



Hinc patram sustinet

Instituto Superior de Agronomia
Universidade Técnica de Lisboa

ESTATÍSTICA

Ano Lectivo - 2011/2012

Aulas práticas


Exercícios com algumas soluções

Nota Introdutória

Apresentamos aqui um conjunto de exercícios preparados para as aulas práticas da unidade curricular *Estatística*, leccionada no 3.^o semestre de todas as licenciaturas de Bolonha do ISA, com excepção de Arquitectura Paisagista.



É por demais reconhecida a importância crescente da Estatística como ferramenta imprescindível na análise, interpretação e previsão de resultados em todas as áreas no domínio das Ciências e Engenharias, nomeadamente as que constituem a formação disponibilizada pelo Instituto Superior de Agronomia

Na reestruturação actual esta unidade curricular passa a ter 5h/semana sendo 2h teóricas e 3h práticas (até 2009/2010 dispunha de 2h teóricas e 4h práticas). Tem como principais objectivos:

- Apresentar os conceitos básicos de Probabilidade (com os principais modelos probabilísticos, imprescindíveis para a modelação e inferência de características de interesse e que estão sujeitas à intervenção do acaso - aleatórias).
- Ensinar métodos de interpretação, descrição, análise e inferência realizados sobre um conjunto de observações obtidas em trabalhos de campo nas diferentes unidades curriculares dos planos curriculares do ISA.
- Introduzir um *software* estatístico adequado, que possua técnicas actualizadas para os tratamentos estatísticos que os alunos necessitem de realizar. Muitos são os programas informáticos existentes. Entendemos, porém, que ao introduzir os nossos alunos no *software*  estamos a colocá-los na vanguarda do que hoje é usado quer nas aplicações quer na investigação.

Estes exercícios integram três capítulos: Estatística Descritiva, Introdução à Teoria da Probabilidade e Introdução à Inferência Estatística.

No final de cada capítulo de exercícios encontram-se tópicos de resolução ou algumas soluções.

Para preparar desde já os alunos na introdução e utilização do *software* estatístico que vamos usar, na 1.^a parte destas folhas de exercícios, dedicadas à Estatística Descritiva, apresentamos **uma lista de exercícios de treino do **. Alguns exercícios de Estatística Descritiva serão também resolvidos com apoio do .


Como material de apoio sobre o  pode utilizar os textos referenciados na página *web* da *Estatística* (<http://www.isa.utl.pt/dm/estat/estat/estat.html>).

Antes de cada aula prática os alunos deverão tentar resolver os exercícios da matéria já leccionada, principalmente os indicados pelos professores.

Capítulo 1 Estatística Descritiva


Nota: Todos os ficheiros e/ou objectos referidos nos exercícios encontram-se na página *web* da *Estatística* (<http://www.isa.utl.pt/dm/estat/estat/estat.html>).

Exercícios de Introdução ao *software* com algumas aplicações em Estatística

Os exercícios seguintes ilustram e fazem a aplicação da breve introdução ao *software*  feita na primeira semana de aulas. Este *software* será utilizado ao longo de todo o semestre, como apoio na resolução de exercícios.

- 1.1. É desencadeado um programa de controlo da poluição de um rio em que são efectuadas medições antes de lançar a campanha antipoluição (Ano0) e um ano após (Ano1). Os resultados destas medições são os seguintes:

Ponto de controlo	1	2	3	4	5	6	7	8	9	10
Ano0	68	88	101	82	96	74	65	74	52	99
Ano1	67	87	90	76	98	69	68	65	59	70

- a) Usando a linguagem  crie dois vectores `Ano0` e `Ano1`, com os dados observados.
- b) Verifique que os dois vectores têm o mesmo número de registos.
- c) Crie o vector `dif` com as diferenças entre os valores de `Ano0` e `Ano1`.
- d) Crie o vector `posdif` com as componentes positivas do vector `dif`.
- 1.2. a) Crie os vectores `nomes`, `idades` e `alturas`, com os nomes e as respectivas idades e alturas de dez dos seus colegas.
- b) Construa o vector lógico `cartac` em que cada componente indica se o aluno da componente homóloga do vector `nomes` possui ou não carta de condução.
- c) Determine o máximo do vector `idades` e o mínimo do vector `alturas`.
- d) Determine a média das alturas e identifique os alunos que têm altura superior a este valor.
- e) Identifique os colegas com carta de condução e determine o seu número.
- f) Construa a *data frame* `colegas` que contenha os vectores `nomes`, `idades`, `alturas` e `cartac`. Resolva novamente as alíneas c), d) e e) utilizando este novo objecto.

- 1.3.** O ficheiro “concelho.txt” contém o nome de todos os concelhos de Portugal Continental, o respectivo número de freguesias e o nome do distrito a que pertencem.
- a) Leia os valores do ficheiro “concelho.txt” e guarde-os na *data frame* `conc`.
 - b) Qual ou quais os concelhos com maior número de freguesias?
 - c) Guarde no objecto `totc` o número total de concelhos existentes em Portugal Continental.
 - d) Crie a tabela `dist` com o nome de todos os distritos de Portugal Continental e o respectivo número de concelhos.
 - e) Construa um diagrama de barras com o número de concelhos por distrito de Portugal Continental.
 - f) Calcule o número médio de concelhos por distrito.
- 1.4.**
- a) Crie o vector `v` com o resultado de 30 lançamentos de um dado equilibrado de seis faces (sugestão: utilize a instrução `sample(1:6,30,rep=TRUE)`)
 - b) Calcule a média $\bar{v} = \frac{\sum_{i=1}^{30} v_i}{30}$ e a variância $s^2 = \frac{\sum_{i=1}^{30} (v_i - \bar{v})^2}{30}$ destes 30 valores.
 - c) Utilize os comandos `mean(v)` e `var(v)` para resolver a questão anterior. Comente os resultados obtidos.
 - d) Faça um diagrama de barras de `v`.
 - e) Crie os vectores `w1` e `w2` com, respectivamente, os resultados de 300 e 3000 lançamentos de um dado de seis faces equilibrado. Construa um diagrama de barras para os valores de cada um destes vectores e compare-os com o diagrama obtido na alínea anterior.
- 1.5.** Importe o ficheiro de objectos “SementesOzono.RData”. Este ficheiro contém os vectores `sementes` e `ozono`. O vector `sementes` contém o número de sementes de um certo cereal que germinaram em cada um de 50 vasos iguais (inicialmente foram semeadas 5 sementes em cada vaso). O vector `ozono` contém 78 valores de concentração de ozono na atmosfera (ppm) obtidos numa dada cidade.

Calcule o número médio de sementes que germinaram por vaso e a concentração média de ozono observada.

- 1.6.** Importe o ficheiro de objectos “DadosMeteo.RData”. Este ficheiro contém os vectores `precip`, `temp` e `vento`, com dados de precipitação (mm), temperatura do ar ($^{\circ}\text{C}$) e velocidade do vento (ms^{-1}), respectivamente, medidos numa estação meteorológica em Évora.
- a) Indique o número total de elementos e o número total de observações não disponíveis (NA) de cada vector.
 - b) Determine indicadores numéricos de localização e de dispersão para estes dados.

Exercícios de Estatística Descritiva a uma dimensão

Nota: Recorde-se que os ficheiros e/ou objectos referidos nos exercícios se encontram na página *web* da *Estatística* (<http://www.isa.utl.pt/dm/estat/estat/estat.html>).

1.7. Classifique, justificando, cada uma das seguintes variáveis quanto ao tipo: qualitativo/quantitativo, contínuas/discretas.

- a) Cor do cabelo de uma pessoa;
- b) Peso de um bebé à nascença;
- c) Número de automóveis que passaram na portagem nos domingos de verão;
- d) Qualidade da comida numa cantina (má, razoável, boa, muito boa);
- e) Temperatura máxima diária em Agosto deste ano;
- f) Número de sementes que germinaram num vaso onde foram semeadas 5 sementes.

1.8. Para cada um dos conjuntos de dados apresentados abaixo, construa uma tabela de frequências absolutas, relativas, relativas acumuladas e absolutas acumuladas.

- a) Conjunto 1 - - número de nemátodos contados em cada uma de 60 placas observadas ao microscópio.

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

- b) Conjunto 2 - - número de laranjas de cada uma das 40 árvores de um laranjal

131	136	150	152	155	156	162	169	170	177
188	196	201	201	205	210	210	211	214	216
217	220	225	226	231	238	240	244	244	247
251	262	268	275	288	297	300	302	303	305

- c) Conjunto 3 - - peso (kg) de 56 ovelhas, após administração de um dado tratamento. Estes dados encontram-se no ficheiro “ovelhas.txt”.

30.28	27.58	27.91	29.33	31.20	28.40	33.3	25.40
34.26	32.55	21.78	25.59	35.08	26.86	33.20	29.70
39.47	30.15	33.40	27.38	30.39	25.85	29.11	26.22
33.54	30.40	29.60	28.82	30.70	30.83	33.84	27.58
29.46	36.15	23.40	24.48	30.35	23.85	27.12	26.42
36.54	20.40	23.30	38.42	31.10	25.83	31.84	22.58
31.54	22.42	33.35	28.22	34.15	26.83	21.24	30.14

1.9. Considere os vectores $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) = (2, 5, 6, 10, 15)$ e $\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (-1, 2, 0, 3, 4)$. Calcule:

$$\begin{array}{lll} \text{a)} \sum_{i=1}^5 x_i & \text{b)} \sum_{i=3}^5 x_i & \text{c)} \sum_{k=1}^5 x_k \\ \text{d)} \sum_{j=1}^5 2x_j & \text{e)} \sum_{k=2}^5 x_k - 4 & \text{f)} \sum_{i=1}^5 x_i^2 \\ \text{g)} (\sum_{i=1}^5 x_i)^2 & \text{h)} \sum_{i=1}^5 x_i y_i & \text{i)} \sum_{i=1}^5 x_i \sum_{i=1}^5 y_i \end{array}$$

1.10. Um viticultor registou o peso diário das uvas recolhidas durante os 15 dias de uma vindima, mas no fim só forneceu o peso médio diário - 515 kg.

- Qual foi a produção total (peso em kg) daquele período?
- Alguém comentou que naqueles 15 dias o peso mínimo diário colhido tinha sido 150 kg e o peso máximo diário 475 kg. O que pensa destas afirmações?
- Constatou-se que num dos dias tinha havido erro no registo do peso de uvas colhidas. Por engano o registo desse dia foi de 20 kg. Qual o valor do peso médio diário, depois de retirado aquele registo?

1.11. Considere dois conjuntos de dados: o primeiro contendo o registo diário, efectuado durante 50 dias, do número de casos de intoxicação ocorridos numa fábrica e o segundo com o registo das preferências relativamente a 5 tipos de mistura de café (designadas por A, B, C, D e E) manifestadas num inquérito feito a 1000 consumidores.

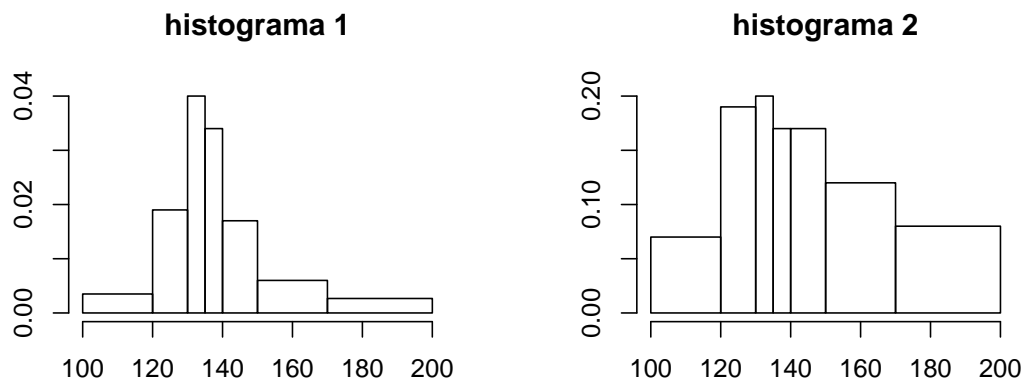
Nº de casos	0	1	2	3	4	5	6	Misturas de café	A	B	C	D	E
Nº de dias	13	15	8	6	5	2	1	Nº de respostas	190	210	180	205	215

- Indique a variável considerada em cada um dos casos e classifique-a.
- Determine uma medida de localização adequada a cada um dos conjuntos de dados.
- Indique o valor do mínimo e do máximo de cada conjunto de dados, caso existam.

1.12. Na tabela que se segue apresenta-se o agrupamento dos dados relativos a uma amostra de alturas (em dm) de 100 árvores de uma mesma espécie.

classe	[100;120[[120;130[[130;135[[135;140[[140;150[[150;170[[170;200[
nº de árvores	7	19	20	17	17	12	8

- Elabore uma tabela de frequências relativas e frequências relativas acumuladas, dos valores observados.
- Indique, justificando, qual dos histogramas apresentado a seguir se pode considerar o mais adequado para descrever o agrupamento de dados apresentado?
- Determine a mediana, aproximada, da altura das árvores observadas.



1.13. No início do 9^o ano de uma escola secundária passou-se um pequeno inquérito aos alunos das quatro turmas existentes, no qual se pedia que assinalassem com uma cruz (×) ou indicassem a resposta correcta à sua situação no final do ano lectivo anterior:

Idade	<input type="text"/>	Sexo	<input type="checkbox"/> F	<input type="checkbox"/> M	Altura (cm)	<input type="text"/>
Classificação(nível) obtida no último período na disciplina de Português	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5					

Os dados recolhidos foram registados no ficheiro “inquerito.csv”.

Para cada uma das variáveis em estudo, faça uma representação gráfica adequada e interprete a informação contida nos dados.

1.14. Considere o quadro seguinte com os dados da altitude das principais serras do Continente (Fonte: Instituto Geográfico e Cadastral e Centro de Estudos Geográficos; dados reproduzidos no *Anuário Estatístico*, I.N.E., Lisboa, 1980):

Designação	Altitude (m)	Designação	Altitude (m)
Peneda	1416	Gardunha	1227
Soajo	1415	Leomil	1008
Gerês	1507	Lapa	953
Barroso	1208	Marofa	973
Larouco	1525	Malcata	1075
Cabreira	1261	Grândola	325
Alvão	1283	Cercal	372
Marão	1415	Espinhaço de Cão	297
Padrela	1146	Monchique	902
Coroa	1273	Caldeirão	577
Montezinho	1438	Mendro	412
Nogueira	1318	Ossa	653
Bornes	1200	S.Mamede	1025
Mogadouro	993	Adiça	522
Montemuro	1382	Sicó	553
Arada	1116	Aire	679
Caramulo	1071	Candeeiros	613
Buçaco	549	Montejunto	664
Lousã	1204	Sintra	528
Açor	1340	Arrábida	501
Estrela	1991	Monte Figo	411
Alvelos	1084		

Os dados foram introduzidos no *software*  e estão disponíveis no objecto `serras`, no ficheiro “Serras.RData”.

- Agrupe os dados em classes. Faça a sua representação gráfica.
- Comente a distribuição das altitudes das serras.
- Averigúe se há candidatos a “outlier” no conjunto dos dados.

1.15. Os valores da precipitação (em mm) registada na Estação Meteorológica de Lisboa, nos 31 dias do mês de Janeiro de um dado ano, foram os seguintes (dados do Instituto de Meteorologia):

Dia	Precip.	Dia	Precip.	Dia	Precip.
1	0.0	11	3.8	21	0.9
2	0.0	12	0.3	22	0.3
3	0.0	13	0.0	23	18.2
4	0.0	14	0.0	24	4.0
5	4.7	15	0.5	25	4.6
6	0.6	16	7.0	26	22.0
7	17.2	17	0.0	27	15.6
8	1.4	18	0.0	28	0.0
9	11.2	19	3.3	29	3.4
10	1.0	20	7.6	30	0.0
				31	0.0

- Construa o histograma para os dados da precipitação e comente-o.
- Obtenha a caixa-de-bigodes dos dados e comente-a.
- Calcule a precipitação média e mediana diária em Lisboa, naquele ano. Compare os valores obtidos da média e da mediana e comente, tendo em atenção que ambos são indicadores de localização.

d) Introduza os dados no *software* \mathbb{R} . Responda às questões anteriores utilizando o \mathbb{R} .

1.16. Num pomar de pêra-rocha registou-se o número de pêras que foram colhidas no ano passado em cada uma das 60 pereiras. Os dados recolhidos foram introduzidos no ficheiro “peras.dat”.

- Construa uma tabela de frequências e faça a representação gráfica do número de pêras em cada pereira daquele pomar.
- Qual a produção média e a produção mediana de cada pereira? Calcule ainda o desvio padrão da produção de cada pereira. Comente.
- Construa a caixa de bigodes dos dados apresentados.

1.17. Um biólogo está a estudar uma unidade de aquicultura de criação de douradas. Num dado dia recolheu 15 douradas no viveiro *A* e obteve como peso médio e variância $\bar{x}_1 = 235 \text{ g}$ e $s_1^2 = 254 \text{ g}^2$; recolheu 20 douradas no viveiro *B* e obteve $\bar{x}_2 = 245 \text{ g}$ e $s_2^2 = 267 \text{ g}^2$.

Indique a média e a variância do conjunto das 35 douradas observadas nos dois viveiros.

1.18. Diga **justificando** se são verdadeiras ou falsas as afirmações que se seguem:

- A amplitude interquartil é metade da amplitude total.
- A média está sempre entre o primeiro e o terceiro quartil.
- A mediana está sempre entre o primeiro e o terceiro quartil.
- O desvio padrão é sempre igual à amplitude interquartil.
- O desvio padrão é menor do que a média dos desvios relativos à média.

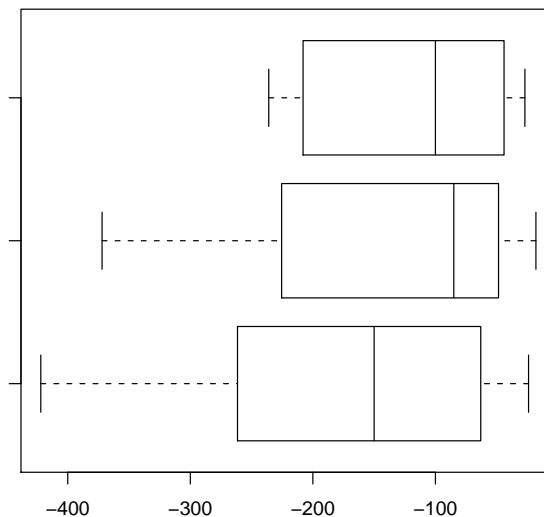
1.19. Num estudo, efectuado nos dois últimos anos, sobre o número de golfinhos observados em cada passeio organizado pela empresa OlhóGolfinho no estuário do Sado, obtiveram-se os seguintes dados:

nº de golfinhos	0	1	2	3	4	5	6	8
nº de passeios	17	45	84	52	23	11	2	1

- Diga qual a variável em estudo e classifique-a.
- Apresente os dados numa tabela de frequências relativas e faça a representação gráfica adequada. Comente.
- Descreva a amostra indicando medidas de localização central e de dispersão.
- Qual é a percentagem de passeios em que no máximo se observaram 2 golfinhos?

1.20. Numa experiência medem-se fluxos de calor de meia em meia hora, das 7h às 18h (inclusivé), durante três dias consecutivos. Os resultados obtidos (em $W m^{-2}$) são indicados na tabela em baixo. Ao lado da tabela estão as caixas-de-bigodes dos três dias, sem qualquer ordem aparente. Os dados foram introduzidos no *software* R e estão disponíveis no objecto `fluxoCalor`, no ficheiro “FluxoCalor.RData”.

DIA 1	DIA 2	DIA 3
-27	-24	-85
-32	-38	-74
-31	-61	-49
-53	-54	-31
-67	-59	-18
-48	-65	-32
-38	-67	-33
-47	-74	-57
-41	-120	-34
-41	-150	-59
-63	-171	-48
-114	-50	-92
-100	-98	-138
-100	-175	-74
-175	-184	-103
-208	-178	-196
-228	-228	-194
-208	-295	-259
-208	-320	-255
-196	-359	-284
-236	-401	-324
-210	-422	-294
-216	-405	-372



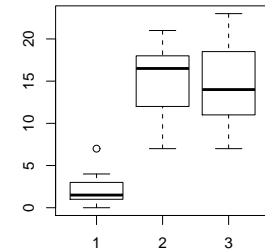
- Associe cada diagrama ao respectivo dia. Justifique.
- Sem fazer contas, diga se a média correspondente ao diagrama do topo será inferior ou superior a -100. Justifique.
- Considere agora o conjunto das observações nos 3 dias.
 - Calcule os indicadores de localização e dispersão destes dados.
 - Desenhe o *boxplot* dos dados e compare com os que lhe são fornecidos.
 - Construa uma tabela de frequências para dados agrupados em classes de amplitude 50.
 - Use a tabela da alínea anterior para calcular valores aproximados da média e da mediana das observações nos três dias. Comente os resultados.

1.21. Num estudo realizado para avaliar o efeito de três *sprays*, *A*, *B* e *C*, em insectos, organizaram-se 3 grupos de 12 recipientes cada, nos quais se colocou o mesmo número de insectos a que se aplicaram aqueles insecticidas. Indicadores relativos ao nº de insectos mortos em cada um deles, encontram-se no quadro no final do exercício.

- Associe cada *boxplot* a cada *spray*, indicando o valor das barreiras de *outliers* no primeiro diagrama. Justifique.

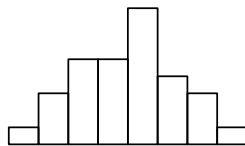
- b) Compare os três conjuntos de dados quanto à localização, dispersão e simetria.
 c) Para a totalidade das observações calcule a média, a variância e a amplitude total.
 d) A totalidade dos dados foi agrupada em classes, com extremos 0, 5, 10, 15, 20 e 25, e frequências absolutas observadas, respectivamente, 11, 5, 9, 8 e 3. Calcule a média, mediana e variância para os dados agrupados.

Spray	$\sum_{i=1}^{12} x_i$	$\sum_{i=1}^{12} x_i^2$	min	Q1	Q2	Q3	max
A	174	2768	7	11.5	14.0	17.8	23
B	184	3022	7	12.5	16.5	17.5	21
C	25	95	0	1.0	1.5	3.0	7



1.22. Na figura que se segue apresentam-se, para 3 conjuntos de dados, os histogramas e respectivas caixas de bigodes sem qualquer ordem. Associe cada histograma à caixa de bigodes relativa ao mesmo conjunto de dados.

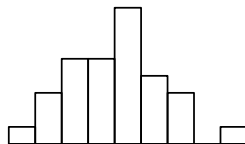
hist 1



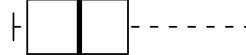
boxplot A



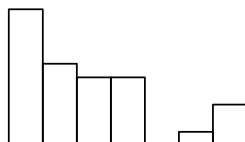
hist 2



boxplot B



hist 3



boxplot C



Exercícios de Estatística Descritiva a duas dimensões

1.23. As colunas da *data frame* `cor` (no ficheiro “Cor.RData”) contêm, respectivamente, a cor dos olhos e do cabelo de 300 portugueses. Com estes dados, construa uma tabela de contingência. Determine as frequências marginais e indique o seu significado.

1.24. A tabela seguinte mostra os valores do índice de preços ao consumidor (IPC) em Portugal nos últimos anos.

i	1	2	3	4	5
Ano(x)	2002	2003	2004	2005	2006
IPC(y)	100	103.3	105.7	108.1	111.5

- Calcule $cov(x, y)$ e a média e a variância de x e y .
 - Se aos anos de observação x se tivesse aplicado a transformação $x/2 - 1000$, i.e., se se considerasse os valores 1, 1.5, 2, 2.5, 3, qual seria o valor de $cov(x', y)$, com $x' = x/2 - 1000$?
 - Comente os resultados obtidos em a) e b) e diga se poderá considerar-se a covariância um bom indicador da existência de uma relação forte entre x e y . Justifique.
 - Independentemente das respostas anteriores, determine a equação da recta de regressão de y sobre x .
 - Calcule a precisão da recta e interprete o seu significado.
 - Qual a variação anual média dos preços, estimada pela regressão, no período 2002-06?
 - Mantendo-se a actual tendência, qual prevê que seja o IPC em 2007?
 - Quais seriam os valores dos coeficientes da recta de regressão se o índice de preços ao consumidor tivesse como base o ano de 2003, i.e. $y'_i = 100y_i/y_2$?
- 1.25.** Para $n = 20$ pares de observações (x_i, y_i) , seja $y = -5.6 + 0.7x$, a equação da recta de regressão dos mínimos quadrados de y sobre x .

- Comente as seguintes afirmações:
 - O coeficiente de correlação entre y e x é positivo porque o declive da recta é positivo.
 - Em média, quando x aumenta y não aumenta, pois o declive da recta é menor do que 1.
- Sendo $\sum_{i=1}^{20} x_i = 200$ determine $\sum_{i=1}^{20} y_i$.

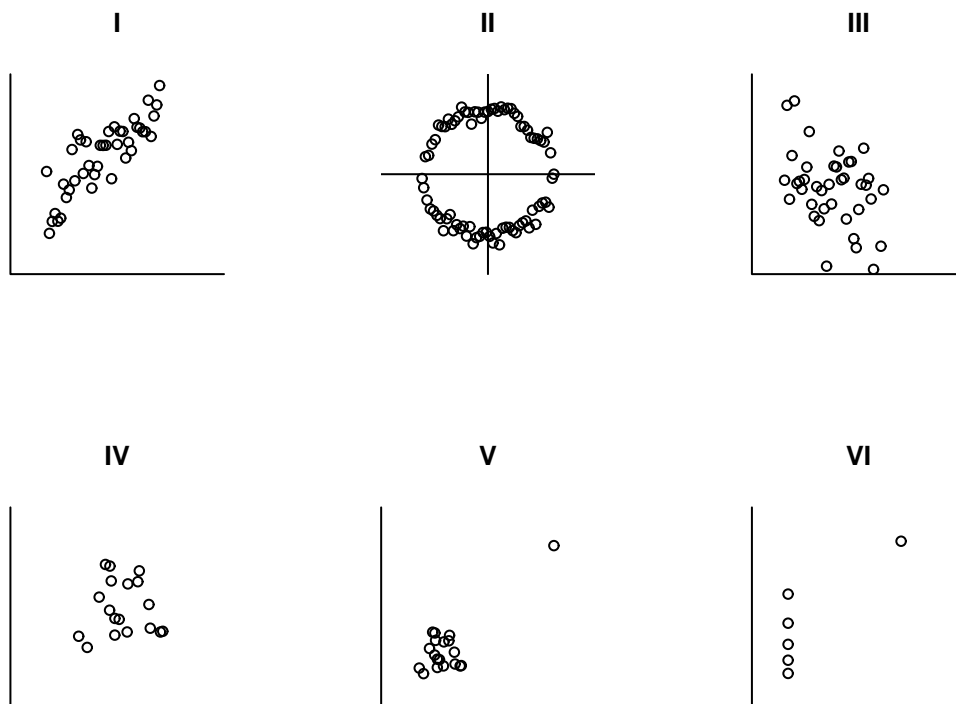
1.26. Indique qual dos valores abaixo indicados se aproxima mais do coeficiente de correlação dos dados descritos nas seguintes nuvens de pontos:

a) 0

b) 0.8

c) -0.5

d) 2.0



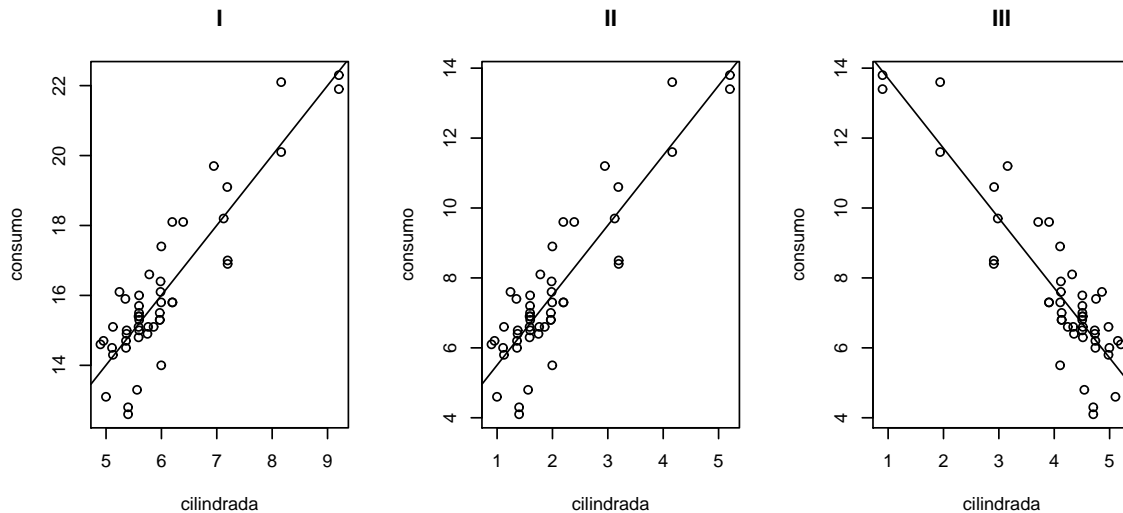
1.27. Num estudo sobre o consumo de gasolina de vários modelos de automóveis ligeiros de passageiros e a cilindrada do respectivo motor, foi estabelecida a seguinte equação da recta de regressão dos mínimos quadrados

$$y = 3.5 + 2x$$

em que x é a cilindrada (em 10^3 cm^3) e y é o consumo (em litros por 100 km percorridos). Sabendo que a precisão desta recta é de 0.803 e que a média e o desvio padrão das cilindradas observadas foram de 2.027 e 0.994 (10^3 cm^3), respectivamente, responda às seguintes questões:

- Determine a média e o desvio padrão dos consumos de gasolina dos automóveis observados.
- Qual é a variação esperada para o consumo de gasolina quando se aumenta a cilindrada de 1000 cm^3 ?

- c) Qual dos seguintes gráficos corresponde à nuvem de pontos e à respectiva recta de regressão do estudo descrito?



- d) Parece-lhe adequada a utilização do modelo linear para descrever a relação entre o consumo de gasolina e a cilindrada do motor nos modelos de automóveis analisados? Justifique.

1.28. A evaporação dos solventes que se usam nas tintas depende da humidade ambiente. O conhecimento desta relação poderá ser útil para melhorar a qualidade da operação de pintura. Foi realizado um estudo para examinar a relação entre x - “humidade relativa ambiente (%)” e y - “quantidade de um determinado solvente evaporado durante a pintura (% do peso)”. Desse estudo resultaram os seguintes dados:

$$n = 20; \quad \bar{x} = 52.5; \quad \bar{y} = 9.5; \quad s_x^2 = 256.5789; \quad s_y^2 = 10.2632; \quad cov(x, y) = -46.0526$$

- Classifique, justificando, a variável x - “humidade relativa ambiente”.
- Poder-se-á admitir a existência de uma relação linear entre as variáveis? Justifique.
- Independentemente da resposta à alínea anterior, determine a recta de regressão dos mínimos quadrados de y sobre x . Indique uma medida da precisão dessa recta e interprete o seu valor.
- Suponha que foi registado o resíduo $e = 0.34$, associado à observação $x = 55$, relativamente à recta de regressão definida em c). Qual o correspondente valor observado para y ?

1.29. A medição directa do calor específico de ramos de macieira é difícil de efectuar. Um investigador propõe prever o calor específico de ramos individuais a partir de

medições (muito mais simples de efectuar) da percentagem de água no ramo, em vez de medir directamente o calor específico.

Para isso recolheu observações da percentagem de água (x) e do calor específico (y) de 21 ramos. Os valores obtidos (registados no ficheiro “CalorEspecifico.RData”) são os seguintes :

x	y	x	y	x	y
49	46	53	57	62	119
58	90	50	44	63	131
59	104	57	100	52	53
51	65	53	89	51	70
56	85	60	96	65	131
61	113	52	69	52	66
56	96	58	111	54	69

- Desenhe o diagrama de extremos e quartis para os valores do calor específico observados. Comente a distribuição dos dados.
- Parece-lhe adequada a existência de uma relação linear entre x e y ? Porquê? Independentemente da sua resposta ajuste aos dados a recta de regressão dos mínimos quadrados.
- Qual o valor que se prevê para o calor específico quando a percentagem de água é de 60? Justifique.
- Sabe-se que, para facilitar os cálculos, os valores originais obtidos para o calor específico dos ramos (y') foram transformados de acordo com a expressão $y = 1000 y' - 600$, sendo os valores de y os registados na tabela dada acima. Suponha que lhe era pedido para escrever a regressão linear entre x e y' ; deduza a relação existente entre os coeficientes da nova recta e os da recta que obteve em b). Haverá alteração na precisão da regressão?

1.30. Numa dada região, registou-se anualmente entre 1998 e 2006 a produção de trigo. Designando por x o ano e por y a produção de trigo, em milhares de toneladas, obtiveram-se os seguintes valores para os 9 pares de observações efectuadas:

$$\bar{x} = 2002; \quad \bar{y} = 270.5; \quad \sum_{i=1}^9 (x_i - \bar{x})^2 = 60$$

$$\sum_{i=1}^9 (y_i - \bar{y})^2 = 1416.2; \quad \sum_{i=1}^9 (x_i - \bar{x})(y_i - \bar{y}) = -203$$

- Determine a recta de regressão dos mínimos quadrados da evolução da produção de trigo em função do tempo. Indique a sua precisão.
- Se se decidisse identificar os anos por 1,...,9 respectivamente, qual seria a precisão da recta de regressão que se obteria considerando esta transformação? Justifique convenientemente.

1.31. A seguinte tabela apresenta o período de gestação (x), em dias, e o tempo médio de vida (y), em anos, registados em 10 mamíferos.

	urso	hipopótamo	canguru	leopardo	leão	macaco	rato	porco	cão	gato
x_i	219	238	42	98	100	164	21	112	61	63
y_i	18	25	7	12	15	15	3	10	12	12

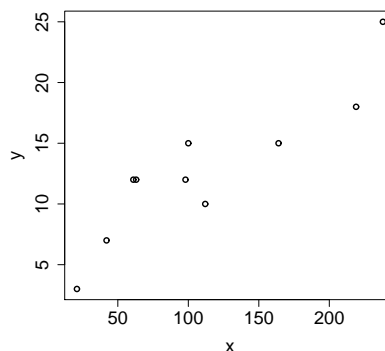
Os dados foram introduzidos no *software* \mathbb{R} , apresentando-se abaixo os resultados obtidos:

```
x<-c(219,238,42,98,100,164,21,112,61,63)
y<-c(18,25,7,12,15,15,3,10,12,12)
```

```
>mean(x)      > mean(y)
[1] 111.8      [1] 12.9
```

```
> var(x)      > var(y)
[1] 5394.622   [1] 36.1
```

```
> cov(x,y)    > plot(x,y)
[1] 396.7556
```



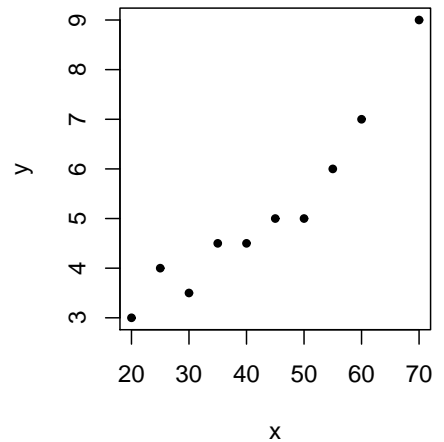
- Parece-lhe adequada a existência de uma relação linear entre x e y ? Justifique.
- Independentemente da resposta à alínea anterior determine a recta de regressão dos mínimos quadrados de y sobre x . Calcule a precisão da recta e interprete o seu significado.
- Interprete, no contexto do problema, o significado do coeficiente de regressão de y sobre x .
- O período de gestação de uma girafa é de 425 dias. Se usasse a recta determinada em b) que previsão obteria para o seu tempo médio de vida? Critique o resultado obtido, sabendo que o tempo médio de vida de uma girafa é de 10 anos.
- Determine a recta de regressão dos mínimos quadrados de “tempo médio de vida” sobre “tempo de gestação”, sendo agora o tempo de gestação, x' , dado em meses ($x' = x/30$). Qual a precisão desta recta?

1.32. Num estudo em que se pretende avaliar a influência da velocidade do vento (m/s) na quantidade de água (centenas de litro) evaporada por dia na albufeira de uma barragem, obtiveram-se os seguintes dados que foram introduzidos no *software* \mathbb{R} .

```

> x<-c(20, 50,30,55,70,45,60,25,40,35) #vel. vento (m/s)
> y<-c(3,5,3.5,6,9,5,7,4,4.5,4.5)      #agua evaporada
> mean(x)
[1] 43
> mean(y)
[1] 5.15
> var(x)
[1] 256.6667
> var(y)
[1] 3.169444
> cov(x,y)
[1] 27
> plot(x,y,pch=20)

```



Com base nos resultados apresentados, responda às seguintes questões.

- Parece-lhe adequada a existência de uma relação linear entre x e y ? Justifique.
- Independentemente da resposta à alínea anterior determine a recta de regressão dos mínimos quadrados de y sobre x . Calcule a precisão da recta e interprete o seu significado.
- Determine a equação da recta de regressão de “quantidade de água evaporada” sobre “velocidade do vento” no caso de os valores da velocidade do vento serem dados em km/h . Qual será a precisão desta recta? Justifique. (**Note que** $1 m/s = 3.6 km/h$).
- Determine $cov(x', y)$.

1.33. Foram seleccionadas aleatoriamente 20 folhas de videira da casta Água Santa, tendo sido medidos, para cada folha, os comprimentos (em mm) da nervura principal (variável NP) e das nervuras laterais esquerda (variável $NLesq$) e direita (variável $NLdir$), bem como a área foliar (variável $Area$, em mm^2). Alguns indicadores associados aos valores observados são:

NLesq		NP		NLdir		Area	
Min.	: 8.20	Min.	: 8.80	Min.	: 8.90	Min.	: 134.00
Median	:10.70	Median	:12.05	Median	:10.80	Median	: 199.00
Mean	:10.70	Mean	:11.97	Mean	:10.71	Mean	: 208.45
Max.	:15.10	Max.	:15.70	Max.	:14.10	Max.	: 356.50
Var	: 3.001053	Var	: 3.031447	Var	: 1.804711	Var	:3188.076316

```

> var(NLdir-NLesq)
[1] 0.8626053

```

- a) Calcule o primeiro quartil da variável NP , sabendo que os valores observados foram:

15.7 15.4 14.0 12.7 12.6 12.6 12.6 12.6 12.5 12.4 11.7 11.6 11.5
11.1 10.8 10.5 10.5 10.2 9.7 8.8

- b) Ajustou-se uma recta de regressão de área foliar ($Area$) sobre comprimento da nervura principal (NP), tendo-se obtido a equação $Area = -137.951 + 28.927NP$.
- Qual a variação esperada na área foliar associada a um aumento de 1 mm no comprimento da nervura principal, estimada pela regressão?
 - Sabe-se que uma das 20 folhas observadas no ajustamento tinha 11.3 mm de nervura principal e uma área foliar de 190.0 mm^2 . Qual o resíduo associado a esta folha?
 - Determine a precisão da recta de regressão e interprete o valor obtido.

1.34. Considere os quatro conjuntos de dados seguintes (dados de Anscombe, 1973):

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

- Calcule as médias e as variâncias de cada uma das oito variáveis. Comente.
- Calcule os coeficientes de correlação entre as variáveis x e as variáveis y de cada um dos quatro pares de variáveis. Comente.
- Calcule as rectas de regressão de y sobre x para cada um dos quatro pares de variáveis (x_i, y_i) , $(i=1, \dots, 4)$.
- Construa as quatro nuvens de pontos correspondentes aos pares de variáveis utilizados nas duas alíneas anteriores. Comente, à luz dos resultados das alíneas anteriores.
- Construa os gráficos dos resíduos para cada conjunto de pares de observações e comente-os.

Exercícios de Revisão de Estatística Descritiva

R1.1. Considere n pares de observações (x_i, y_i) . Seja $z_i = \frac{x_i - \bar{x}}{s_x}$, $i = 1, \dots, n$.

- Mostre que as observações z_i têm média nula e variância unitária.
- Determinou-se a recta de regressão dos mínimos quadrados de y sobre x tendo-se obtido $y = 3 + 1.5x$. Sabendo que $\bar{y} = 10.5$ e $s_x^2 = 0.25$, determine a equação da recta de regressão de y sobre z .

R1.2. A densidade óptica (d) de uma solução de um dado produto químico, medida para oito níveis diferentes de concentração (c), está registada na seguinte tabela (considerando unidades de medição adequadas):

c_i	1	2	4	5	8	10	12	15
d_i	4	9	18	20	35	41	42	60

$$\sum_{i=1}^8 c_i = 57 \quad \sum_{i=1}^8 d_i = 229 \quad \sum_{i=1}^8 c_i d_i = 2288 \quad \sum_{i=1}^8 c_i^2 = 579 \quad \sum_{i=1}^8 d_i^2 = 9091$$

- Pretende-se ajustar uma recta de regressão aos dados obtidos. Parece-lhe admissível tal ajustamento? Justifique convenientemente.
- Independentemente da sua resposta à alínea anterior, escreva a equação da recta de regressão dos mínimos quadrados que relaciona as variáveis envolvidas na experiência. Qual a precisão da recta que obteve? Comente.
- Sem efectuar novos cálculos, altere uma única observação de c , de modo que se verifique uma diminuição da média \bar{c} , um aumento da variância s_c^2 e um valor idêntico para a mediana \tilde{c} . Justifique.

R1.3. Pretende-se estudar a relação existente entre a superfície florestal (y) e a superfície territorial (x), expressas em milhares de hectares, nos 18 distritos do Continente. A equação da recta de regressão ($y = a + bx$), calculada a partir dos dados das Estatísticas Agrícolas do INE, num dado ano, tem os seguintes coeficientes: $a = 13.1$; $b = 0.32$.

- Comente as seguintes afirmações:
 - Em média, quando a superfície territorial aumenta, não aumenta a superfície florestal, pois b é menor do que 1;
 - O coeficiente de correlação entre a superfície florestal e a superfície territorial tem de ser positivo, porque o coeficiente a é positivo.
- Sendo a superfície territorial total do Continente 8892.7 milhares de hectares, diga qual a superfície florestal total.

R1.4. Num projecto de construção de mesas para computadores, verificou-se ter interesse avaliar a distância entre o assento e os cotovelos, estando uma pessoa sentada. Designando essa quantidade por y , procura-se relacioná-la com a altura total da pessoa (x). Os valores de uma amostra de dimensão $n = 22$ são dados na tabela seguinte:

altura(x) (cm)	distância ao cotovelo (y) (cm)			
159	22	23		
160	25	25	27	
161	24	27	25	26
162	23	26	27	29
166	27	23	28	
168	27	29	31	31
172	34	35		

(Nota: $\sum_{i=1}^{22} x_i^2 = 590754$ $\sum_{i=1}^{22} y_i^2 = 16288$ $\sum_{i=1}^{22} x_i y_i = 97547$)

- Calcule a distância média entre os assentos e os cotovelos e a altura média dos indivíduos observados.
- Calcule a mediana da variável altura.
- Estime um modelo de regressão linear simples da distância (y) em função da altura das pessoas. Indique a precisão da recta e comente-a.
- Considere que o par (166, 23) resulta de uma medição errada e foi decidido retirá-lo. Deduza à custa dos somatórios dados na Nota os parâmetros da recta construída com as observações restantes.
- Qual o aumento esperado para a distância entre o assento e o cotovelo por cada aumento unitário na altura de uma pessoa?

R1.5. Dados n pares de observações (x, y) , seja $y = a + bx$ a recta de mínimos quadrados ajustada.

- Defina coeficiente de correlação, $r_{x,y}$ e indique uma sua propriedade.
- Sendo $s_x^2 = 5.1$; $b = -3$; $\bar{x} = 3$; $\bar{y} = 2.8$ e $r^2 = 0.92$ determine s_y e a equação da recta de regressão.
- Prove que o declive da recta é invariante quando se efectua uma mesma transformação de escala a ambas as variáveis.

R1.6. Foi efectuado um estudo para analisar a relação entre o número de dias após a eclosão do ovo (variável x , em dias) e o comprimento das asas de crias de pardal doméstico (*Passer domesticus*) (variável y , em cm). Alguns indicadores associados aos dados observados são:

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Variância
x	3	6	10	10	14	17	21.83333
y	1.400	2.400	3.200	3.415	4.500	5.200	1.638077

A covariância entre as variáveis x e y é $5.9 \text{ cm} \times \text{dia}$.

- Averigúe se existem candidatos a *outliers* para os valores observados do comprimento das asas (y) e desenhe a respectiva caixa de bigodes, indicando os valores utilizados na sua construção.
- Poder-se-á admitir a existência de uma relação linear entre as variáveis? Justifique.
- Independentemente da resposta à alínea anterior, determine a recta de regressão dos mínimos quadrados de y sobre x . Qual é a variação diária média do comprimento das asas de crias de pardal doméstico prevista pela regressão?
- Para a recta de regressão determinada na alínea anterior obteve-se o resíduo -0.21538 associado à observação $x = 10$. Calcule o correspondente valor observado para y .
- Suponha que os dados do comprimento das asas foram registados em dm. Deduza a relação entre os coeficientes da recta de regressão neste caso e a obtida na alínea c).

R1.7. Para oito pares de observações $\{(x_i, y_i)\}_{i=1}^8$ determinou-se a recta de regressão dos mínimos quadrados, $y = 2.45 - 1.2x$, cuja precisão é 0.9604 . Responda, **justificando convenientemente**, se são verdadeiras ou falsas as afirmações nas seguintes alíneas.

- Sabendo que $\sum_{i=1}^8 x_i = 15$ então $\bar{y} = 0.2$.
- O coeficiente de correlação é $r = 0.98$.
- Se $s_x^2 = 0.5522$ então $s_y^2 = 0.828$.

Soluções de alguns Exercícios

- 1.7. a) Qualitativa nominal
b) Quantitativa contínua
c) Quantitativa discreta
d) Qualitativa ordinal
e) Quantitativa contínua
f) Quantitativa discreta
- 1.9. a) 38 b) 31 c) 30 d) 76 e) 32
f) 390 g) 1444 h) 98 i) 304
- 1.10. a) 7725 kg
c) 550.3571 kg
- 1.11. a) 1º conjunto - Conjunto da esquerda: variável - nº de casos de intoxicação em cada dia - quantitativa discreta
2º conjunto - Conjunto da direita: variável - tipo de mistura de café - variável qualitativa nominal
b) 1º conjunto: pode ser a média $\bar{x} = 1.7$ casos/dia ou a mediana $me = 1$ caso/dia ou a moda $mo = 1$ caso/dia
2º conjunto: moda $mo =$ tipo B.
c) Só é possível indicar o mínimo e o máximo para o 1º conjunto:
mínimo - 0 casos/dia; máximos 6 casos/dia
- 1.12. b) As classes dadas têm amplitudes diferentes e como a área de cada retângulo deverá representar a frequência relativa associada à respectiva classe, isto é, $A_i = f_i$, tem-se, por exemplo $A_1 = f_1 = 0.07$, $A_2 = f_2 = 0.19$, etc.
Como $f_3 = 0.2 \Rightarrow 0.2 = alt_3 \times 5 \Rightarrow alt_3 = 0.04$. Efectuando raciocínio análogo tem-se $alt_4 = 0.034$. Só poderá ser o histograma 1.
c) 1º calcular classe mediana que é $[135; 140[$.
 $me \approx 136.1765$
- 1.17. Para as 35 douradas tem-se $\bar{x} = 240.71g$ e $s^2 = 281.13g^2$
- 1.19. a) variável - nº de golfinhos em cada passeio - variável discreta porque toma valores em \mathbb{N}_0 .
c) média $\bar{x} = 2.28$, $mo = 2$ e $\tilde{x} = 2$ golfinhos por passeio.
Medidas de dispersão - variância e desvio padrão.

- d) Pretende-se $F(2) = \frac{17+45+84}{235} = 0.6213$, frequência relativa acumulada, logo é 62%
- 1.20.** a) Para Dia 1 tem-se $\min(x_i) = -236$ e $\max(x_i) = -27$; Dia 2 $\min(x_i) = -422$ e $\max(x_i) = -24$ e Dia 3 $\min(x_i) = -372$ e $\max(x_i) = -18$.
Então Dia 1 corresponde ao diagrama 1 (topo), Dia 2 corresponde ao diagrama 3 (o que está em baixo) e Dia 3 corresponde ao diagrama do meio.
- 1.22.** O histograma 1 corresponde ao *boxplot* C; o histograma 2 corresponde ao *boxplot* A e o histograma 3 corresponde ao *boxplot* B.
- 1.24.** a) $\sum_{i=1}^5 x_i = 10020$; $\sum_{i=1}^5 y_i = 528.6$; $\sum_{i=1}^5 x_i^2 = 20080090$;
 $\sum_{i=1}^5 y_i^2 = 55961.24$; $\sum_{i=1}^5 x_i y_i = 1059342$
 $cov(x, y) = 6.95$; $\bar{x} = 2004$; $\bar{y} = 105.72$; $s_x^2 = 2.5$; $s_y^2 = 19.412$
- b) $cov(x', y) = 3.475$
- d) $y = -5465.4 + 2.78x$
- e) $r^2 = 0.9953$, i.e., 99% da variabilidade do índice é explicada pela recta de regressão, portanto como é próximo de 1, a recta é muito precisa.
- g) $\tilde{y}_{2007} = 114.06$
- h) $b' = 2.69$; $a' = 5290.8$
- 1.25.** a) i) A afirmação é verdadeira. A relação $r = b s_x / s_y$, ($s_x > 0, s_y > 0$) estabelece que r e b têm o mesmo sinal. Logo $b > 0 \Rightarrow r > 0$.
ii) A afirmação é falsa, pois b é positivo. O coeficiente b representa a variação esperada para y quando x aumenta de uma unidade.
- b) Uma consequência da recta dos mínimos quadrados é: $\bar{y} = a + b\bar{x}$, o que é equivalente a $\sum y_i = na + b \sum x_i = 20 \times (-5.6) + 0.7 \times 200 = 28$.
- 1.26.** Nuvem I corresponde a b); nuvem II corresponde a a); nuvem III corresponde a c); nuvem IV corresponde a a); nuvem V corresponde a b); nuvem VI corresponde a b).
- 1.27.** a) $\bar{y} = 7.554$ litros por 100 Km; $s_y = 2.2185$ litros por 100 Km.
b) Espera-se que o consumo de gasolina aumente 2 litro por 100 Km.
c) Gráfico II.
- 1.30.** a) $y = 7043.43 - 3.383 x$; a precisão da recta é $R^2 = 0.485$.
b) A mesma precisão.
- 1.31.** a) O diagrama de dispersão sugere a existência de uma relação linear entre as variáveis x e y . Como $r = 0.899$ se pode considerar não muito afastado de 1, é de admitir a existência de uma relação linear entre x e y .

b) $y = 4.677 + 0.0735x$.

A precisão da recta é dada por $r^2 = 0.899^2 = 0.808$, o que significa que 80.8% da variabilidade de y é explicada pela regressão de y sobre x .

- c) O coeficiente de regressão de y sobre x , $b = 0.0735$, significa que, para aqueles mamíferos, por cada dia de aumento no período de gestação se espera um aumento de 0.0735 anos no seu tempo médio de vida.
- d) A previsão feita pela recta de regressão da alínea b) para o tempo médio de vida de uma girafa (sabendo que o seu período de gestação é de 425 dias) é

$$\hat{y} = 4.677 + 0.0735 \times 425 = 35.9 \text{ anos.}$$

Contudo, a utilização desta recta de regressão para prever o tempo médio de vida de uma girafa não é aconselhável, já que o valor da variável preditora ($x = 425$) não pertence à gama de valores observados de x ([21, 238]). Sendo assim, aquela recta não permite efectuar esta previsão, pelo que, a grande diferença entre o valor ajustado e o valor real (10 anos) é justificável.

1.33. a) 10.65 mm.

- b) i) Espera-se que a área foliar aumente 28.927 mm²
 ii) $e_{(NP=11.3)} = 1.076$ mm.
 iii) $R^2 = 0.798$.

R1.1. b) Seja $y = a' + b'z$ recta de y sobre z

$a' = \bar{y} - b'\bar{z}$ e como $\bar{z} = 0 \Rightarrow a' = \bar{y} = 10.5$

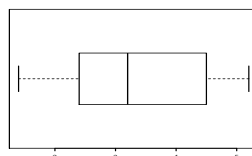
$b' = \frac{cov(z,y)}{s_z^2}$, como $z = \frac{x}{s_x} - \frac{\bar{x}}{s_x}$ $cov(z, x) = \frac{1}{s_x} cov(x, y)$ e $s_z^2 = 1$, então

$b' = \frac{1}{s_x} cov(x, y) = 0.75$

R1.6. a) Tem-se a barreira inferior $B_I = 2.4 - 1.5(4.5 - 2.4) = -0.75$ e a barreira superior $B_S = 4.5 + 1.5(4.5 - 2.4) = 7.65$, portanto não há, nos dados, valores inferiores à barreira inferior e também não há valores superiores à barreira superior, portanto não há candidatos a *outliers*.

Os valores necessários à construção da caixa de bigodes são:

$max(y) = 5.2$, $min(y) = 1.4$, $Q_1 = 2.4$, $Q_3 = 4.5$ e a mediana $\tilde{y} = 3.2$.



- b) Como não dispomos dos dados não podemos construir a nuvem de pontos, mas o coeficiente de correlação, $r = \frac{cov(x,y)}{s_x s_y} = \frac{5.9}{\sqrt{21.83333 \times 1.638077}} = 0.9866$, apresenta um valor próximo de 1 ($-1 \leq r \leq 1$), pelo que podemos admitir a existência de uma relação linear entre as variáveis.

c) A recta de regressão é $y = 0.713 + 0.2702x$.

O comprimento das asas aumenta por dia, em média, 0.2702 cm.

d) Como $y_i = a + b x_i + e_i$, onde e_i é o resíduo, então para $x_i = 10$ tem-se $y_i = 0.713 + 0.2702 \times 10 - 0.21538 = 3.1996$ cm.

e) $y' = 0.1y$ designa o comprimento das asas expresso em dm.

Sendo assim, consideremos b' e a' os coeficientes da recta de regressão de y' em x .

$$b' = \frac{\text{cov}(x,y')}{s_x^2} = \frac{0.1\text{cov}(x,y)}{s_x^2} = 0.1 b$$

$$a' = \overline{y'} - b'\overline{x} = 0.1\overline{y} - 0.1 b \overline{x} = 0.1(\overline{y} - b\overline{x}) = 0.1 a.$$