Estatística (2011/2012)

Slides de apoio às Aulas

Página *web* da U.C *Estatística*
http://www.isa.utl.pt/dm/estat/estat/estat.html

Endereço para descarregar o *software* R
http://cran.r-project.org/
(ver mais indicações na página da U.C)

Estatística (2011/2012)

Docentes:

Manuela Neves (manela@isa.utl.pt)(coordenadora)
Maria Emília Pinto (mila@isa.utl.pt)
Maria João Martins (mjmartins@isa.utl.pt)
Marta Mesquita (martaoliv@isa.utl.pt)
Mariana Mota (mariana@isa.utl.pt)

O que é a Estatística ?

É a ciência que se ocupa da recolha e tratamento de informação, i.e., da obtenção de uma amostra, sua descrição e interpretação e, com apoio da teoria da probabilidade efectuar inferências para a população e previsões da evolução futura do fenómeno em estudo.

Principais tópicos da Unidade Curricular e Calendarização:

- Estatística Descritiva. A Regressão Linear Simples. Introdução ao *software* R (3 semanas)
- Introdução aos Modelos Probabilísticos (6 semanas)
- Introdução à Inferência Estatística – intervalos de confiança e testes de hipóteses (5 semanas)

Objectivos de cada capítulo

Estatística Descritiva:
– sumariar e descrever os aspectos relevantes num conjunto de dados. Recurso a tabelas, gráficos e indicadores numéricos. Introdução à regressão linear simples.

Teoria da Probabilidade:
— apresentar os modelos mais usuais de fenómenos naturais nos quais se supõe intervir o acaso - fenómenos aleatórios.

Inferência Estatística:
— tirar conclusões para a população a partir do estudo da amostra;
– tomar decisões quanto ao(s) valor(es) de características importantes da população de onde foi retirada a amostra.

Veja-se o seguinte esquema...

Referências Bibliográficas

Neves, M. M. (2008) - *Introdução à Estatística e à Probabilidade.* Apontamentos de Apoio à Unidade Curricular.

—— (2011) - *Folhas de exercícios de apoio às aulas práticas* com algumas resoluções.

—— (2011) - *Colectânea de exames com algumas resoluções*. Edição da AEISA.

Murteira, B.; Ribeiro, C.S.; Silva, J.A. e Pimenta, C.(2002)- *Introdução à Estatística*, Mc Graw Hill - **cota Bisa - U10-681** (existe edição de 2008).

Murteira, Bento (1993) - *Análise exploratória de dados. Estatística Descritiva*. Mc Graw-Hill -**cota Bisa - U10-401**.

Pestana, D.D. e Velosa, S.F. (2002)- *Introdução à Probabilidade e à Estatística* . Fundação Calouste Gulbenkian - **cota Bisa - U10-677**(existe edição de 2008).

Daniel W. W. (1991)- *Biostatistics: A Foundation for analysis in the Health Sciences*. John-Wiley & Sons - **cota Bisa - U10-481**.

...Antes de começar!!!...

- Os capítulos **I - Estatística Descritiva** e **II - Introdução à Teoria da Probabilidade** iniciam-se com assuntos que são leccionados na disciplina *Matemática* do Ensino Secundário.

  Essa parte inicial contém matéria de revisão. Os *slides* são preparados com o objectivo de facilitar essa revisão, bem como uniformizar os símbolos e notações que iremos usar.

- A unidade curricular ***Estatística*** é leccionada no 3º semestre comum a todas as licenciaturas do ISA (com excepção de Arquitectura Paisagista).

  O acompanhamento adequado dos assuntos que trataremos necessita que os alunos tenham adquirido formação em Cálculo e Análise, especificamente tenham conhecimentos de:

...Antes de começar!!!...

- sucessões, funções reais de variáveis reais, diferenciabilidade, primitivação e cálculo integral em $\mathbb{R}$ e em $\mathbb{R}^2$;
- embora muitos resultados da teoria da probabilidade e da estatística necessitem de conceitos de séries numéricas e séries de funções, a sua utilização será omitida sempre que fosse necessária para a dedução de resultados, atendendo a que esta matéria não foi leccionada nas unidades curriculares Matemática e Informática e Álgebra Linear e Análise Matemática.

  Ainda assim optámos por incluir a utilização deste tópico nos apontamentos teóricos preparados para apoio à unidade curricular.

...Antes de começar!!!...

Várias unidades curriculares dos actuais planos de licenciatura (1º Ciclo) do ISA e dos mestrados (2º Ciclo) utilizam os conhecimentos leccionados nesta unidade curricular quer como conceitos base, quer para o tratamento das suas aplicações.

Queremos, por isso, deixar aqui um alerta aos nossos alunos – existindo no plano curricular do 1º Ciclo apenas esta unidade curricular de Probabilidade e Estatística, é fundamental cumprir-se o programa proposto.
Tal exige de alunos e professores um trabalho sistemático e coerente que tem que se iniciar no 1º dia de aulas.
Os assuntos considerados revisão constituem trabalho individual, sob pena de não se conseguir atingir as metas propostas.

Objectivos da Estatística Descritiva:

- condensar, sob a forma de tabelas, os dados observados;
- fazer a representação gráfica;
- calcular indicadores de localização e de dispersão.

Conceitos básicos em Estatística (definição e um exemplo):

- **população ou universo** $\longrightarrow$ conjunto de todos os elementos que têm uma característica de interesse em comum (ex: todas as árvores de uma dada espécie)
- **unidades estatísticas** $\longrightarrow$ são os elementos da população (ex: as árvores)
- **variável** $\longrightarrow$ característica de interesse (ex: $X \rightarrow$ altura de árvores de uma espécie e $x \rightarrow$ altura observada de uma árvore).
- **amostra** $\longrightarrow$ subconjunto da população, efectivamente observado.

Estatística descritiva a uma dimensão

Ao(s) valor(es) da(s) característica(s) de interesse observadas nos elementos da amostra costuma chamar-se **dado(s)**.

Os **dados** podem ser de natureza

- **quantitativa** $\longrightarrow$ **discreta** (contagens - n$^{\underline{o}}$ de peras em cada pereira, n$^{\underline{o}}$ de machos por ninhada de coelhos) ou
  $\longrightarrow$ **contínua** (peso, comprimento, altura, tempo)
- **qualitativa** $\longrightarrow$ **nominal** (sexo de um indivíduo, categoria taxonómica de uma espécie) ou
  $\longrightarrow$ **ordinal** (avaliação numa escala de A (óptima) a E (péssima) da qualidade do almoço numa cantina)

Exemplo 1.
Num estudo para analisar a taxa de germinação de um certo tipo de cereal foram semeadas cinco sementes em cada um de 50 vasos iguais com o mesmo tipo de solo.
O número de sementes germinadas em cada vaso está registado a seguir:

$$\begin{array}{ccccccccccccccc}
1 & 0 & 1 & 2 & 1 & 3 & 2 & 0 & 0 & 1 & 4 & 0 & 2 & 1 & 0 \\
2 & 4 & 1 & 2 & 0 & 3 & 5 & 3 & 0 & 2 & 1 & 3 & 3 & 0 & 4 \\
0 & 2 & 5 & 3 & 0 & 2 & 5 & 1 & 1 & 0 & 4 & 4 & 1 & 2 & 1 \\
0 & 5 & 1 & 2 & 3
\end{array}$$

Neste caso os **dados são de natureza discreta, com um número pequeno de valores distintos**.
Dados deste tipo podem ser condensados numa tabela da forma

Descrição dos dados por tabelas

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	12	0.24	0.24
1	12	0.24	0.48
2	10	0.20	0.68
3	7	0.14	0.82
4	5	0.10	0.92
5	4	0.08	1

$x_i \longrightarrow$ n º de sementes germinadas;
$n_i \longrightarrow$ frequência absoluta;
$f_i = \frac{n_i}{n} \longrightarrow$ frequência relativa;
$F_i \longrightarrow$ frequência relativa acumulada

Exemplo 2.

Um dos principais indicadores da poluição atmosférica nas grandes cidades é a concentração de ozono na atmosfera. Num dado Verão registou-se 78 valores dessa concentração (em $\mu g/m^3$), numa dada cidade:

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						

Agora estamos em presença de dados de **natureza contínua**

Descrição dos dados por tabelas

Para dados de natureza contínua - como é este caso - (ou quando temos dados de natureza discreta com um elevado número de valores distintos) elabora-se a **tabela de frequências** procedendo assim:

- Determina-se $\max(x_i)$ e $\min(x_i)$,

  $\max(x_i) - \min(x_i) \longrightarrow$ **amplitude total**.

- Escolhe-se um número de subintervalos $\longrightarrow$ **classes**

- Para cada classe calcula-se a **frequência absoluta**, n_i

  e a **frequência relativa**, f_i

Exemplo de uma regra para escolha do número de classes:
Regra de Sturges $\longrightarrow$ toma-se como número de classes
o inteiro m mais próximo de $\quad \mathbf{1 + (\log_2 n) = 1 + \frac{\log_{10} n}{\log_{10} 2}}$

Descrição dos dados por tabelas

Voltemos ao exemplo: $min(x_i) = 1.1 \qquad max(x_i) = 11.7$
Pela regra de Sturges $m \approx 7.285 \longrightarrow$ considere-se $m = 7$
amplitude das classes $h = 1.51 \longrightarrow$ considere-se $h = 1.5$
(veremos que, com esta escolha, será necessário considerar 8 classes para se incluírem todas as observações)

Uma tabela de frequências possível é:

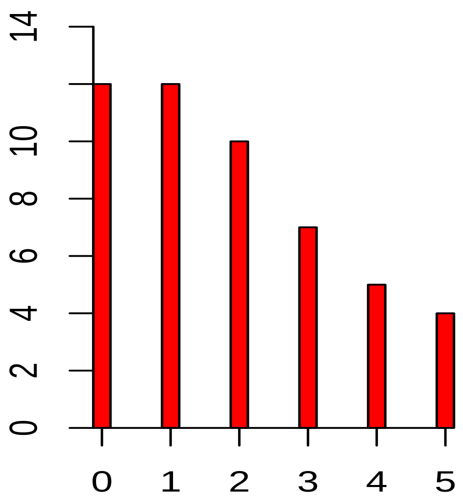
c_i	$x_i^{'}$	n_i	f_i	F_i
$]1.0, 2.5]$	1.75	10	0.128	0.128
$]2.5, 4.0]$	3.25	16	0.205	0.333
$]4.0, 5.5]$	4.75	18	0.231	0.564
$]5.5, 7.0]$	6.25	26	0.333	0.897
$]7.0, 8.5]$	7.75	5	0.064	0.962
$]8.5, 10.0]$	9.25	2	0.026	0.987
$]10.0, 11.5]$	10.75	0	0.00	0.987
$]11.5, 13.0]$	12.25	1	0.013	1

$x_i^{'} \longrightarrow$ ponto médio da classe c_i

Métodos gráficos

Métodos gráficos usados para representar um conjunto de dados $\longrightarrow$ dois dos principais são:

- **o diagrama de barras** $\longrightarrow$ para dados de natureza discreta, com um número pequeno de valores distintos e
- **o histograma** $\longrightarrow$ para dados de natureza contínua, ou quando o n$^{\underline{o}}$ de valores distintos é muito elevado.



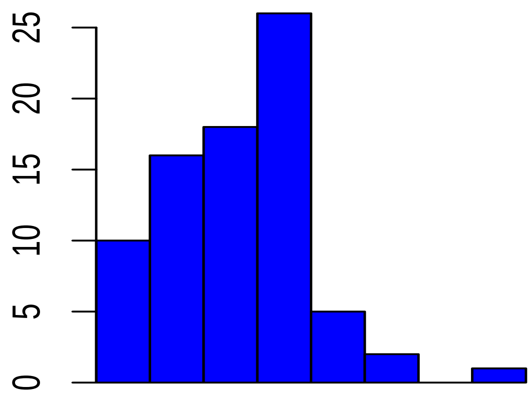


Diagrama de barras (exemplo 1) e histograma (exemplo 2) das frequências absolutas

Indicadores numéricos

As tabelas e gráficos constituem um primeiro conjunto de ferramentas usadas pela Estatística Descritiva para resumir e descrever um conjunto de dados

Outro conjunto de ferramentas que permite caracterizar um conjunto de dados é constituído pelos indicadores numéricos também chamados indicadores amostrais. Falaremos nas:

- medidas de localização e
- medidas de dispersão.

Medidas de localização que iremos estudar:
média, **mediana**, **quantis** e **moda**

A média. Propriedades

Considere-se $x_1, x_2, \cdots, x_n$, uma amostra de n observações.

Definição Chama-se **média aritmética**, **média empírica** ou simplesmente **média** e representa-se por $\overline{x}$ a

$$\boxed{\overline{x} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}}$$

Propriedades da média

- Sejam $x_1, x_2, ..., x_n$ observações cuja média é $\overline{x}$ e considere-se $y_i = a + bx_i, \quad {\scriptstyle i = 1, ..., n}$.
  As observações transformadas $y_1, y_2, ..., y_n$ têm média $\overline{y} = a + b\overline{x}$.
- Se $x_1, ..., x_n$ são n observações de média $\overline{x}$ e $y_1, ..., y_m$ são m observações de média $\overline{y}$,
  a média das $n + m$ observações é dada por $\dfrac{n\,\overline{x} + m\,\overline{y}}{n+m}$.

A mediana e a moda

Definição A **mediana** é o valor que divide a amostra ordenada em duas partes iguais (i.e., com o mesmo número de observações cada).

Dada a amostra $x_1, ..., x_n$, seja $x_{(1)} \leq ... \leq x_{(n)}$ a amostra ordenada.
A **mediana** é dada por:

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

Definição A **moda,** $\boldsymbol{mo}$**,** é a observação mais frequente (se existir).

Caso discreto $\rightarrow$ é a observação que tem maior frequência.
Caso contínuo $\rightarrow$ só faz sentido definir-se sobre dados agrupados $\rightarrow$ é um valor da classe que tem maior frequência (ver medidas para dados agrupados)

Os quantis empíricos

Se considerarmos a amostra ordenada dividida em quatro partes, cada uma com o mesmo número de observações, os pontos da divisão chamam-se **quartis empíricos** ou apenas **quartis** e costumam representar-se por Q_1, Q_2 e Q_3. É claro que $Q_2 \equiv \tilde{x}$.

Generalização do conceito de quartil

Definição Chama-se **quantil de ordem $\boldsymbol{\theta}$**, $(0 \leq \theta \leq 1)$, o valor $\boldsymbol{Q^*_\theta}$ tal que há uma proporção θ de observações inferiores ou iguais a $\boldsymbol{Q^*_\theta}$ e uma proporção $(1-\theta)$ de observações maiores ou iguais a esse valor. Uma fórmula de cálculo pode ser

$$\boxed{\mathrm{Q}^*_\theta = \begin{cases} \dfrac{x_{(n\ \theta)} + x_{(n\ \theta+1)}}{2} & \text{se } n\ \theta \text{ inteiro} \\ x_{([n\ \theta]+1)} & \text{se } n\ \theta \text{ não inteiro} \end{cases}}$$

onde $[n\ \theta]$ designa o maior inteiro contido em $n\ \theta$.

Nota: $Q^*_{0.25} \equiv Q_1$; $Q^*_{0.5} \equiv Q_2$ e $Q^*_{0.75} \equiv Q_3$

Medidas de localização – dados agrupados

Dados agrupados em c $(c < n)$ classes (ou grupos). Sejam
$x_1^{'}, x_2^{'}, ..., x_c^{'}$ pontos médios de cada classe (ou valores de cada grupo);
$n_1, n_2, ..., n_c$ as frequências absolutas de cada classe (ou grupo)

Média agrupada $= \tilde{\bar{x}} = \frac{n_1 x_1^{'} + n_2 x_2^{'} + \cdots + n_c x_c^{'}}{n} = \frac{\sum_{i=1}^{c} n_i \; x_i^{'}}{n}$

Moda amostral para **dados agrupados:**
1º determina-se a **classe modal** $\rightarrow$ classe com maior frequência.
2º de várias fórmulas que existem, vamos aqui considerar:

$$\boxed{\boldsymbol{mo} \simeq x_k^{min} + h \, \frac{f_{k+1}}{f_{k-1} + f_{k+1}}}$$

sendo k a classe modal; f_{k-1} e f_{k+1} a frequência relativa da classe anterior e posterior à classe modal, respectivamente, x_k^{min} limite inferior da classe k e h amplitude da classe k.

Quantil de ordem θ :

- Identifica-se a primeira classe cuja frequência relativa acumulada seja superior ou igual a $\theta \longrightarrow$ seja k essa classe e F_k a frequência relativa acumulada correspondente.

- Uma das fórmulas usadas para determinar o quantil de ordem θ é:

$$\boxed{\boldsymbol{Q^*_\theta} \simeq x_k^{min} + h\ \frac{\theta - F_{k-1}}{f_k}}$$

com $F_{k-1} \longrightarrow$ frequência relativa acumulada da classe anterior à classe k

Nota: A **mediana** para dados agrupados obtém-se considerando na fórmula acima $\boldsymbol{\theta = 0.5}$.

Indicadores de dispersão

- **Amplitude total** $A_{tot} = max(x_i) - min(x_i)$
- **Amplitude inter-quartil** $AIQ = Q_3 - Q_1.$
- **Variância**[1] $$\boxed{s_x^2 = s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$
- **Desvio padrão** $s_x = s = \sqrt{\text{Variância}}$

Outra fórmula de cálculo da variância: $s^2 = \dfrac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)}$

Uma medida de dispersão relativa (as acabadas de indicar são medidas de dispersão absolutas) é o **coeficiente de variação** que só se calcula quando as observações têm todas o mesmo sinal. Permite a comparação entre distribuições e define-se como :

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

[1] Vamos considerar esta definição de variância

Variância e desvio padrão

Propriedades

- $s_x^2 \geq 0$
- Sejam $x_1, ..., x_n$, observações com variância s_x^2 considere-se $y_i = a + bx_i, \qquad i = 1, ..., n$.
  As observações transformadas têm como variância $s_y^2 = b^2\ s_x^2$.

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

Dados agrupados em c classes - a variância calcula-se:

$$\frac{\sum_{i=1}^{c} n_i {x'_i}^2}{n} - \overline{x}^2$$

A caixa de bigodes

Um modo gráfico que permite facilmente interpretar a localização e a dispersão de um conjunto de dados, efectuando em simultâneo a sua síntese $\longrightarrow$ **o diagrama de extremos e quartis**.

Se nesse gráfico identificarmos as observações que se afastam do padrão geral dos dados (candidatos a *outliers*) é hábito designá-lo por **caixa de bigodes**. Existem vários critérios para classificar uma observação como um *outlier*, vamos considerar:

Definição Um valor x_i é um candidato a **outlier** se

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

sendo B_I **barreira inferior** e B_S **barreira superior** definidas como:

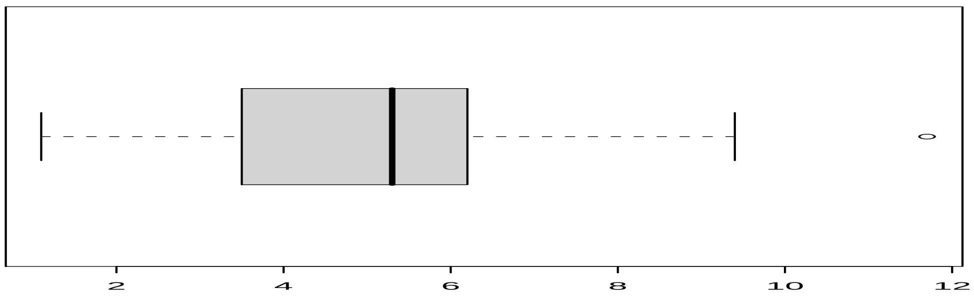
$$B_I = Q_1 - 1.5(Q_3 - Q_1) \qquad B_S = Q_3 + 1.5(Q_3 - Q_1)$$

A caixa de bigodes

Como desenhar **uma caixa de bigodes?**

- Marcar **o valor adjacente inferior** $\longrightarrow$ é o **menor** valor do conjunto dos dados (podendo ser o *mínimo*) maior ou igual à barreira inferior;
- Marcar **o valor adjacente superior** $\longrightarrow$ é o **maior** valor do conjunto dos dados (podendo ser o *máximo*) menor ou igual à barreira superior.
- Marcar **a mediana, primeiro e terceiro quartis** (que vão permitir desenhar uma "caixa") e marcar os candidatos a **"outliers"**

Exemplo Caixa de bigodes referente os dados do exemplo 2.



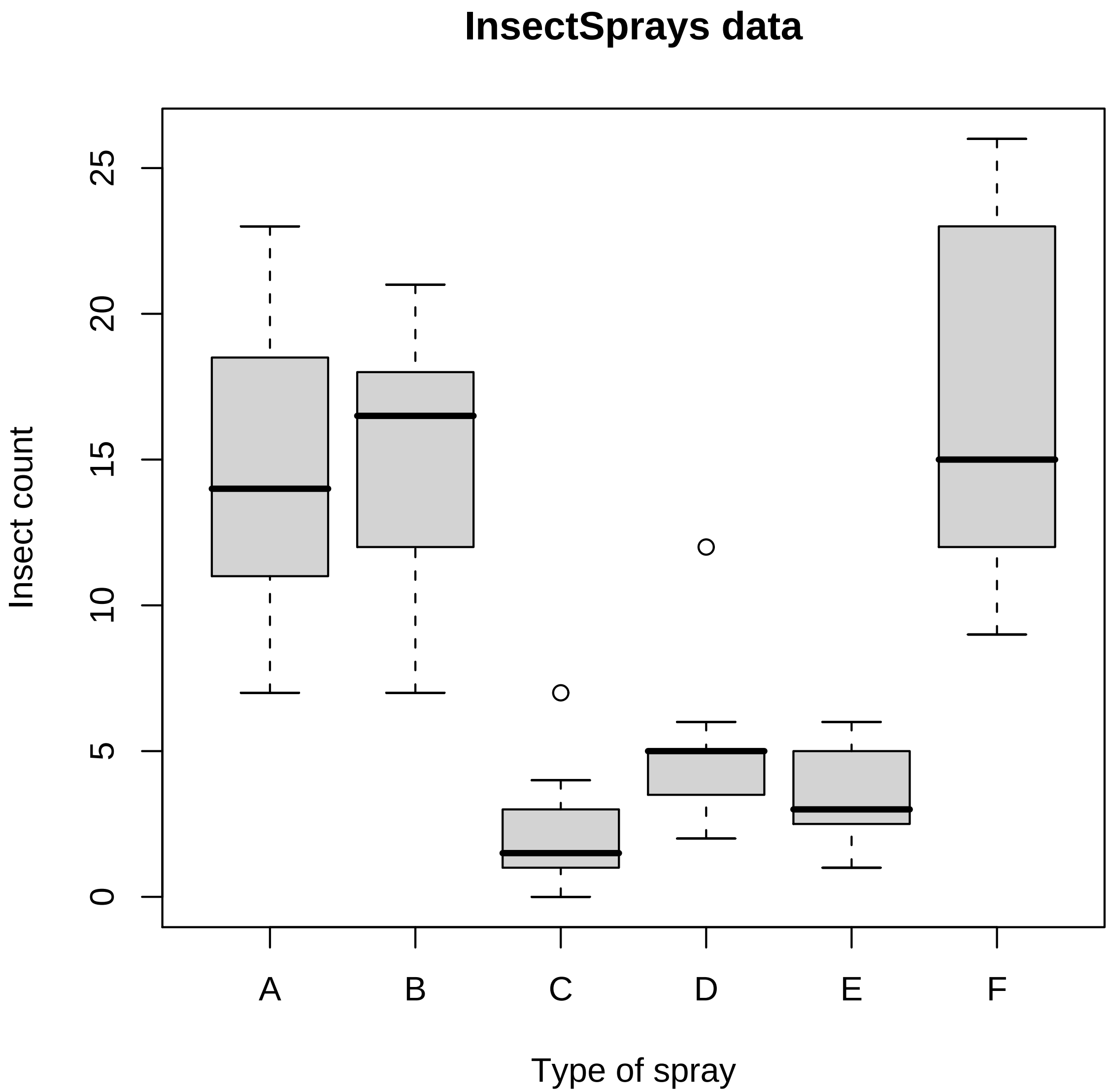
Caixas de bigodes paralelas

Quando se pretende comparar várias amostras, o recurso a caixas de bigodes paralelas é uma ferramenta muito útil, permitindo de forma fácil obter uma primeira interpretação e comparação dos conjuntos de dados.

Exemplo As seguintes caixas de bigodes referem-se a um conjunto de dados `InsectsSprays` disponíveis no *package* `datasets` do R. São contagens de insectos em unidades agrícolas experimentais, às quais foram aplicados 6 tipos de insecticidas.

Referência: Beall, G., (1942) The Transformation of data from entomological field experiments, Biometrika, 29, 243;262.

Caixas de bigodes paralelas



Estatística descritiva a duas dimensões

Nas aulas anteriores, em cada unidade estatística, estudámos uma única variável. Muitas vezes, porém, interessa registar os valores de mais do que uma variável e procurar a existência de alguma relação entre as variáveis. Vamos tratar neste curso o caso de **duas variáveis** observadas na unidade estatística.

Exemplo Peso e altura de uma pessoa; Comprimento e largura das folhas de uma espécie vegetal, etc.

Consideremos o seguinte exemplo, retirado de *Estatística, Teoria e Métodos*, Pierre Dagnielie, 1º volume (1973).

Exemplo Foram registados os pesos das folhas e das raízes de 1000 pés de *Cichorium intybus*, sendo os valores dos pesos das folhas e das raízes agrupados em classes de 80 g e 40 g, respectivamente.

Exemplo (cont.)

Construíu-se então o seguinte **quadro de correlação**, **quadro de dupla entrada** ou **tabela de contingência**.

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

Objectivos Estudo em simultâneo de duas séries de observações, pondo em evidência “relações” existentes entre elas.

Não são relações determinísticas que interessam à Estatística, mas é o comportamento em média (relação estatística) das duas características.
Se duas variáveis estão ligadas por uma relação estatística diz-se haver **correlação** entre elas.

Correlação **positiva** se as duas características variam no mesmo sentido e **negativa** caso contrário.

Tabelas e representação gráfica

Sejam $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ observações efectuadas em n unidades estatísticas.

Para o estudo das características e pesquisa de existência de relação entre as variáveis:
elaboração de tabelas; representação gráfica e cálculo de indicadores.

- Se n é grande é útil considerar uma **tabela de contingência** (como no exemplo do wideslide 29).
- Se n não for muito elevado, as observações podem representar-se graficamente num **diagrama de dispersão** (*scatterplot*) ou **nuvem de pontos** (aqui cada par observado (x_i, y_i) é marcado num sistema de eixos cartesianos).

Tabela de contingência

$$
\begin{array}{c|cccc|c}
 & y_1 & y_2 & \dots & y_q & \\
\hline
x_1 & n_{11} & n_{12} & \dots & n_{1q} & n_{1.} \\
x_2 & n_{21} & n_{22} & \dots & n_{2q} & n_{2.} \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
x_p & n_{p1} & n_{p2} & \dots & n_{pq} & n_{p.} \\
\hline
 & n_{.1} & n_{.2} & \dots & n_{.q} & n
\end{array}
$$

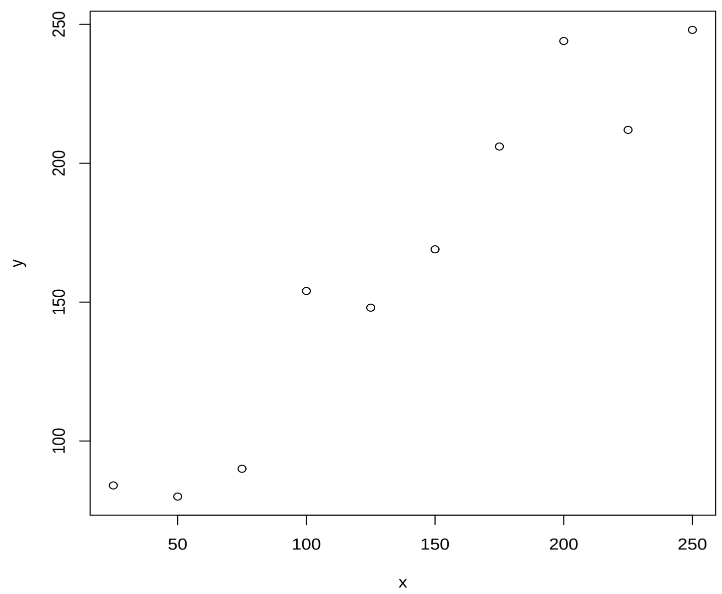
$\boldsymbol{n_{ij}} \longrightarrow$ número de indivíduos para os quais foi observado o par (x_i, y_j).

$\boldsymbol{n_{i.} = \sum_{j=1}^{q} n_{ij}}$ e $\boldsymbol{n_{.j} = \sum_{i=1}^{p} n_{ij}}$ **frequências marginais** de x e y, respectivamente.

Nuvem de pontos

Exemplo Pretende-se estudar o efeito da aplicação de diferentes quantidades de um dado fertilizante (x) na produção de relva (y). A relva é semeada uniformemente numa dada área na qual são marcados ao acaso 10 talhões de 1 m^2, a cada um dos quais é aplicada uma certa quantidade de fertilizante. A relva é depois cortada, seca e pesada sendo os dados obtidos e a **nuvem de pontos** correspondente:

x (g/m^2)	y (g/m^2)
25	84
50	80
75	90
100	154
125	148
150	169
175	206
200	244
225	212
250	248



Indicadores numéricos

Médias marginais de x e y, respectivamente, são

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

$(\overline{x}, \overline{y}) \longrightarrow$ centro de gravidade da nuvem de pontos.

Dispersões marginais de x e y, respectivamente

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

Mas... há uma medida que dá informação sobre as duas variáveis em simultâneo.

Definição Dadas as variáveis x e y, chama-se **covariância de x e y** a

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

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

Propriedades da covariância

1. Seja (x_i, y_i) uma série de n observações e considere-se:

$$x_i^{'} = a + bx_i \qquad y_i^{'} = c + dy_i.$$

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

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

Nota 1

Importância da covariância $\rightarrow cov(x, y) > 0$ – há correlação positiva; $cov(x, y) < 0$ – há correlação negativa.

Desvantagem da covariância $\rightarrow$ fortemente afectada por mudanças de escala nas observações (ver propriedade 1.)

Nota 2 $\quad |cov(x, y)| = s_x s_y \Longleftrightarrow (y_i - \overline{y}) - b(x_i - \overline{x}) = 0 \qquad \forall i$

portanto, se $|cov(x, y)| = s_x s_y$ todos os pontos observados se encontram sobre uma recta definida como $y - \overline{y} = b(x - \overline{x})$

O coeficiente de correlação. Propriedades

Definição O **coeficiente de correlação** é definido como

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

Propriedades do coeficiente de correlação

1. $\boldsymbol{r}$ tem sempre o mesmo sinal da covariância;
2. $\boldsymbol{-1 \leq r \leq 1}$; (se $|r_{xy}| = 1$ todos os valores observados se encontram sobre uma recta).
3. Se (x, y) têm coeficiente de correlação r_{xy} e

   $x_i^{'} = a + bx_i \quad \text{e} \quad y_i^{'} = c + dy_i \quad (bd > 0)$, tem-se

   $$r_{x^{'}y^{'}} = r_{xy} \quad \text{se} \quad (bd > 0)$$
   $$r_{x^{'}y^{'}} = -r_{xy} \quad \text{se} \quad (bd < 0)$$

Então **o coeficiente de correlação não é afectado, em valor absoluto, por transformações lineares.**

O coeficiente de correlação. Interpretação

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

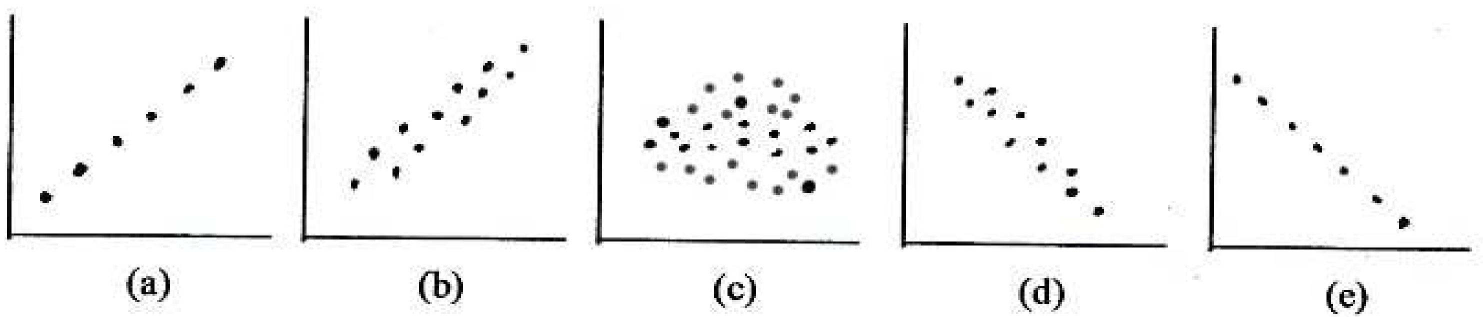
(b) $r \simeq 1$ todos os pontos observados se encontram próximos de uma recta de declive positivo.

(c) $r \simeq 0$ a nuvem apresenta um aspecto arredondado ou alongado segundo um dos eixos.

(d) $r \simeq -1$ todos os pontos observados se encontram próximos de uma recta de declive negativo.

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

Nota: O coeficiente de correlação mede *a nitidez da ligação* existente entre duas variáveis, quando essa ligação é linear ou aproximadamente linear



A regressão linear simples

Se $|r| \simeq 1$ e a nuvem de pontos sugere a existência de uma relação linear entre os valores observados.

Faz sentido determinar a equação de uma recta que possa traduzir bem a relação observada, i.e., pretende-se determinar $\boldsymbol{y = a + bx}$
$\longrightarrow$ **recta de regressão**, que permita:

- descrever a relação entre y (variável resposta ou dependente) e x (variável explicativa ou independente);
- prever um valor de y para um dado valor de x.

Mas ... a equação $y = a + bx$ não é verificada para todos os pares (x_i, y_i) (note-se que só o seria se $|cov(x, y)| = s_x s_y$)

Na verdade para cada par (x_i, y_i) tem-se $\boldsymbol{y_i = a + b\, x_i + \ e_i}$

A regressão linear simples

A $a + b\,x_i$ designe-se por $\hat{y}_i$
são os valores de y estimados pela recta para cada x_i.

Então pode-se escrever $y_i = \hat{y}_i + e_i$

$e_i = y_i - \hat{y}_i$ são chamados **resíduos**.

Portanto $\longrightarrow$ obter a recta $\Longleftrightarrow$ determinar a e b.
Método usado $\longrightarrow$ **método dos mínimos quadrados** $\longrightarrow$ a e b são determinados de modo a minimizar a soma dos quadrados dos resíduos ou seja, minimizar

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

A regressão linear simples

Pretende-se então determinar os minimizantes de uma função de duas variáveis. As condições de estacionaridade são:

$$\begin{cases} \frac{\partial Q}{\partial a} = 0 \\ \frac{\partial Q}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{cases} 2 \sum (y_i - a - b\, x_i) = 0 \\ 2 \sum x_i\, (y_i - a - b\, x_i) = 0 \end{cases}$$

A estas equações chama-se **equações normais**
Algumas conclusões podem ser tiradas destas equações:

- $\sum (y_i - a - b\, x_i) = 0 \Rightarrow \sum (y_i - \hat{y}_i) = \sum \boldsymbol{e_i = 0}$ a soma dos resíduos é nula.
- $\sum (y_i - \hat{y}_i) = 0 \Rightarrow \boldsymbol{\overline{y} = \overline{\hat{y}}}$ a média dos valores observados é igual à média dos valores estimados.

A regressão linear simples

- a recta de regressão passa no ponto $(\overline{x}, \overline{y})$.
- **Solução do sistema**

$$\boldsymbol{b} = \frac{n\sum x_i\ y_i - \sum x_i\ \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{y_i} - \overline{\boldsymbol{y}})}{\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2} = \frac{cov(x,y)}{s_x^2} = r\ \frac{s_y}{s_x}$$

$$\boldsymbol{a = \overline{y} - b\overline{x}}$$

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

Observações:

- b tem o mesmo sinal que $cov(x, y)$ e r.
- Dado x_i e sendo $x_i' = x_i + 1$ tem-se
  $\hat{y}_i = a + b\ x_i \qquad \hat{y_i'} = a + b\ (x_i + 1)$.

  $\boldsymbol{b = \hat{y_i'} - \hat{y}_i,} \longrightarrow$ $\boldsymbol{b}$ representa a variação esperada para y quando x aumenta uma unidade.

Precisão da recta de regressão

Um dos objectivos da recta de regressão é o de **predizer** o valor de uma variável conhecendo o valor assumido pela outra **mas** é necessário avaliar o **grau de precisão** atingido pelas estimativas.

O método dos mínimos quadrados permite uma importante decomposição de $\sum(y_i - \overline{y})^2$.

$\sum(y_i - \overline{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2$ cujas parcelas se costuma representar por

$\boldsymbol{SQ_T} \quad = \quad \boldsymbol{SQ_{RE}} \quad + \quad \boldsymbol{SQ_R}$, isto é:

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

O coeficiente de determinação

Vamos designar por

$R^2 = \frac{SQ_R}{SQ_T}$ **a percentagem de variabilidade "explicada" pela regressão**

A $\boldsymbol{R^2}$ chama-se **coeficiente de determinação** $\longrightarrow$ e é uma **medida de precisão** da recta de regressão.

Observe-se que no contexto que estamos a considerar - a regressão linear simples se tem

$$\boldsymbol{R^2} = \frac{SQ_R}{SQ_T} = \frac{b^2 \sum(x_i - \overline{x})^2}{\sum(y_i - \overline{y})^2} = \frac{cov^2(x,y)}{s_x^4} . \frac{s_x^2}{s_y^2} = \frac{cov^2(x,y)}{s_x^2 \; s_y^2} = \boldsymbol{r^2}.$$

Últimas notas ...

Tratámos aqui a regressão linear simples como uma técnica descritiva.
Em Estatística e Delineamento voltar-se-á a abordar a regressão mas em contexto inferencial.

Nessa altura é necessário recorrer a modelos de probabilidade o que exige admitir certos pressupostos. O gráfico dos resíduos, e_i, v.s. os valores ajustados, $\hat{y}_i$, constitui uma ferramenta essencial na validação desses pressupostos.
Por exemplo, nesse gráfico :

– não deve existir qualquer padrão aparente;
– não deve verificar-se um aspecto de "funil";
– a existência de um ou mais resíduos destacados, alerta para a ocorrência de observações que estejam a afectar o ajustamento;
– ...