeira é bem definida, a segunda é uma relação vaga, a qual nos permite apenas dizer algo da forma : – em média quanto maior for a altura maior é o peso.

É de facto o comportamento em média de duas características que vai ser o objectivo da estatística descritiva a duas dimensões. O tipo de relações estudadas neste campo designam-se por **relações estatísticas** e entre as variáveis ligadas por uma relação estatística diz-se haver **correlação**. Se forem duas variáveis em estudo, a correlação é **simples**, havendo mais de duas a correlação é **múltipla**.

A correlação diz-se **positiva** se as duas características variam no mesmo sentido e **negativa** caso contrário.

Tal como para a estatística descritiva a uma dimensão há três aspectos que devem ser considerados para o estudo das relações existentes entre duas séries de observações, tomadas simultaneamente

- elaboração de tabelas condensando a informação sob a forma de distribuições de frequências;
- representação gráfica das observações;
- cálculo de parâmetros servindo para caracterizar numericamente as relações entre as variáveis.

Tabelas de frequência e representação gráfica

Consideremos as observações feitas em n indivíduos relativas a duas características, que iremos representar por n pares:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n).$$

O facto de escrevermos pares é independente da correlação que possa existir entre as duas variáveis, significa apenas a observação de duas características de um mesmo indivíduo, ou observações emparelhadas por alguma situação. Assim, por exemplo, podemos estar interessados em registar a altura e o peso de um grupo de n indivíduos.

Se n é grande é útil condensar os dados numa tabela de frequências bivariada, **quadro de dupla entrada** ou **tabela de contingência**, que é um quadro da forma

	y_1	y_2	...	y_q	
x_1	n_{11}	n_{12}	...	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2q}	$n_{2.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
x_p	n_{p1}	n_{p2}	...	n_{pq}	$n_{p.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.q}$	n

onde se supôs existirem q valores distintos de y e p valores distintos de x; n_{ij} designa o número de indivíduos para os quais foi observado o par (x_i, y_j).

A $n_{i.}$ e $n_{.j}$, soma dos elementos da linha i e dos elementos da coluna j , chamamos **frequências marginais** de x e y, respectivamente, sendo

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \qquad n_{.j} = \sum_{i=1}^{p} n_{ij} \qquad \text{e} \qquad \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = n. \tag{15}$$

Quando o número de linhas e o número de colunas é muito elevado é possível condensar os dados formando classes para os valores de x e para os valores de y. Veja-se o seguinte exemplo *Estatística, Teoria e Métodos. Pierre Dagnielie, 1º volume, 1973.*:

Exemplo 4.

Foram registados os pesos das folhas e das raízes de 1000 pés de *Cichorium intybus*, sendo alguns dos valores obtidos:

Folhas	Raízes	Folhas	Raízes
71	56	.	.
76	51	.	.
106	40	.	.
108	174	658	253
109	62	660	276
111	59	662	174
.	.	673	290
.	.	679	290
.	.	741	230

Uma distribuição de frequências destes dados, consistiu em agrupar os valores dos pesos das folhas em classes de 80 g e os pesos das raízes em classes de 40 g. Tem-se então

Raízes / Folhas	40 a 79	80 a 119	120 a 159	160 a 199	200 a 239	240 a 279	280 a 319	320 a 359	Totais
0 a 79	2								2
80 a 159	49	46	5	2					102
160 a 239	86	137	46	11					280
240 a 319	27	153	89	25	7				301
320 a 399	5	45	91	40	6				187
400 a 479		10	33	21	16	1	1		82
480 a 559		1	4	11	10	3			29
560 a 639			2	1	2	4		1	10
640 a 719				1		3	2		6
720 a 799					1				1
Totais	169	392	270	112	42	11	3	1	1000

Vejamos, como exemplo, algumas observações que é possível fazer com uma análise rápida deste quadro:

– Podemos dizer que, 'em média', quando o peso das folhas aumenta, também aumenta o peso das raízes. Observa-se ainda que há uma concentração de valores correspondendo a plantas que apresentam folhas com pesos situados entre 160 e 320 e raízes entre 40 e 160 (em gramas).

– Os totais 2,102,...,6,1 e 169,392,...,3,1 representam, respectivamente, o n $^{\underline{o}}$ total de plantas com pesos das folhas e pesos das raízes situados nos intervalos considerados no quadro, etc..

Representação gráfica

A representação gráfica sob a forma de histograma necessitava do uso de projecções o que torna difícil a sua visualização e interpretação.

Se o número n não for muito elevado, a série estatística das observações pode representar-se graficamente por meio de **diagramas de dispersão** ou **nuvens de pontos**, dando-nos uma ideia grosseira da correlação que poderá existir. Para isso marca-se num sistema de eixos cartesianos cada par (x_i, y_i).

Exemplo 5.
A percentagem (x) de caroteno em semente de trigo e a percentagem (y) de caroteno na farinha de trigo determinadas para 10 variedades de trigo , encontram-se no seguinte quadro:

x	1.18	2.13	1.41	1.42	1.50	1.25	1.65	1.24	1.48	1.35
y	2.39	3.11	2.15	1.96	2.02	1.76	2.10	2.12	2.28	1.86

A nuvem de pontos é

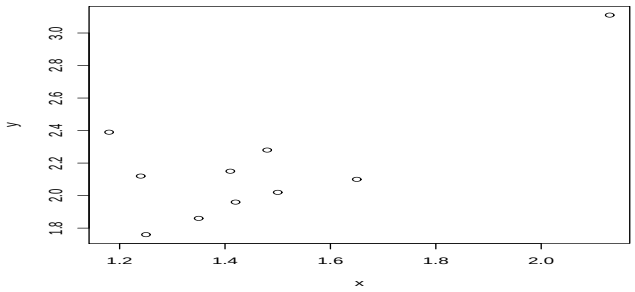


Figura 6: Nuvem de pontos para os dados do exemplo 5.

Na Figura 7. estão representados alguns exemplos de outros diagramas de dispersão e indicada a correlação existente entre as variáveis.

Figura 7: Nuvem de pontos e correlação associada.

Indicadores para dados bidimensionais

Consideremos novamente a série de n observações (x_i, y_i), $i = 1, ..., n$.

Localização da nuvem de pontos

As **médias marginais** de x e y, respectivamente

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

dão-nos o ponto $(\overline{x}, \overline{y})$ que é o centro de gravidade da nuvem de pontos.

Dispersão da nuvem de pontos

As dispersões marginais de x e y

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \qquad s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

dão-nos uma ideia da dispersão de cada uma das variáveis.

Interessa porém haver uma medida que dê informação sobre as duas variáveis em simultâneo, i.e., que consiga traduzir alguma relação que exista entre as variáveis.

Suponhamos a seguinte nuvem de pontos, na qual marcámos o centro de gravidade $(\overline{x}, \overline{y})$. Sobre ela tracemos rectas paralelas aos eixos passando por $(\overline{x}, \overline{y})$. A nuvem de pontos fica então dividida nas quatro regiões, que designaremos por I, II, III, IV.

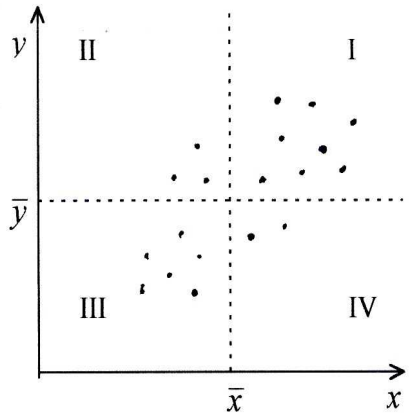


Figura 8: Divisão da nuvem de pontos por rectas paralelas aos eixos passando pelo centro de gravidade.

Como facilmente se verifica, para os pontos situados nas regiões I e III tem-se $(x_i - \overline{x})(y_i - \overline{y}) > 0$, enquanto para os pontos situados nas regiões II e IV se verifica $(x_i - \overline{x})(y_i - \overline{y}) < 0$.

Tendo em conta o que acabámos de observar, pode definir-se uma medida que caracterize o tipo de correlação existente entre duas séries de observações, baseada no produto dos desvios das variáveis a que se chama **covariância da amostra**, assim definida:

$$\boldsymbol{cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}}. \tag{16}$$

Se $cov(x,y) > 0$, significa que há predominância de elementos nas regiões I e III, i.e., entre as variáveis existe uma correlação positiva. Se $cov(x,y) < 0$ significa que existe uma correlação negativa entre as variáveis.

A $cov(x,y)$ é nula ou quase nula se há compensação entre o conjunto dos pontos em I/III e II/IV ou ainda, se os pontos observados se situam em torno, mas próximos, das rectas $x = \overline{x}$ ou $y = \overline{y}$.

Uma outra fórmula de cálculo da covariância, cuja dedução se deixa como **exercício** é

$$\boldsymbol{cov(x,y) = \frac{n\sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n(n-1)}}.$$

Propriedades da covariância

1. Seja (x_i, y_i) uma série de n observações e admitamos a transformação afim das variáveis $x'_i = a + bx_i \quad y'_i = c + dy_i$, com $a, b, c, d \in \mathbb{R}$ e $b \neq 0$, $d \neq 0$. Tem-se

$$cov(x', y') = bd\ cov(x, y).$$

Dem: Considerando a definição de covariância

$$cov(x', y') = \frac{\sum_{i=1}^{n}(x'_i - \overline{x'})(y'_i - \overline{y'})}{n-1} = \frac{\sum_{i=1}^{n}\ b(x_i - \overline{x})\ d(y_i - \overline{y})}{n-1} = bd\ cov(x, y).$$

A covariância, tal como a variância, é afectada por uma mudança de escala, mas não o é por uma mudança de origem.

2. $|cov(x,y)| \leq s_x s_y$

Dem:

Consideremos a seguinte expressão não negativa

$$\frac{1}{n-1}\ \sum_{i=1}^{n} [m(x_i - \overline{x}) - (y_i - \overline{y})]^2 \geq 0.$$

Desenvolvendo os quadrados obtemos

$$\frac{1}{n-1}\ \sum_{i=1}^{n}\left[m^2(x_i-\overline{x})^2-2m\ (x_i-\overline{x})(y_i-\overline{y})+(y_i-\overline{y})^2\right]\geq 0$$

$$m^2\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}-2m\ \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}+\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}\geq 0$$

$$m^2\ s_x^2-2m\ cov(x,y)+s_y^2\geq 0.$$

Trata-se então de uma desigualdade do 2ºgrau em m, que é não negativa sse

$$4cov^2(x,y)-4s_x^2s_y^2\leq 0 \qquad \text{ou seja}$$

$$|cov(x,y)|\leq s_xs_y.$$

A igualdade, $|cov(x,y)|=s_xs_y$, só se verifica se $m(x_i-\overline{x})-(y_i-\overline{y})=0$, condição esta que significa que todos os pontos observados se encontram sobre uma recta da forma

$$y-\overline{y}=m(x-\overline{x}),$$

não paralela aos eixos coordenados.

Exercício 3. Definindo

$$u_i=\frac{x_i-\overline{x}}{s_x} \qquad \text{e} \qquad v_i=\frac{y_i-\overline{y}}{s_y}$$

i.e., (u,v) são as variáveis reduzidas correspondentes a (x,y), provar que

$$cov(u,v)=\frac{cov(x,y)}{s_x\ s_y}.$$

Do que ficou dito atrás, a covariância é uma medida importante pela informação que ela nos dá sobre a correlação existente entre as variáveis: $cov(x,y)>0$ – há correlação positiva; $cov(x,y)<0$ – há correlação negativa.

Porém, apresenta a grande desvantagem de, tal como a variância, ser fortemente afectada por mudanças de escala nas observações. Sendo assim, a importância de podermos dispor de medidas independentes das unidades dos dados, leva-nos a considerar o resultado do exercício 3. Efectivamente, $cov(u,v)$ é uma medida independente das unidades, que se revela então de grande importância no nosso estudo.

Temos então uma nova medida a que se chama **coeficiente de correlação** [3] e se define como

[3]Também chamado **coeficiente de correlação de Pearson.**

$$\boldsymbol{r = r_{x,y} = \frac{cov(x,y)}{s_x\ s_y}} \qquad \text{com } s_x \neq 0 \text{ e } s_y \neq 0. \tag{17}$$

Vejamos agora algumas propriedades importantes do coeficiente de correlação , que justificam a sua importância como medida da relação existente entre duas variáveis

Propriedades do coeficiente de correlação

1. r tem sempre o mesmo sinal da covariância, o que é imediato da definição (17);

2. $-1 \leq r \leq 1$, basta ter em conta que $-s_x\ s_y \leq cov(x,y) \leq s_x\ s_y$ (propriedade 2. da covariância);

3. Dada a série de n observações (x_i, y_i), consideremos a transformação $x_i' = a + b\ x_i$ e $y_i' = c + d\ y_i$, com $b \neq 0$ e $d \neq 0$.

Se $(bd > 0)$ então $r_{x',y'} = r_{x,y}$. Caso $(bd < 0)$ $r_{x',y'} = -r_{x,y}$.

Verifiquemos o caso $(bd > 0)$, ficando como exercício a situação em que $(bd < 0)$. Para isso basta ter em conta que

$$r_{x',y'} = \frac{cov(a+bx, c+dy)}{s_{a+bx}\ s_{c+dy}} = \frac{bd\ cov(x,y)}{|b|s_x\ |d|s_y} = r_{x,y} \qquad \text{se } bd > 0.$$

Esta propriedade diz-nos que o coeficiente de correlação é independente de qualquer transformação linear positiva (o que ocorre em particular quando efectuamos a estandardização ou redução das variáveis em estudo).

4. O coeficiente de correlação é igual a 1, em valor absoluto (ou seja, $|cov(x,y)| = s_x\ s_y$), se todos os valores observados se encontram sobre uma recta; de declive positivo se $r = 1$, de declive negativo se $r = -1$.

Exercício 4. Mostrar que o coeficiente de correlação também se pode calcular usando a fórmula

$$\boldsymbol{r = \frac{n\ \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i\ \sum_{i=1}^n y_i}{\sqrt{\left[n\ \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2\right]\left[n\ \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2\right]}}}.$$

Podemos agora resumir as situações seguintes:

- $r = 1$ todos os pontos observados se encontram sobre uma recta de declive positivo (a).

- $r \simeq 1$ sugere que todos os pontos observados se encontram próximos de uma recta de declive positivo (b).
- $r \simeq 0$ significa ausência de associação linear, a nuvem apresenta um aspecto arredondado ou alongado segundo um dos eixos (c).
- $r \simeq -1$ sugere que todos os pontos observados se encontram próximos de uma recta de declive negativo (d).
- $r = -1$ todos os pontos observados se encontram sobre uma recta de declive negativo (e).

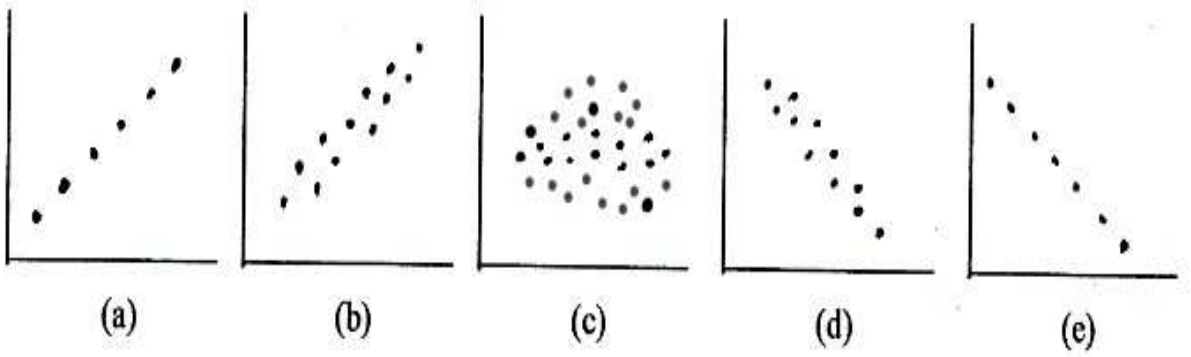


Figura 9: Possíveis nuvens de pontos para os valores do coeficiente de correlação.

Observação:

O coeficiente de correlação mede *a nitidez da ligação* existente entre duas variáveis, quando essa ligação é linear ou aproximadamente linear [(4)].

É importante porém, ter em conta o seguinte:

Ao estudarmos a relação existente entre duas variáveis x e y, um valor elevado para r nem sempre significa que x seja causa de y ou y seja causa de x. Afirmar, portanto, que há correlação entre duas variáveis não nos permite dizer que exista relação causal entre elas. Basta considerarmos o seguinte exemplo (Murteira e Black) de uma conclusão ridícula que se poderia ser levado a tirar:

– Da existência de uma correlação elevada entre o número anual de casos de insolação e a produção de trigo, não se deve concluir que a produção de trigo faz aumentar os casos de insolação (ou ao contrário). O que acontece é que ambos os fenómenos têm uma causa comum – os verões quentes (efectivamente, nesta situação, verificam-se boas colheitas e casos de insolação).

[4]Estatística, Teoria e Métodos. Pierre Dagnielie, 1$^{\underline{o}}$volume, 1973.

A regressão linear simples

Suponhamos que o diagrama de dispersão sugere a existência de uma relação linear entre as observações, i.e., os pontos (x_i, y_i) se situam em torno de uma recta. A essa recta é costume chamar-se **recta de regressão** e diz-se então que existe uma regressão linear simples [5] entre as duas características observadas. Pretendemos agora determinar a equação da recta de regressão, i.e., a equação de uma recta que representaremos por

$$\boldsymbol{y = b_0 + b_1 x}$$

que seja uma estimativa da verdadeira recta de regressão entre as duas variáveis em estudo e tendo como finalidade

- descrever a relação entre y e x;
- prever um valor de y para um dado valor de x.

Considerando os valores observados (x_i, y_i), designaremos por

$$\hat{y}_i = b_0 + b_1\ x_i$$

os valores de y estimados pela recta para cada x. Usa-se $\hat{y}_i$ para indicar que regra geral a ordenada da recta não coincide com a observação y_i. De facto o que se verifica é que se tem para cada par (x_i, y_i) a relação

$$\boldsymbol{y_i = b_0 + b_1\ x_i +\ e_i}, \tag{18}$$

sendo $e_i = y_i - \ \hat{y}_i$ designados por **resíduos**.

Para obter a recta é necessário determinar as estimativas dos coeficientes b_0 e b_1. Sendo assim, interessa como é óbvio, que os resíduos tenham os menores valores possíveis.

Um dos métodos usados para a determinação daqueles parâmetros é o **método dos mínimos quadrados** que consiste em determinar a e b por forma a minimizar a soma dos quadrados dos resíduos, ou seja, a minimizar

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \ \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1\ x_i)^2 = Q(b_0, b_1).$$

Como se pretende minimizar a função de duas variáveis $Q(b_0, b_1)$, as condições de estacionaridade são:

[5] Regressão linear simples quando existem apenas duas variáveis em estudo, uma dita explicativa, controlada, independente ou regressora, que regra geral se designa por x e a outra y, que se diz explicada, resposta ou dependente.
A regressão linear diz-se múltipla quando para uma variável y (dependente) se consideram duas ou mais variáveis independentes.

$$\left\{\begin{array}{l}\frac{\partial Q}{\partial b_0}=0\\ \frac{\partial Q}{\partial b_1}=0\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}2\ \sum_{i=1}^n (y_i-b_0-b_1\ x_i)=0\\ 2\ \sum_{i=1}^n x_i\ (y_i-b_0-b_1\ x_i)=0\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}nb_0+\ b_1\sum_{i=1}^n x_i=\sum_{i=1}^n y_i\\ b_0\sum_{i=1}^n x_i+\ b_1\sum_{i=1}^n x_i^2=\sum_{i=1}^n x_i\ y_i\end{array}\right. .$$

Estas equações são chamadas **equações normais**.
Vejamos algumas conclusões que é possível tirar da análise destas equações:

- da primeira tem-se $\sum_{i=1}^n (y_i-b_0-b_1\ x_i)=\sum_{i=1}^n (y_i-\hat{y}_i)=\sum_{i=1}^n e_i=0 \Leftrightarrow \overline{y}=\overline{\hat{y}}$, i.e., a soma dos resíduos é nula ou ainda, a média dos valores observados é igual à média dos valores estimados.

  Ainda da primeira equação tem-se, depois de dividir ambos os membros por n $b_0+b_1\ \overline{x}=\overline{y}$, i.e., a recta de regressão passa no ponto $(\overline{x},\overline{y})$.

- considerando a segunda equação e substituindo b_0 por $(\overline{y}-b_1\ \overline{x})$ tem-se

$$\sum_{i=1}^n x_i(y_i-\overline{y}+b_1\ \overline{x}-b_1\ x_i)=0$$

$$\sum_{i=1}^n x_i(y_i-\overline{y}+b_1(\overline{x}-x_i))=0$$

$$\sum_{i=1}^n (x_i\ y_i-x_i\ \overline{y}-b_1\ x_i(x_i-\overline{x}))=0$$

$$b_1=\frac{\sum_{i=1}^n x_i(y_i-\overline{y})}{\sum_{i=1}^n x_i(x_i-\overline{x})},$$

que pode escrever-se da forma

$$b_1=\frac{n\sum_{i=1}^n x_i\ y_i-\sum_{i=1}^n x_i\ \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2-(\sum_{i=1}^n x_i)^2}=\frac{\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^n (x_i-\overline{x})^2}; \tag{19}$$

ou ainda

$$\boldsymbol{b_1=\frac{cov(x,y)}{s_x^2}=r\ \frac{s_y}{s_x}}. \tag{20}$$

A $\boldsymbol{b_1}$ chama-se **coeficiente de regressão de** y **sobre** x.

Exercício 5. Sendo $y=b_0+b_1\ x$ a recta de regressão de y sobre x, determine os coeficientes da recta de regressão de v sobre u, sendo (u,v) as variáveis reduzidas correspondentes a (x,y).

Observações:

- Atendendo à expressão (20), o declive da recta, b_1, tem o mesmo sinal que $cov(x, y)$ e r. Além disso, como em muitas situações a variável x é uma variável controlada, deve evitar-se s_x^2 pequeno.

- Sendo $y = b_0 + b_1\ x$ a equação da recta de regressão, vejamos quais os valores estimados pela recta para x_i e $x_i + 1$

$$\widehat{y}_i = b_0 + b_1\ x_i \qquad \widehat{y'}_i = b_0 + b_1\ (x_i + 1).$$

Subtraindo as duas igualdades tem-se

$$b_1 = \widehat{y'}_i - \widehat{y}_i,$$

i.e., b_1 representa a variação esperada para y quando x aumenta de uma unidade.

Coeficiente de determinação

As equações de regressão determinam-se com diversos objectivos, sendo um deles o de predizer o valor de uma variável conhecendo o valor assumido pela outra.

Sendo assim, há a preocupação de avaliar o grau de precisão atingido pelas estimativas.

Para se definir uma medida de precisão, vejamos uma importante decomposição de $\sum_{i=1}^n (y_i - \overline{y})^2$.

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i - \overline{y}})^{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i - \hat{y}_i})^{\boldsymbol{2}} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{\hat{y}_i - \overline{y}})^{\boldsymbol{2}}. \tag{21}$$

Dem:
$\sum_{i=1}^n (y_i - \overline{y})^2$ pode escrever-se na forma

$$\sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + \sum_{i=1}^n (\hat{y}_i - \overline{y})^2.$$

Como $\hat{y}_i - \overline{y} = b_1(x_i - \overline{x})$, tem-se $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_{i=1}^n [y_i - \overline{y} - b_1(x_i - \overline{x})] b_1(x_i - \overline{x}) = b_1 \sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x}) - b_1^2 \sum_{i=1}^n (x_i - \overline{x})^2 = 0$ (tendo em conta a definição de b_1); temos, portanto, a relação (21). Aquela decomposição é costume também exprimir-se na forma

$$\boldsymbol{SQ_T = SQ_{RE} + SQ_R}, \quad \text{i.e.,} \tag{22}$$

soma dos quadrados totais =
= soma dos quadrados devidos aos resíduos
+ soma dos quadrados devidos à regressão.

Da relação (22) tem-se

$$1 = \frac{SQ_{RE}}{SQ_T} + \frac{SQ_R}{SQ_T} \quad \text{donde}$$

$$\frac{SQ_R}{SQ_T} = 1 - \frac{SQ_{RE}}{SQ_T} = 1 - \frac{s_e^2}{s_y^2} = \frac{s_y^2 - s_e^2}{s_y^2} = \frac{cov^2(x,y)}{s_x^2\ s_y^2} = r^2.$$

O quociente $\boldsymbol{\frac{SQ_R}{SQ_T} = r^2}$ dá-nos então **a proporção da variabilidade de y que é explicada pela regressão**, i.e., **põe em relevo em que medida o conhecimento de x serve para através de $\hat{y} = b_0 + b_1\ x$ estimar ou explicar a variação de y.**

A $\boldsymbol{r^2}$ chama-se **coeficiente de determinação** e trata-se então de **uma medida de precisão da recta de regressão**.

Por exemplo, suponhamos que para uma dada série de observações se tem $r = 0.70$, o que indicia uma correlação linear positiva entre as variáveis em estudo, que até parece razoável. Porém $r^2 = 0.49$, o que significa que recta de regressão não permite afinal obter resultados muito precisos.

A regressão pela origem

Em muitos problemas exige-se que a recta de regressão passe pela origem, i.e., que a equação da recta seja da forma $y = b_1\ x$.

Novamente o cálculo de b é feito por forma a minimizar a soma dos quadrados dos resíduos

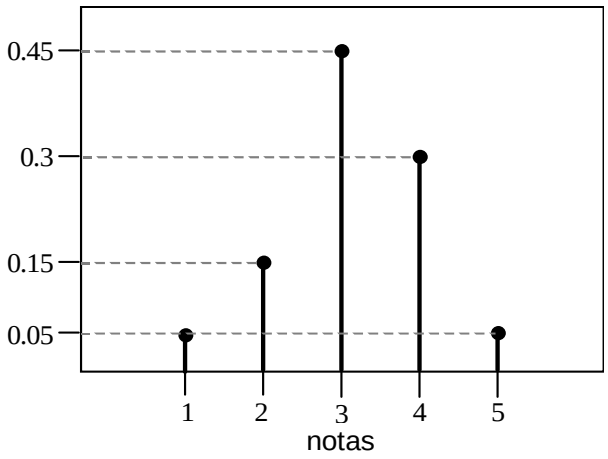
$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_1\ x_i)^2.$$

Como agora se trata de minimizar uma função de apenas uma variável, facilmente se obtém

$$\boldsymbol{b_1 = \frac{\sum_{i=1}^{n} x_i\ y_i}{\sum_{i=1}^{n} x_i^2}}. \tag{23}$$

Exercícios resolvidos

1. Para avaliar os conhecimentos de Estatística dos alunos do 12º ano em todo o País, seleccionou-se aleatoriamente uma amostra de 100 alunos, que foram submetidos a um teste cuja classificação é 1, 2, 3, 4 ou 5. Considera-se que um aluno adquiriu os conhecimentos suficientes se a sua classificação for superior ou igual a 3. O seguinte diagrama de barras representa a distribuição de frequências das notas obtidas pelos alunos.



a) Construa uma tabela de frequências absolutas, relativas e relativas acumuladas correspondente ao diagrama de barras.

b) Calcule a nota média e a nota mediana.

c) Calcule a percentagem de alunos da amostra que adquiriram os conhecimentos suficientes.

Resolução

Seja x_i a nota obtida pelo aluno i no referido teste, com $i = 1, \ldots, 100$, sendo cinco os valores distintos dessas notas, que designaremos por x'_j, com $j = 1, \ldots, 5$.

a) A tabela pedida é.

j	x'_j	Frequência absoluta (n_j)	Frequência relativa (f_j)	Frequência relativa acumulada (F_j)
1	1	$5{=}0.05 \times 100$	0.05	0.05
2	2	$15{=}0.15 \times 100$	0.15	$0.20{=}(5{+}15)/100$
3	3	$45{=}0.45 \times 100$	0.45	$0.65{=}(5{+}15{+}45)/100$
4	4	$30{=}0.3 \times 100$	0.3	$0.95{=}(65{+}30)/100$
5	5	$5{=}0.05 \times 100$	0.05	$1{=}(95{+}5)/100$

b) $\overline{x} = \sum_{j=1}^{5} x'_j \times f_j =$

$1 \times 0.05 + 2 \times 0.15 + 3 \times 0.45 + 4 \times 0.3 + 5 \times 0.05 = 3.15.$

$\widetilde{x} = 3$ (o primeiro x'_j com $F_j \geq 0.5$).

c) $1 - F_2$=1-0.20=0.8, ou seja, 80%.

2. Num estudo sobre a relação entre a produção de trigo - y (t/ha) e a quantidade de adubo - x (kg/ha), obtiveram-se os seguintes resultados :

x (kg/ha)	400	410	420	430	440	450	460
y (t/ha)	40	50	50	60	65	65	70

a) Estime a recta de regressão dos mínimos quadrados de y sobre x e diga qual é a precisão dessa recta. Comente.

b) Que valor prevê para a produção de trigo para uma quantidade de adubo de 450 kg/ha?

c) A estimação da recta de regressão pode ser feita considerando os valores centrados da variável x, i.e., $z_i = x_i - \overline{x}$, $i = 1, \cdots, 7$. Utilizando os resultados da alínea a) determine a equação da recta de regressão dos mínimos quadrados de y sobre z e a respectiva precisão.

Resolução

Temos $n = 7$ pares de observações para as quais

$\sum_{i=1}^{n} x_i = 3010 \qquad \sum_{i=1}^{n} y_i = 400$

$\sum_{i=1}^{n} x_i y_i = 173350 \qquad \sum_{i=1}^{n} x_i^2 = 1297100 \qquad \sum_{i=1}^{n} y_i^2 = 23550.$

a) Então $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 430$ kg/ha e $\overline{y} = 57.14$ t/ha.
e considerando a fórmula de cálculo habitual para a variância
$s_x^2 = \frac{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)} = 466.67$ (kg/ha)^2, do mesmo modo $s_y^2 = 115.476$ (t/ha)^2 e para a covariância

$$cov(x, y) = \frac{n \sum_{i=1}^{n} x_i \; y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(n-1)} = 225.$$

As estimativas dos coeficientes da recta de regressão dos mínimos quadrados são

$$b_1 = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = 0.482$$

$$b_0 = \overline{y} - b_1\overline{x} = -150.179$$

A precisão da recta é dada por $R^2 = r^2 = \left(\frac{cov(x,y)}{s_x s_y}\right)^2 = 0.939$, isto é, aprox. 94% da variabilidade total é explicada pela recta de regressão, portanto apresenta uma boa precisão.

b) Uma vez que a recta de regressão permite obter $\hat{y_i} = -150.179 + 0.482\ x_i$, então para $x_i = 450$ a recta prevê uma produção média de trigo de $\hat{y} = -150.179 + 0.482 \times 450 = 66,721\ t/ha$.

c) Considerando agora $z_i = x_i - \overline{x} \quad i = 1, ..., 7$.

Dado que a recta de regressão dos mínimos quadrados $y = b_0 + b_1 x$ se pode apresentar na forma

$y - \overline{y} = b_1(x - \overline{x})$ (tem declive b_1 e passa no ponto $(\overline{x}, \overline{y})$)

temos $y - \overline{y} = b_1\ z$. Então a equação da recta de regressão de y em z é (note-se que $\overline{z} = 0$) $\quad y = \overline{y} + b_1 z$,

portanto a nova recta tem o mesmo coeficiente de regressão da anterior e a ordenada na origem é $\overline{y}$.

Vejamos o valor do coeficiente de correlação

$$r_{y,z} = \frac{cov(y,z)}{s_y\ s_z} = \frac{cov(y, x-\overline{x})}{s_y\ s_{x-\overline{x}}} = \frac{cov(y,x)}{s_y\ s_x} = r_{x,y}$$

pois a covariância e o desvio padrão não são afectadas por uma mudança de localização, isto é $cov(a+x, b+y) = cov(x,y)$ e $s_{a+x} = s_x$.

Recorde-se que uma das propriedades do coeficiente de correlação é que ele não é afectado por mudanças de localização, portanto nem era necessário efectuar aqueles cálculos.

Portanto a precisão $R^2 = r^2$ é a mesma.

Exercícios propostos

1. Os dados seguintes, que já se encontram ordenados, representam os tempos de resposta, em segundos, obtidos quando se trabalha num dado terminal de computador:

1.26 1.28 1.30 1.37 1.43 1.43 1.43 1.46 1.47 1.48 1.48 1.49 1.51 1.51 1.51
1.51 1.52 1.53 1.53 1.55 1.56 1.57 1.60 1.60 1.61 1.64 1.64 1.65 1.68 1.74

a) Determine as seguintes características amostrais: média, variância, mediana e o terceiro quartil.

b) Escolhendo uma amplitude de classe conveniente, construa uma tabela de frequências relativas para os dados observados.

c) Usando os cálculos da alínea anterior represente o histograma correspondente. Interprete-o.

2. Sejam (x_i, y_i), n pares de observações. Considere $z_i = b_0 + b_1\ x_i$, com b_0 e b_1 constantes. Exprima o coeficiente de correlação de z e y em função do coeficiente de correlação de x e y.

3. Da análise do consumo médio de energia por agregado familiar durante 10 dias de um mês de Inverno numa dada cidade obtiveram-se os seguintes resultados:

Temp. diária média(^{o} C)	15	14	12	14	12	11	11	10	12	13
Cons. médio de energia (KW)	4.3	4.4	5.3	4.6	5.5	5.9	5.7	6.2	5.2	5.0

O modelo de regressão linear simples foi usado para estudar a relação entre o consumo médio de energia por agregado familiar e a temperatura diária média.

a) Escreva a equação da recta de regressão dos mínimos quadrados. Diga o valor do coeficiente de regressão e interprete o seu significado.

b) Qual o consumo médio previsto num dia de temperatura média igual a 10^o C? E num dia de temperatura média de 20^o C? Comente os resultados obtidos.

c) Suponha que lhe é solicitado que o valor do consumo médio de energia seja expresso em W. Deduza a relação existente entre o coeficiente de regressão e a ordenada na origem obtidos com os dados apresentados e com os dados depois de considerada a transformação proposta.

4. Num estudo sobre a relação entre a produção de trigo - y (t/ha) e a quantidade de adubo - x (kg/ha), recolheram-se 20 observações que conduziram aos seguintes resultados (nas unidades respectivas):

n	$\overline{x}$	$\overline{y}$	$var(x)$	$var(y)$	r_{xy}
20	450	57.5	466.56	115.56	0.97

a) Estime a recta de regressão dos mínimos quadrados de y sobre x e diga qual é a precisão dessa recta. Comente.

b) Que valor prevê para a produção de trigo para uma quantidade de adubo de 460 kg/ha?

c) Suponha que os valores observados tinham sido registados todos na mesma unidade (kg/ha). Actualize o quadro anterior de modo a ficar coerente com aquela unidade. Justifique.

Referências bibliográficas

Bhattacharyya, G.K. and Johnson R.A.(1977), *Statistical Concepts and Methods*, John Wiley & Sons Inc.

Dagnelie, P.(1973), *Estatística, Teoria e Métodos*, trad. do Prof. Doutor A. St.Aubyn, Europa América, vol I e II.

Daniel, W. W.(1995), *Biostatistics: A Foundation for Analysis in the Health Sciences*. John Wiley.

Galvão de Mello, F. (1993)- *Probabilidades e Estatística. Conceitos e métodos fundamentais*. Vol I. Escolar Editora.

Hoaglin,D.C., Mosteller,F. e Tukey, J.W.(1992), *Análise Exploratória de Dados. Técnicas Robustas* (trad. por Dinis Pestana e outros), Edições Salamandra.

Murteira, B. (1993), *Análise Exploratória de Dados. Estatística Descritiva*, Mc Graw Hill.

Murteira, B., Ribeiro, C.S., Silva, J.A. e Pimenta C.(2007), *Introdução à Estatística*, 2ª edição, Mc Graw Hill.

Pestana, D.D. e Velosa, S.F. (2006), *Introdução à Probabilidade e à Estatística* . 2ª edição, Fundação Calouste Gulbenkian.