Matemática II

Exercícios de Estatística Descritiva

Fernanda Valente e Marta Mesquita

INSTITUTO SUPERIOR DE AGRONOMIA

- 2014/2015 -

I - ESTATÍSTICA DESCRITIVA

1. Considere os seguintes conjuntos de dados:
   (I) o registo semanal do número de operações de manutenção nos jardins de uma grande cidade;

Nº de operações	0	1	2	3	4	5
Nº de semanas	13	15	8	6	5	2

   (II) o registo dos lotes de sementes de relva vendidas por uma loja num ano;

Lote de sementes	A	B	C	D	E
Nº de embalagens	191	215	205	160	259

   (III) o registo diário da precipitação em mm na Tapada da Ajuda em Janeiro de 2011;

dia	1	2	3	4	5	6	7	8	9	10	11	12
(mm)	1	0	0	0	3	17.8	14.8	7.1	9.5	0	8	0
dia	13	14	15	16	17	18	19	20	21	22	23	24
(mm)	0	0	0	0	0	0	0	0	0	0	0	4.4
dia	25	26	27	28	29	30	31					
(mm)	0	0	0.6	9.5	2.4	0	0					

   (IV) o registo das respostas a um inquérito ao grau de satisfação dos utentes do Centro de Atendimento ao Munícipe de Lisboa.

Grau satisfação	Mau	Razoável	Bom	Muito bom	Excelente
Nº de respostas	8	79	225	171	17

   (a) Para cada um dos conjuntos de dados:
      i. indique a variável em estudo e classifique-a;
      ii. faça a representação gráfica adequada.
   (b) Nos casos em que é possível,
      i. determine uma medida de localização e uma medida de dispersão adequada;
      ii. desenhe a caixa de bigodes.

2. Uma câmara municipal avalia, através de um questionário com 50 perguntas, a qualidade dos espaços verdes existentes no concelho. Cada pergunta tem uma resposta de 1 a 5, em que uma maior nota significa maior desempenho. Foram inquiridas 42 pessoas e para cada uma foi calculada a nota média das suas respostas. Os resultados obtidos foram os seguintes:

4.2	2.7	4.6	2.5	3.3	4.7	4.0
2.4	3.9	1.2	4.1	4.0	3.1	2.4
3.8	3.8	1.8	4.5	2.7	2.2	3.7
2.2	4.4	2.8	2.3	1.9	3.6	3.9
3.3	3.4	3.3	1.8	3.5	4.1	2.2
3.0	4.1	3.4	3.2	2.2	2.0	2.8

(a) Proceda à organização dos dados, construindo um quadro de frequências.

(b) Faça uma representação gráfica adequada.

(c) Calcule medidas de localização e de dispersão adequadas a este conjunto de dados.

(d) Desenhe a caixa de bigodes.

(e) Calcule valores aproximados da média e do desvio padrão para os dados agrupados. Compare os resultados obtidos com os da alínea c) e comente.

3. A partir do Plano Director Municipal (PDM) de um concelho registaram-se as áreas de ocupação do uso do solo destinadas a Espaços Verdes tendo-se obtido os seguintes dados:

Área (m^2)	Espaços Verdes
]0 , 25]	13
]25 , 50]	9
]50 , 75]	12
]75 , 100]	5
]100 , 125]	1

(a) Identifique e classifique a variável em estudo.

(b) Desenhe o histograma da distribuição de frequências relativas.

(c) Calcule valores aproximados e explique o significado de cada um dos seguintes indicadores: (i) média (ii) mediana (iii) moda (iv) 1º quartil (v) desvio padrão (vi) amplitude interquartil

4. Os vencimentos, em euros, dos funcionários de três *ateliers* de arquitectura são descritos pela seguinte tabela:

Atelier	1	2	3
Nº de funcionários	5	12	8
Média	2187	1733	1236
Desvio padrão	1230	739	942

Pretende-se constituir um gabinete de estudo que englobe estes três *ateliers*.

(a) Calcule o vencimento médio e o respectivo desvio padrão para este novo gabinete.

(b) Face aos valores obtidos na alínea anterior, decidiu-se alterar linearmente os vencimentos de forma a que a média e o desvio padrão dos salários de todos os funcionários passem a ser 1750 e 900 euros, respectivamente. Sabendo que um funcionário do *atelier* 1 ganha actualmente 1000 euros, calcule o seu vencimento na nova estrutura. E qual será o valor de um vencimento de 4500 euros na nova escala ?

5. Do Recenseamento Geral Agrícola (RGA) retirou-se o seguinte quadro com os dados referentes ao número de explorações agrícolas da região do Alentejo, em função da superfície agrícola utilizada (SAU) (Murteira *et al.*, 2002):

SAU (ha)	$]0, 0.5[$	$[0.5, 1[$	$[1, 2[$	$[2, 5[$	$[5, 10[$	$[10, 20[$	$[20, 50[$	$[50, 100[$	$[100, 200[$	$[200, +\infty[$
Nº explorações	1639	4392	8476	9575	5666	3995	3389	1566	1053	559

   (a) Elabore uma tabela de frequências.

   (b) Faça uma representação gráfica adequada.

6. Os dados que se seguem referem-se à precipitação observada (em mm) em Lisboa e no Porto no mês de Janeiro, dos anos de 1980 a 1999 (Murteira *et al.*, 2002):

Lisboa	1980	53.5	1985	224.6	1990	63.1	1995	57.3
	1981	11.6	1986	52.2	1991	64.9	1996	394.0
	1982	112.8	1987	130.0	1992	61.5	1997	136.6
	1983	6.2	1988	107.7	1993	19.0	1998	62.3
	1984	40.6	1989	62.5	1994	79.7	1999	107.5
Porto	1980	118.2	1985	203.3	1990	164.5	1995	165.8
	1981	3.6	1986	171.3	1991	164.2	1996	331.4
	1982	94.4	1987	107.7	1992	99.3	1997	162.5
	1983	16.4	1988	217.1	1993	54.1	1998	149.1
	1984	279.0	1989	35.5	1994	232.1	1999	101.7

   (a) Compare a pluviosidade em Lisboa e no Porto, no mês de Janeiro, recorrendo a medidas de localização e de dispersão.

   (b) Compare os dois conjuntos de dados recorrendo a caixas de bigodes.

   (c) Poder-se-á admitir a existência de uma relação linear entre as precipitações de Lisboa e do Porto no mês de Janeiro.

7. Resolva computacionalmente o exercício anterior utilizando a seguinte metodologia:

   (a) Introduza os dados anteriores em três colunas de numa folha de cálculo e guarde o ficheiro no formato csv (valores separados por vírgulas).

   (b) Na aplicação R, leia os valores do ficheiro csv criado na alínea anterior e guarde-os numa *data frame*.

   (c) Responda às questões do exercício anterior utilizando o R.

   (d) Construa uma nuvem de pontos da precipitação observada em Lisboa (eixo horizontal) e no Porto (eixo vertical), no mês de Janeiro para os vários anos.

8. Pretende fazer-se uma análise do consumo médio de energia por agregado familiar. Durante 10 dias de um Inverno registou-se numa cidade a temperatura média diária (x, em ^{o}C) e o consumo médio de energia (y, em kW h) tendo se obtido os seguintes resultados:

i	1	2	3	4	5	6	7	8	9	10
x_i	15	14	12	14	12	11	11	10	12	13
y_i	4.3	4.4	5.3	4.6	5.5	5.9	5.7	6.2	5.2	5.0

$$\sum_{i=1}^{10} x_i = 124 \quad \sum_{i=1}^{10} y_i = 52.1 \quad \sum_{i=1}^{10} x_i y_i = 637.1$$

$$\sum_{i=1}^{10} x_i^2 = 1560 \quad \sum_{i=1}^{10} y_i^2 = 275.13$$

   (a) Calcule a covariância entre as variáveis x e y.

   (b) Calcule novamente a covariância mas considerando agora que o consumo médio de energia está expresso em Watts hora (W h). Compare o resultado obtido com o da alínea anterior e comente.

   (c) Parece-lhe adequada uma relacão linear entre x e y? Justifique.

   (d) Independentemente da resposta à alinea anterior, determine a recta de regressão dos mínimos quadrados de y sobre x.

   (e) Calcule a precisão da recta e interprete o seu significado.

   (f) Interprete, no contexto do problema, o significado do coeficiente de regressão de y sobre x.

   (g) Qual o valor do consumo médio de energia previsto para um dia em que a temperatura média é de 12^{o}C? E para um dia em que é de 25^{o}C? Comente.

   (h) Diga a que observação(ões) corresponde o resíduo -0.06875kW h.

   (i) Determine a recta de regressão dos mínimos quadrados do consumo médio de energia em kW h sobre a temperatura média diária em ^{o}F (^{o}F = 32+1.8$\times^{o}$C). Qual a precisão desta recta?

9. A regeneração natural nas florestas tropicais tem de sobreviver e crescer utilizando as manchas de luz intermitentes que chegam ao solo por entre as copas das árvores. Com o objectivo de estudar o efeito da duração dessas manchas na taxa relativa de crescimento de plantas de uma espécie arbórea (*Shorea leprosula*) da floresta tropical do Sudeste Asiático foi efectuada uma experiência cujos resultados foram os seguintes (dados na *data frame* `fleck` do ficheiro `SunFleck.RData`):

planta	duração (min)	taxa relativa de crescimento (semana^{-1})
1	3.4	0.010
2	3.2	0.008
3	3.0	0.007
4	2.7	0.005
5	2.8	0.003
6	3.2	0.003
7	2.2	0.005
8	2.2	0.003
9	2.4	0.000
10	4.4	0.009
11	5.1	0.010
12	6.3	0.015
13	7.3	0.017
14	6.0	0.016
15	5.9	0.020
16	7.1	0.021
17	8.8	0.024
18	7.4	0.019
19	7.5	0.016
20	7.5	0.014
21	7.9	0.014

(a) Construa uma nuvem de pontos com as observações anteriores.

(b) Poder-se-á admitir a existência de uma relação linear entre a duração das manchas de sol e a taxa relativa de crescimento? Justifique.

(c) Ajuste a recta de regressão dos mínimos quadrados da taxa relativa de crescimento sobre a duração das manchas. Indique a precisão da recta e interprete o seu valor.

(d) Qual é a variação média na taxa relativa de crescimento da planta quando se aumenta a duração da mancha de sol num minuto?

(e) Qual é a taxa relativa de crescimento prevista para uma planta que esteja exposta a uma mancha de sol durante 8 minutos. E durante 20 minutos?

10. Num estudo sobre a influência da velocidade do vento (x, em m s^{-1}) na quantidade de água que se evapora por dia na albufeira de uma certa barragem (y, em centenas de litros) foi estabelecida a seguinte equação da recta de regressão dos mínimos quadrados

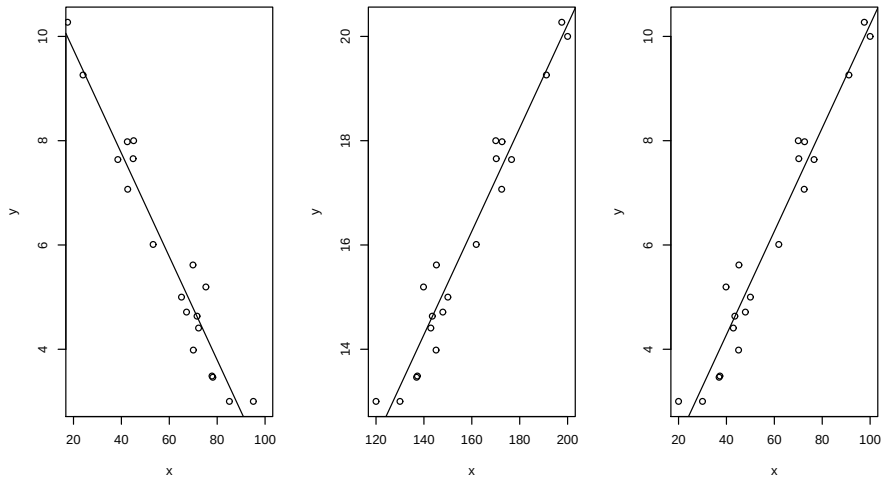
$$y = 0.313 + 0.099x.$$

Sabendo que a precisão da recta é de 0.9508 e que $\overline{x} = 57.54962$ e $s_x = 22.78833$, responda às seguintes questões:

(a) Determine a média e o desvio padrão da quantidade de água que se evapora por dia na albufeira desta barragem.

(b) Qual a variação esperada na quantidade de água evaporada, em litros, quando a velocidade do vento aumenta 1 m s^{-1}?

(c) Diga, justificando, qual dos seguintes gráficos corresponde à nuvem de pontos e à recta de regressão do estudo efectuado.



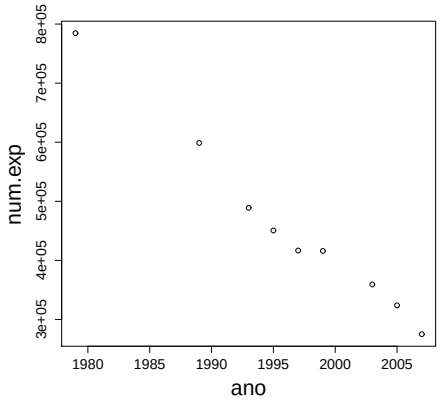
(d) Parece-lhe adequada a utilização do modelo linear para descrever a relação entre a quantidade de água que se evapora por dia na albufeira e a velocidade do vento? Justifique.

11. Com o objectivo de estudar a variação do número de explorações agrícolas em Portugal no período de 1979 a 2007, os dados disponibilizados em Pordata (www.pordata.pt) foram introduzidos na aplicação R. Alguns dos comandos e resultados obtidos foram os seguintes:

```
> ano <- c(1979,1989,...,2007)
> num.exp <- c(784497,598742,...,275083)
> plot(num.exp~ano)
> lm(num.exp~ano)

Coefficients:
(Intercept)          ano
   35576216       -17592

> cor(ano,num.exp)^2
[1] 0.9796289
```



(a) Determine o coeficiente de correlação entre as variáveis `num.exp` (número de explorações agrícolas) e `ano` e justifique porque é admissível considerar a existência de uma relação linear entre essas duas variáveis.

(b) Escreva a equação da recta de regressão linear dos mínimos quadrados do número de explorações sobre o ano e indique qual é a proporção da variabilidade do número de explorações que é explicada por esta regressão.

(c) De acordo com o modelo linear ajustado, quantas explorações agrícolas prevê que existiam no ano 2000?

(d) Qual é o significado, no contexto do problema em estudo, do valor -17592?

12. Num estudo americano sobre a influência da temperatura média do mês de Julho (x) na produção de milho (y), foram analisados dados destas duas variáveis para vários anos. No quadro que se segue apresentam-se os resultados obtidos para vários indicadores e para os coeficientes da recta de regressão dos mínimos quadrados de y sobre x.

    As unidades dos dados utilizados foram graus Fahrenheit (ºF), para a temperatura, e *bushels* por acre (bu/acre), para a produção de milho.

	x	y
média	75.17	50.0
variância	8.936553	173.9081
desvio padrão	2.989407	13.18742
covariância	-22.85813	
coef. correlação	-0.5798233	
recta de regressão	$y = 242.2708 - 2.5578\,x$	

    Para uma melhor interpretação dos resultados obtidos, pretende-se que estes venham expressos em graus Celsius (ºC), para a temperatura, e quilogramas por metro quadrado (kg/m^2), para a produção de milho.

    Tendo em conta estas novas unidades, recalcule os valores apresentados no quadro anterior, indicando as unidades respectivas.

    Conversões:

    1 bu = 35.2390702 kg

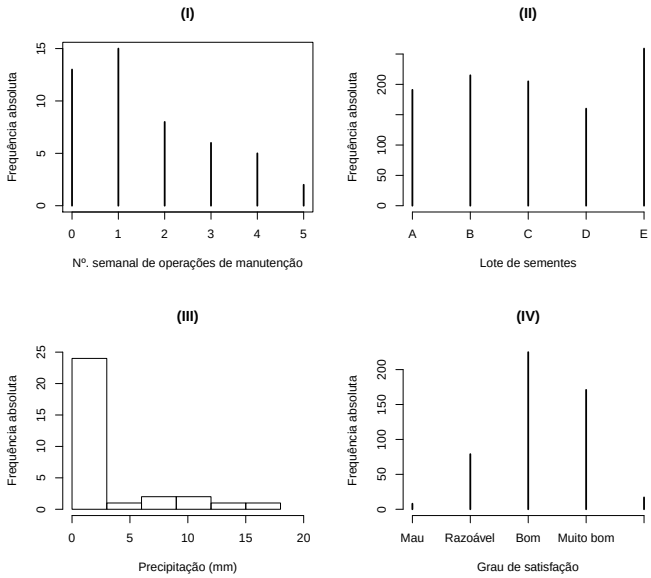
    1 acre = 4046.8564224 m^2

    T(ºC) = (T(ºF) $-32) \times 5/9$.

Soluções de alguns Exercícios

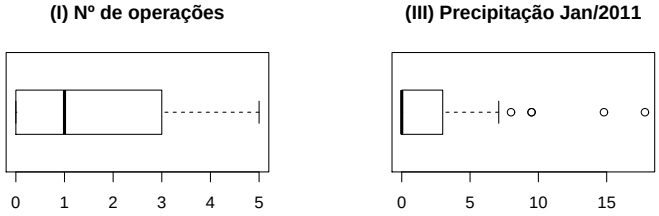
1. (a) i. (I) Número semanal de operações de manutenção nos jardins de uma grande cidade - variável quantitativa discreta;
(II) Lotes de sementes de relva vendidos numa loja num ano - variável qualitativa nominal;
(III) Precipitação diária na Tapada da Ajuda em Janeiro de 2011 (mm) - variável quantitativa contínua;
(IV) Grau de satisfação dos utentes do Centro de Atendimento ao Munícipe de Lisboa - variável qualitativa ordinal.

ii.



(b) i. Medidas de localização:
(I) $\bar{x} = 1.612, \quad \tilde{x} = 1, \quad Q_{0.25} = 0, \quad Q_{0.75} = 3, \quad \text{moda} = 1$
(II) $\text{moda} = E$
(III) $\bar{x} = 2.519, \quad \tilde{x} = 0, \quad Q_{0.25} = 0, \quad Q_{0.75} = 3$
(IV) $\text{moda} = \text{Bom}$
Medidas de dispersão:
(I) $A_{tot} = 5, \quad AIQ = 3, \quad s^2 = 2.159, \quad s = 1.469$
(II) —
(III) $A_{tot} = 17.8, \quad AIQ = 3, \quad s^2 = 22.317, \quad s = 4.724$
(IV) —

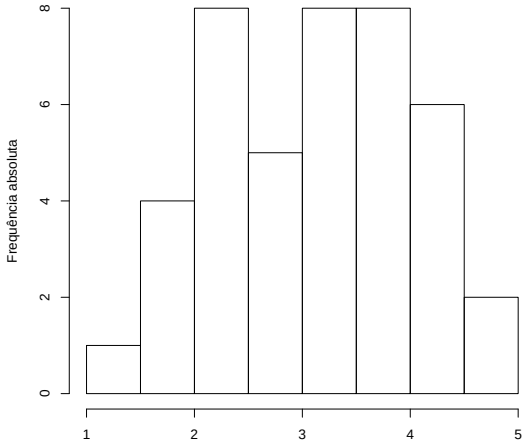
ii.



2. (a)

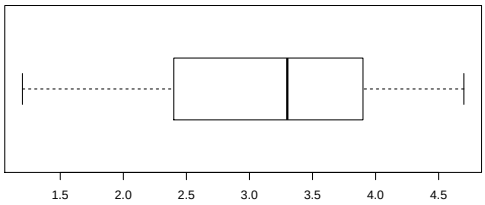
Classe	n_i	f_i	F_i
$]1.0, 1.5]$	1	0.023810	0.023810
$]1.5, 2.0]$	4	0.095238	0.119048
$]2.0, 2.5]$	8	0.190476	0.309524
$]2.5, 3.0]$	5	0.119048	0.428571
$]3.0, 3.5]$	8	0.190476	0.619048
$]3.5, 4.0]$	8	0.190476	0.809524
$]4.0, 4.5]$	6	0.142857	0.952381
$]4.5, 5.0]$	2	0.047619	1
Total	42	1	

(b)



(c) Medidas de localização: $\bar{x} = 3.16667$, $\tilde{x} = 3.3$, $Q_{0.25} = 2.4$, $Q_{0.75} = 3.9$;
Medidas de dispersão: $A_{tot} = 3.5$, $AIQ = 1.5$, $s^2 = 0.80032$, $s = 0.89461$.

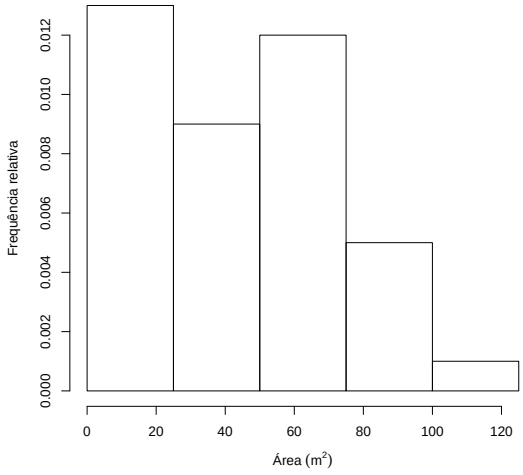
(d)



(e) $\bar{x}' = 3.119048$, $s' = \sqrt{0.81023} = 0.90012$.

3. (a) Área de ocupação do uso do solo destinada a espaço verdes no PDM de um concelho - variável quantitativa contínua.

(b)



(c) (i) $\bar{x}' = 45$ m^2 (ii) $\tilde{x}' = 44.44$ m^2 (iii) moda$= 25$ m^2 (iv) $Q'_{0.25} = 19.2308$ m^2
(v) $s' = \sqrt{787.5} = 28.06243$ m^2 (vi) $AIQ \approx 47.4259$ m^2.

4. (a) $\bar{x} = 1664.76$ euros, $\quad s = 938.9869$ euros

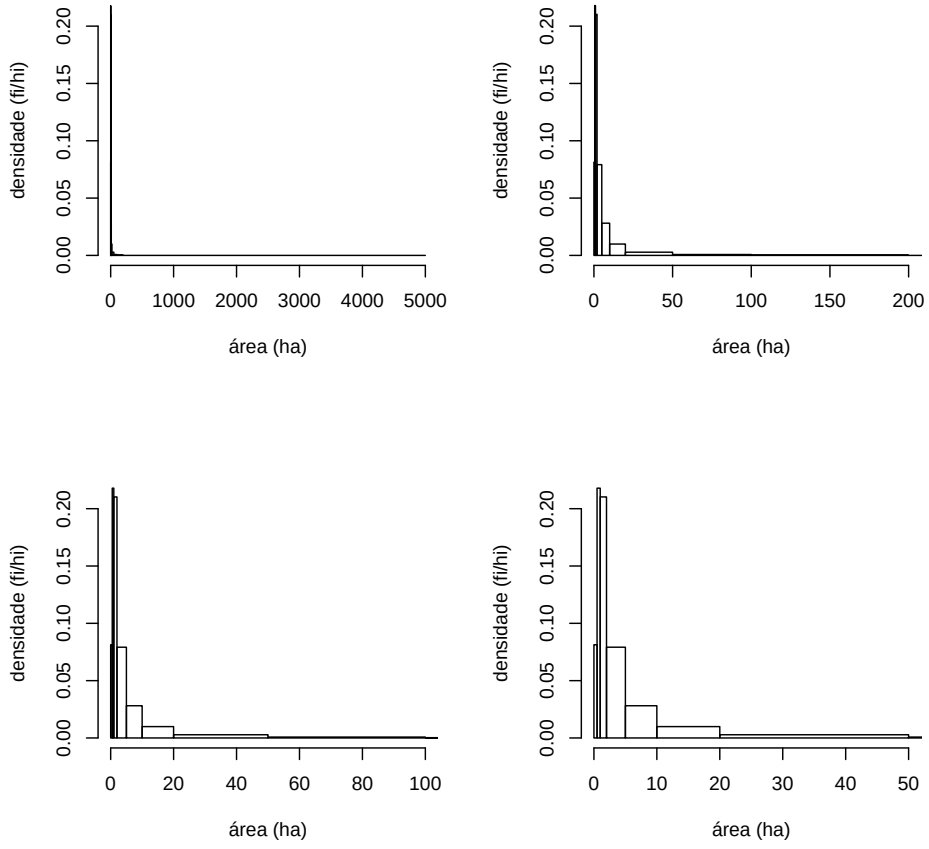
   (b) Salário actual = 1000 euros $\longrightarrow$ Salário futuro = 1112.841 euros

   Salário actual = 4500 euros $\longrightarrow$ Salário futuro = 4467.52 euros

5. (a)

classe	n_i	f_i	F_i	h_i	f_i/h_i
$]0, 0.5]$	1639	0.04066	0.041	0.5	0.082
$]0.5, 1]$	4392	0.108956	0.151	0.5	0.22
$]1, 2]$	8476	0.21027	0.361	1	0.21
$]2, 5]$	9575	0.237534	0.601	3	0.08
$]5, 10]$	5666	0.140561	0.742	5	0.0282
$]10, 20]$	3995	0.099107	0.841	10	0.0099
$]20, 50]$	3389	0.084073	0.921	30	0.002667
$]50, 100]$	1566	0.038849	0.96	50	0.00078
$]100, 200]$	1053	0.026123	0.986	100	0.00026
$]200, \mathbf{5000}]$	559	0.013868	1	4800	2.92E-06
Total	40310	1			

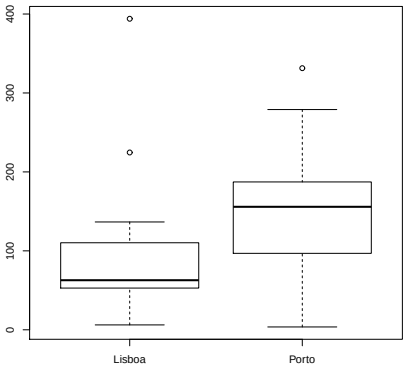
   (b) Notas: 1) para a construção do histograma foi necessário atribuir um limite superior à última classe; 2) Os 4 histogramas que se apresentam só diferem na escala do eixo dos xx.



6. (a)

Lisboa		Porto	
Min.	6.20	Min.	3.60
1st Qu.	52.85	1st Qu.	96.85
Median	62.80	Median	155.80
Mean	92.38	Mean	143.56
3rd Qu.	110.25	3rd Qu.	187.30
Max.	394.00	Max.	331.40
var	7558.27	var	7181.99
sd	86.94	sd	84.75

(b)



(c) $r = 0.4031$

8. (a) $cov(x, y) = -0.9933333$ °C kW h

(b) $cov(x, y^*) = -993.3333$ °C W h

(c) $r = -0.9834656$

(d) $y = 10.1589286 - 0.3991071\,x$

(e) $r^2 = 0.9672046$; Interpretação: 96.7% da variabilidade de y é explicada pela relação linear existente entre y e x.

(f) $b_1 = -0.3991$; Interpretação: Espera-se que o consumo médio de energia por agregado familiar diminua em 0.3991 kW h quando a temperatura média diária aumenta um grau Celcius.

(g) $\hat{y}|_{x=12^oC} = 5.3696$ kW h

(h)

i	x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	15	4.3	4.1723221	0.127678
2	14	4.4	4.5714292	-0.17143
3	12	5.3	5.3696434	-0.06964
4	14	4.6	4.5714292	0.028571
5	12	5.5	5.3696434	0.130357
6	11	5.9	5.7687505	0.13125
7	**11**	**5.7**	**5.7687505**	**-0.06875**
8	10	6.2	6.1678576	0.032142
9	12	5.2	5.3696434	-0.16964
10	13	5	4.9705363	0.029464

(i) $y = 17.2542 - 0.2217\,x^*$; $\quad r^2_{x^*y} = r^2_{xy} = 0.9672046$

9. (b) $r = 0.9076658$

(c) $y-$ taxa relativa de crescimento, $x-$ duração das manchas de sol
$y = -0.002743 + 0.002790\,x$; $\quad r^2 = 0.8238572$

(d) $b_1 = 0.002790$

(e) $\hat{y}|_{x=8} = 0.01957883$

10. (a) $\bar{y} = 6.01041(\times 10^2$ l); $\qquad s_y = 2.313679(\times 10^2$ l)

(b) 9.9 l

(c) Terceiro gráfico

(d) $r = 0.975$

11. (a) $r = -0.98976$

    (b) $\texttt{num.exp} = 35576216 - 17592\ \texttt{ano}$; $r^2 = 0.9796289$

    (c) $\widehat{\texttt{num.exp}}|_{\texttt{ano}=2000} = 392216$

    (d) Em média, houve um decréscimo anual de 17592 explorações agrícolas no período de 1979 a 2007.

12. x^* - temperatura média do mês de Julho (ºC); y^* - produção de milho (kg/m^2)

	x^*	y^*
média	23.98316 (ºC)	0.43539 (kg/m^2)
variância	2.758195 (ºC^2)	0.01319 (kg^2/m^4)
desvio padrão	1.660782 (ºC)	0.1148330 (kg/m^2)
covariância	-0.1105795 (ºC kg/m^2)	
coef. correlação	-0.5798233	
recta de regressão	$y^* = 1.39690(\text{kg/m}^2) - 0.04009(\text{kg/m}^2/^\text{o}\text{C})\, x^*$	