Departamento de Ciências e Engenharia de Biossistemas
Agrupamento de Matemática

Matemática II

Estatística Descritiva

2013/2014

(F. Valente e M. Mesquita)

Revisões de Estatística

- Estatística - ciência que se ocupa da recolha, organização e análise de informação, com a finalidade de inferir de um conjunto limitado de informação para o todo e eventualmente prever a evolução futura de um fenómeno.
- Subáreas da Estatística:

Obtenção dos dados	Análise dos dados	Modelação	Indução
•*Amostragem*	•*Estatística descritiva*	•*Teoria da Probabilidade*	•*Inferência Estatística*
•*Planeamento de Experiências*	•*Análise Exploratória*		Prescrutação do futuro •*Previsão*

Conceitos básicos

- **Dados** – conjunto de informação que constitui o objecto de estudo da Estatística.
- **População** (ou universo) – conjunto de todos os elementos/valores que se pretendem estudar.
- **Unidade estatística** (ou amostral) – elemento da população que é objecto de observação.
- **Variável** – característica comum aos elementos da população (cujo valor pode ser diferente de elemento para elemento).
- **Amostra** – subconjunto de elementos extraídos de uma população (conjunto de todas as observações da característica em estudo efectivamente recolhidas).

Dados estatísticos

- Os dados (e respectiva variável) podem ser de natureza
  - qualitativa
    - nominal (a ordem das categorias não tem significado)
    - ordinal (há uma ordenação natural das categorias)
  - quantitativa
    - discreta (provêm de contagens)
    - contínua (provêm de medições)
- Exemplos:

Classes de rendimento	Número de famílias
0-500	50
500-1000	20
1000-2000	20
+2000	10

Número de filhos	Número de casais
0	4
1	6
2	5
3	3

Níveis de classificação	Número de alunos
A	10
B	20
C	15
D	5

Sexo	Número de alunos
F	30
M	20

Exemplo 1

Para avaliar a taxa de sucesso no primeiro semestre de um curso universitário com **cinco** disciplinas, foram inquiridos 50 alunos inscritos na totalidade dessas disciplinas. As respostas quanto ao número de disciplinas realizadas com aproveitamento por aluno foram as seguintes:

1 5 1 4 1 3 2 5 5 1 4 5 2 4 5
4 4 1 2 5 3 5 3 0 2 1 3 4 5 4
5 2 5 3 0 4 5 1 4 0 4 4 1 2 1
5 5 5 2 3

Exemplo 2

Um dos principais indicadores da poluição atmosférica nas grandes cidades é a concentração de ozono na atmosfera. Num dado Verão, e numa dada cidade, registaram-se 78 valores dessa concentração, tendo-se obtido os seguintes valores:

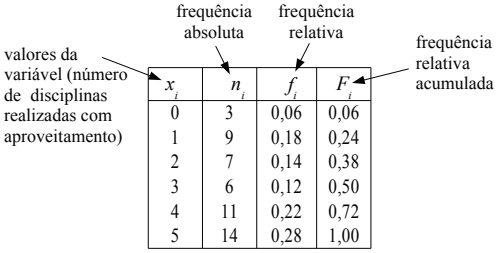
3.5 6.2 3.0 3.1 5.1 6.0 7.6 7.4 3.7 2.8 3.4 3.5
1.4 5.7 1.7 4.4 6.2 4.4 3.8 5.5 4.4 2.5 11.7 4.1
6.8 9.4 1.1 6.6 3.1 4.7 4.5 5.8 4.7 3.7 6.6 6.7
2.4 6.8 7.5 5.4 5.8 5.6 4.2 5.9 3.0 3.3 4.1 3.9
6.8 6.6 5.8 5.6 4.7 6.0 5.4 1.6 6.0 9.4 6.6 6.1
5.5 2.5 3.4 5.3 5.7 5.8 6.5 1.4 1.4 5.3 3.7 8.1
2.0 6.2 5.6 4.0 7.6 4.7

Revisões de Estatística

- A análise inicial dos dados tem como principais objectivos:
  - a exploração dos dados para descobrir/identificar aspectos ou padrões de maior interesse;
  - a representação dos dados de modo a destacar esses aspectos ou padrões:
    - condensar os dados observados sob a forma de quadros;
    - representar graficamente os dados;
    - calcular indicadores numéricos de localização e de dispersão.

Agrupamento dos dados

- Tabela de frequências
  - caso de dados de natureza discreta, com um número pequeno de valores distintos



x_i	n_i	f_i	F_i
0	3	0,06	0,06
1	9	0,18	0,24
2	7	0,14	0,38
3	6	0,12	0,50
4	11	0,22	0,72
5	14	0,28	1,00

Frequência

- **Frequência absoluta** (n_i) – número de observações iguais a x_i
- **Frequência relativa** (f_i) – fracção do número total de observações iguais a x_i $(f_i = n_i / \sum n_i)$
- **Frequência relativa acumulada** (F_i) – fracção do número total de observações menores ou iguais a x_i $(F_i = \sum_{k=1}^{i} f_k)$

Agrupamento dos dados (cont.)

- Tabela de frequências
  - dados de natureza contínua ou dados de natureza discreta com um número elevado de valores distintos
    - neste caso há que agrupar os dados em classes seguindo, por exemplo, o procedimento seguinte:
      1) determinar o máximo e o mínimo do conjunto dos dados (avaliando a amplitude total = max - min);
      2) escolher o número de classes;
      3) definir os vários intervalos de classe fixando os seus limites: os intervalos têm de ser disjuntos e o domínio da variável tem de estar contido na união de todos os intervalos;
      4) contar os valores pertencentes a cada classe, determinando a frequência absoluta e relativa de cada classe.

Número de classes

- A escolha do número de classes, além de se basear na experiência e nos objectivos do investigador, depende de dois factores: a dimensão (n) e a amplitude total da amostra.
- Exemplo de uma regra para escolha do número de classes com amplitude constante:

  Regra de Sturges – toma-se como número de classes o inteiro (m) mais próximo de

$$1 + (\log_2 n) = 1 + \frac{\log n}{\log 2}$$

Exemplo 2 (cont.)

1) min = 1.1 e max = 11.7
2) Pela regra de Sturges $m = 7 \leftarrow 1 + (\log_2 n) = 7.285$
3) Amplitude das classes $h = 1.5 \leftarrow$ (max-min)/7 =1.51

Tabela de frequências

Classe i	Intervalo de classe	Ponto médio	n_i	f_i	F_i
1	]0.0, 1.5]	0,75	4	0,05	0,05
2	]1.5, 3.0]	2,25	9	0,12	0,17
3	]3.0, 4.5]	3,75	20	0,26	0,42
4	]4.5, 6.0]	5,25	24	0,31	0,73
5	]6.0, 7.5]	6,75	15	0,19	0,92
6	]7.5, 9.0]	8,25	3	0,04	0,96
7	]9.0,10.5]	9,75	2	0,03	0,99
8	]10.5,12.0]	11,25	1	0,01	1,00

Representação gráfica

- **Diagrama de barras** – num referencial ortonormado marcam-se no eixo das abcissas todos os valores que a variável pode tomar e, em cada um, traçam-se barras verticais de ordenada igual à frequência absoluta/relativa observada
  - utiliza-se quando o conjunto de dados é discreto com um número moderado de valores possíveis (exemplo 1).
- **Histograma** – diagrama de áreas formado por uma sucessão de rectângulos adjacentes, cuja base é a amplitude de cada classe e cuja área representa a frequência relativa na classe
  - utiliza-se com dados agrupados em classes (exemplo 2).

Indicadores numéricos

- Em geral, é importante resumir os dados de natureza quantitava, calculando algumas características numéricas da amostra de modo a ter informação sobre a sua
  - **localização**
    - média, mediana, quantis e moda
  - **dispersão**
    - amplitude total, amplitude inter-quartil, variância, desvio padrão e coeficiente de variação

Indicadores de localização: média

- Seja $\{x_1, x_2, \ldots, x_n\}$ uma amostra com n observações . Define-se **média aritmética**, ou simplesmente média, como

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Nota: sempre que possível deve-se calcular a média a partir de dados não classificados; no entanto, quando apenas se dispõe de dados classificados pode-se calcular a

$$\textbf{média agrupada} = \bar{x}' = \frac{\sum_{i=1}^{c} n_i x'_i}{n}$$

em que c é o número de classes, n_i é a frequência absoluta da classe i e x'_i é o ponto médio da classe i.

Algumas propriedades da média

- A média é o ponto de equilíbrio ("centro de gravidade") das observações.
- A média de uma transformação linear dos dados é a transformação linear da média dos dados originais, i.e., $x_i^* = a + b\,x_i \Rightarrow \overline{x^*} = a + b\,\bar{x}$.
- A média é muito sensível a valores extremos (*outliers*)

Indicadores de localização: mediana

- Seja $\{x_1, x_2, \ldots, x_n\}$ uma amostra com n observações, e sejam $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ as observações ordenadas $(x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)})$. Define-se **mediana** como

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

- A mediana é o valor que separa as 50% das observações inferiores das 50% superiores.
- Ao contrário da média, a mediana é uma medida de localização resistente a erros grosseiros e *outliers*.

Indicadores de localização: quantil

- Define-se **quantil de ordem *p*** ($0 < p < 1$) como

$$Q_p = \begin{cases} x_{([np]+1)} & \text{se } np \text{ não é inteiro} \\ \dfrac{x_{(np)} + x_{(np+1)}}{2} & \text{se } np \text{ é inteiro} \end{cases}$$

em que $[np]$ designa o maior inteiro contido em np.

Indicadores de localização: quantil (cont.)

- Para dados agrupados, o cálculo do quantil de ordem p (Q_p) pode ser feito de diversas maneiras. A mais usual corresponde a localizar a primeira classe (k) cuja frequência relativa acumulada é superior ou igual a p ($F_k \geq p$) e considerar o seu extremo inferior ($x_k^{\min}$) como uma aproximação inicial a que é preciso adicionar uma correcção. Essa correcção é proporcional à amplitude da classe (h_k) e ao número de observações que faltam para atingir os $p \times 100\%$ de observações inferiores Q_p

$$Q'_p = x_k^{\min} + h_k \frac{p - F_{k-1}}{f_k}$$

Indicadores de localização: moda

- Para dados qualitativos ou dados discretos não agrupados, a **moda** define-se como o valor mais frequente.
- Para dados agrupados em classes de amplitude constante (h), a moda é um valor da classe com frequência mais elevada (**classe modal**) calculado por regras empíricas, por exemplo

$$\text{mod} = x_k^{\min} + h \frac{f_{k+1}}{f_{k-1} + f_{k+1}},$$

em que $x_k^{\min}$ é o limite inferior da classe modal, f_{k-1} e f_{k+1} são as frequências, respectivamente, da classe anterior e da classe posterior à classe modal.

Indicadores de dispersão

- **Amplitude total**: $A_{tot} = \max - \min$
- **Amplitude inter-quartil**: $AIQ = Q_{0.75} - Q_{0.25}$
- **Variância**: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}{n(n-1)}$
- **Desvio padrão**: $s = \sqrt{(s^2)}$

Indicadores de dispersão

- Para dados agrupados em c classes:
  - **Amplitude total**: $A_{tot} \simeq x_c^{max} - x_1^{min}$
  - **Variância**:

$$s'^2 = \frac{\sum_{i=1}^{c} (x'_i - \bar{x}')^2 n_i}{n} = \frac{\sum_{i=1}^{c} n_i x'^2_i}{n} - \bar{x}'^2 = \sum_{i=1}^{c} x'^2_i f_i - \bar{x}'^2$$

em que x_c^{max} é o limite superior da última classe, x_1^{min} é o limite inferior da primeira classe, n_i é a frequência absoluta da classe i, $\bar{x}'$ é a média agrupada, x'_i é o ponto médio da classe i e f_i é a frequência relativa da classe i.

Algumas propriedades dos ind. de dispersão

- A amplitude total, a variância e o desvio padrão são medidas de dispersão muito influênciadas por valores extremos.
- A amplitude inter-quartil é uma medida mais resistente a erros grosseiros e *outliers*.
- $s^2 \geq 0$ e $s \geq 0$
- Seja $\{x_1, x_2, \ldots, x_n\}$ uma amostra com n observações com variância s_x^2 e $x_i^* = a + bx_i$, $i = 1, \ldots, n$.
  A variância das novas observações é $s_{x^*}^2 = b^2 s_x^2$
  e o desvio padrão é $s_{x^*} = |b| s_x$.

Dispersão absoluta e relativa

- As medidas de **dispersão** anteriores são chamadas de **absoluta**s pois dependem da unidade da variável a que se referem.
- Medidas de **dispersão relativa** são independentes da unidade da variável, permitindo comparar dados cujas unidades são diferentes ou que diferem consideravelmente em grandeza.
- De um modo geral, as medidas de dispersão relativa são definidas como

$$\text{dispersão relativa} = \frac{\text{dispersão absoluta}}{\text{localização}}$$

Coeficiente de variação

- A medida de dispersão relativa mais usada é o **coeficiente de variação**,

$$CV = \frac{s}{\bar{x}} \times 100\%.$$

Nota: CV só pode ser ser usado quando a variável toma valores de um mesmo sinal, i.e., só positivos ou só negativos.

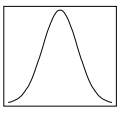
- O CV pode ser interpretado como a fracção da dispersão pela qual a localização é responsável
  - em muitas situações, quanto maior é a localização maior tende a ser a dispersão.

Assimetria

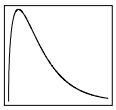
- As distribuições podem classificar-se em:
  - **simétricas**
    média = mediana = moda
  - **enviesadas à esquerda ou assimétricas positivas**
    média > mediana > moda
  - **enviesadas à direita ou assimétricas negativas**
    média < mediana < moda
- A **assimetria** (afastamento da simetria) é tanto maior quanto mais afastadas estiverem média, mediana e moda.

Assimetria (exemplos)

- Distribuição **simétrica**



- Distribuição **assimétrica positiva**



- Distribuição **assimétrica negativa**

Outliers

- ***Outlier*** é um elemento que se afasta do padrão geral dos dados e a que se deve dar atenção especial.
- Regra prática para identificação de (candidatos) a *outliers*:
  - Um valor x_i é considerado um ***outlier*** quando

$$x_i < Q_1 - 1.5(Q_3 - Q_1) \quad \text{ou} \quad x_i > Q_3 + 1.5(Q_3 - Q_1),$$

em que $Q_1 = Q_{0.25}$ e $Q_3 = Q_{0.75}$ são o 1º e o 3º quartil, respectivamente;

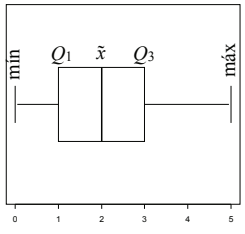
Nota: os valores $Q_1 - 1.5(Q_3 - Q_1)$ e $Q_3 + 1.5(Q_3 - Q_1)$ são as chamadas barreiras inferior e superior, respectivamente;

Outliers (cont.)

- A identificação e interpretação de *outliers* é uma tarefa complexa e altamente subjectiva.
  - Este tipo de observações pode resultar:
    - de erros humanos cometidos ao medir ou ao registar os dados,
    - da própria natureza do fenómeno em estudo.
- A eliminação de potenciais *outliers* deve fazer-se com prudência.
  - É aconselhável proceder à análise estatística com e sem eles e avaliar a sua influência na análise e interpretação dos resultados. Se as diferenças forem apreciáveis há que relatar este facto e, eventualmente, recolher mais dados e recomeçar a análise.

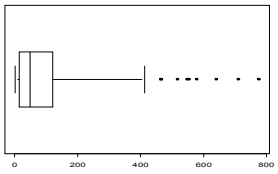
Diagrama de extremos-e-quartis

- O **diagrama de extremos-e-quartis** (ou **caixa-de-bigodes** , *box-and-whiskers plot*) é um gráfico que permite representar em simultâneo medidas de localização e dispersão de uma amostra:
  - transmite de modo imediato uma ideia da localização, dispersão e forma da população de que foi extraída a amostra;
  - é um instrumento adequado para comparar várias amostras.



Caixa-de-bigodes com *outliers*

- A presença de *outliers* leva a modificar a caixa-de-bigodes. Os "bigodes" vão somente até aos elementos do conjunto de dados mais extremos que não são *outliers.* Os *outliers* são marcados com pontos.



Estatística descritiva a duas dimensões

- Quando se consideram colecções de pares de variáveis (x_i, y_i), $i = 1,...,n$, deixamos de estar interessados em explorar isoladamente cada uma das variáveis. O **objectivo** passa a ser **o estudo da variação conjunta dessas variáveis**, procurando pôr em evidência "relações" eventualmente existentes entre elas.
- Em Estatística, não são relações determinísticas (relações funcionais) que interessam, mas sim a variação em média das duas variáveis (**relação estatística**).

Correlação

- Entre duas variáveis quantitativas ligadas por uma relação estatística diz-se que existe **correlação**.
  - Quando existe correlação, os fenómenos observados não estão indissoluvelmente ligados, mas a intensidade de um é acompanhada tendencialmente pela intensidade do outro, no mesmo sentido (**correlação positiva**) ou em sentido inverso (**correlação negativa**).
- Exemplos:
  a) Relação entre o preço do vinho e o montante da colheita em cada ano;
  b) Relação entre o peso e a altura de um ser humano adulto.

Diagrama de dispersão

- **Diagrama de dispersão** (ou nuvem de pontos) é a representação gráfica de um conjunto de pares de observações (x_i, y_i), $i = 1,..., n$, num sistema de eixos cartesianos.
  - O ponto $(\bar{x}, \bar{y})$ é o centro de gravidade da nuvem de pontos.
- Através da análise gráfica obtém-se uma ideia inicial da associação estatística entre as duas variáveis:
  - linear ou não;
  - crescente ou decrescente.

Indicadores para dados bidimensionais: covariância

- Para além dos indicadores numéricos que caracterizam individualmente cada uma das amostras $(\bar{x}, \bar{y}, s_x^2, s_y^2, ...)$, podem-se definir novos parâmetros para descrever as relações existentes numa amostra bivariada.

  Define-se **covariância** de x e y como

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

Algumas propriedades da covariância

- Seja $\{(x_i, y_i)\}_{i=1}^{n}$ uma amosta bivariada e sejam $x_i^* = a + bx_i$ e $y_i^* = c + dy_i$ transformações lineares dos dados. Então $cov(x^*, y^*) = bd\, cov(x, y)$.
- A covariância é um indicador da associação (linear) entre duas variáveis:
  - quando $cov(x, y) > 0$ há correlação positiva;
  - quando $cov(x, y) < 0$ há correlação negativa;
- A covariância tem, no entanto, um forte inconveniente: depende da unidade de medida usada, sendo fortemente afectada por mudanças de escala nas observações.

Indicadores para dados bidimensionais: coeficiente de correlação

- Define-se **coeficiente de correlação** de uma amostra bivariada como

$$r = r_{xy} = \frac{cov(x, y)}{s_x s_y} \quad \text{com } s_x \neq 0 \text{ e } s_y \neq 0.$$

- O coeficiente de correlação é também um indicador numérico que permite avaliar o grau de **associação linear** entre duas variáveis. É, no entanto, uma indicação meramente ordinal:
  - podemos ordenar diferentes valores de *r* e com isso ter uma ideia sobre quais as variáveis que têm um grau de associação(linear) mais forte; não faz, no entanto, sentido dizer que *r* = 0.90 representa uma correlação duas vezes mais forte do que a dada por *r* = 0.45.

Propriedades do coeficiente de correlação

- r tem sempre o mesmo sinal da covariância.
- $-1 \leq r \leq 1$.
- O valor de *r* não é afectado, em valor absoluto, por transformações lineares das variáveis, i.e., se r_{xy} é o coeficiente de correlação da amostra bivariada $\{(x_i, y_i)\}_{i=1}^{n}$ e $x_i^* = a + bx_i$ e $y_i^* = c + dy_i$ são transformações lineares das variáveis, então

$$r_{x^* y^*} = r_{xy} \quad \text{se } bd > 0$$

ou

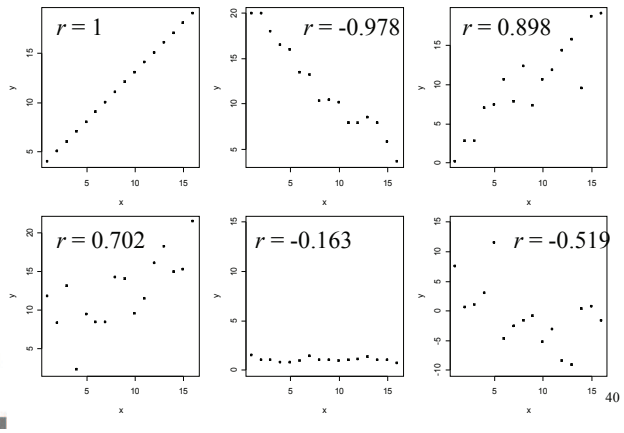
$$r_{x^* y^*} = -r_{xy} \quad \text{se } bd < 0.$$

- $|r_{xy}| = 1$ se todos os valores observados se encontram sobre uma recta.

Interpretação do valor de *r*

- r = 1 se todos os pontos observados se encontram sobre uma recta de declive positivo.
- r ≈ 1 se todos os pontos observados se encontram próximos de uma recta de declive positivo.
- r ≈ 0 se a nuvem apresenta um aspecto arredondado ou alongado segundo um dos eixos.
- r ≈ −1 se todos os pontos observados se encontram próximos de uma recta de declive negativo.
- r = −1 se todos os pontos observados se encontram sobre uma recta de declive negativo.

Exemplos



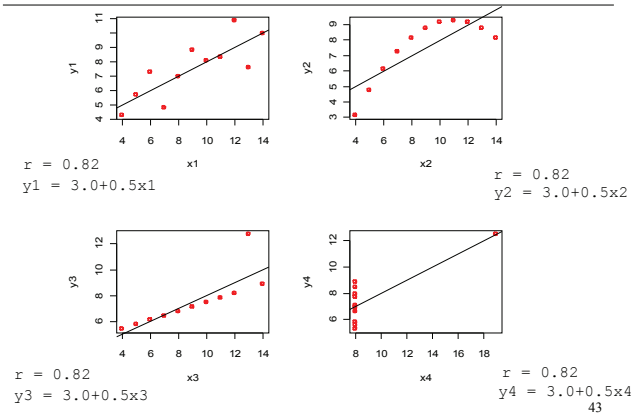
Interpretação do valor de *r* (cont.)

- Um valor de *r* elevado não significa, necessariamente, uma associação linear forte.
  - Pode ser uma consequência da estrutura da nuvem de pontos ou da existência de pontos afastados.
- r ≈ 0 não significa mais do que a ausência de qualquer relação ou tendência linear entre as variáveis.
  - Uma das variáveis pode ser completamente determinada pela outra e a correlação ser nula.
- Não confundir associação estatística com causalidade:
  - um valor elevado de *r* não significa que *x* seja causa de *y* ou que *y* seja causa de *x*.

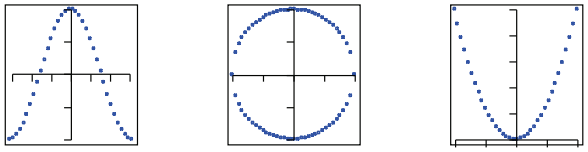
Dados de Anscombe

```
> anscombe

   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
4   9  9  9  8  8.81 8.77  7.11  8.84
5  11 11 11  8  8.33 9.26  7.81  8.47
6  14 14 14  8  9.96 8.10  8.84  7.04
7   6  6  6  8  7.24 6.13  6.08  5.25
8   4  4  4 19  4.26 3.10  5.39 12.50
9  12 12 12  8 10.84 9.13  8.15  5.56
10  7  7  7  8  4.82 7.26  6.42  7.91
11  5  5  5  8  5.68 4.74  5.73  6.89

média
   x1    x2    x3    x4    y1    y2    y3    y4
 9.00  9.00  9.00  9.00  7.50  7.50  7.50  7.50
variância
   x1    x2    x3    x4    y1    y2    y3    y4
11.00 11.00 11.00 11.00  4.13  4.13  4.12  4.12
```

Dados de Anscombe



Exemplos: $r = 0$



Situações determinísticas: $y = f(x)$

Quadro de correlação

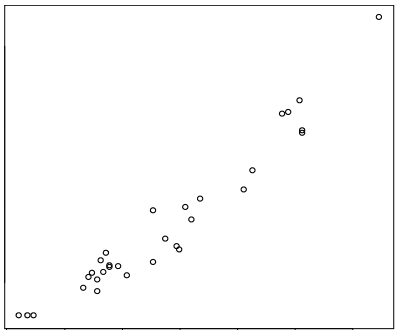
- Quando a amostra bivariada é muito grande, pode ser útil condensar os dados numa distribuição (agrupada ou não) de frequências bivariadas, formando um **quadro de correlação** ou quadro de dupla entrada.
- A presença de frequências mais elevadas sobre ou na vizinhança de uma das diagonais do quadro de correlação evidência a existência de correlação (positiva ou negativa) entre as duas variáveis em questão.
  - Quando só se tem acesso a dados agrupados é possível calcular valores (aproximados) de covariância e do coeficiente de correlação.

Regressão linear simples – contexto descritivo

- A regressão linear estima a relação linear entre duas variáveis:
  x – variável preditora (independente)
  y – variável resposta (dependente)
- Por vezes não é claro qual das variáveis é a resposta e qual é a variável independente (altura, peso). Neste caso deve usar-se uma análise de correlação.
- A regressão linear simples estima relações da forma

$$y = b_0 + b_1 x$$

Regressão linear simples - exemplo



$n = 31$ pares de observações em cerejeiras

x – diâmetro à altura do peito (DAP)
y - volume tronco

R package
`trees(datasets)`

Regressão linear simples – contexto descritivo

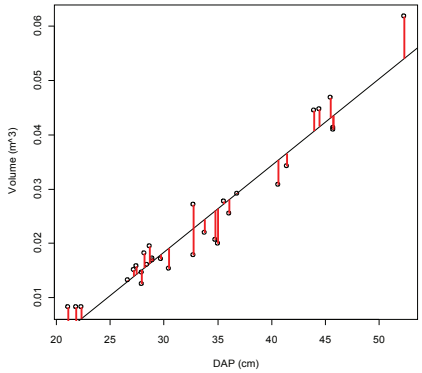
- Considerem-se n pares de observações (x_i, y_i), $i = 1, ..., n$.
  Pretende-se determinar a recta $y = b_0 + b_1 x$ que melhor se ajusta às n observações de acordo com o critério dos mínimos quadrados.

  Sejam $\hat{y}_i = b_0 + b_1 x_i$ os y_i ajustados pela recta.

  O erro ou resíduo é dado por $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$.

  Pretende-se determinar b_0 e b_1 que minimizem a soma dos quadrados dos resíduos.

Regressão linear simples – exemplo



Regressão linear simples – contexto descritivo

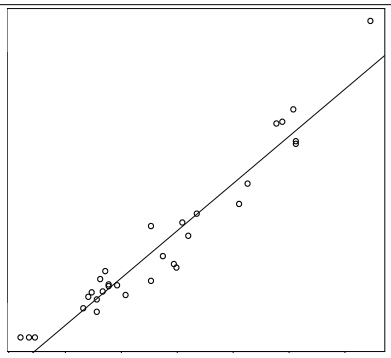
Seja $SQRE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$

Determinar b_0 e b_1 que minimizem $SQRE$ $\Leftrightarrow$

$$\begin{cases} b_1 = \dfrac{cov(x,y)}{s_x^2} = r\dfrac{s_y}{s_x} \\ b_0 = \bar{y} - b_1\bar{x} \end{cases}$$

Nota: b_1 tem o sinal de $cov(x, y)$ e r.

Regressão linear simples – exemplo



$y = -1.04624 + 0.05648\,x$

Regressão linear simples – contexto descritivo

- A recta de regressão passa no ponto $(\bar{x}, \bar{y})$
- O declive da recta b_1 :
  - chama-se **coeficiente de regressão de *y* sobre *x***,
  - representa a variação esperada para *y* quando *x* aumenta 1 unidade.
- **Precisão da recta:**
  Uma medida da precisão da recta de regressão é dada pelo **coeficiente de determinação** $R^2 = (r)^2$ que mede a percentagem de variabilidade de *y* que é explicada pela regressão.
  - Exemplo: $R^2 = 0.9671194^2 = 0.9353199$

Bibliografia

- Murteira, B.J.F. (1993) Análise Exploratória de Dados: Estatística Descritiva, McGraw-Hill, Lisboa.
- Murteira, B., Ribeiro, C.S., Silva, J.A. e Pimenta C. (2010). Introdução à Estatística. Escolar Editora.
- Neves, M. (2009) Introdução à Estatística e Probabilidade. Apontamentos de apoio às aulas.
  - Estatística Descritiva: *www.isa.utl.pt/dm/estat/estat/seb1.pdf*
- Pestana, D.D. e Velosa, S.F. (2002) Introdução à Probabilidade e à Estatística, vol.I, Fundação Caloustre Gulbenkian, Lisboa.