ESTATÍSTICA

Caderno de Exercícios de Estatística Descritiva

com soluções

2024/25

Estatística Descritiva Univariada

Notas:

1. Os dados referidos em alguns exercícios encontram-se em ficheiros que estão comprimidos (*zipados*) no ficheiro `https://fenix.isa.ulisboa.pt/downloadFile/281547991179894/DadosEstatDescr.zip`.
2. Os exercícios marcados com ✓ são de resolução fortemente recomendada.
3. Os exercícios marcados com ⊗ não poderão ser resolvidos com a matéria lecionada em 2024/25.
4. Em 2023/24 foi incluída no *curriculum* dos cursos do ISA a UC **Introdução à Programação** onde se leciona uma introdução à linguagem Python. Por razões de continuidade esta linguagem substitui o R.

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

a) Estado civil 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 golos, por jogo, de uma equipa de futebol.

✓ **1.2.** Para cada um dos conjuntos de dados apresentados abaixo, classifique a variável em estudo e construa uma tabela de frequências absolutas, relativas, relativas acumuladas e absolutas acumuladas. Faça uma representação gráfica de cada tabela de frequências. Os dados encontram-se nos ficheiros "nematodes.txt", "laranjas.txt" e "ovelhas.txt".

a) Conjunto 1 – número de nemátodes 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.

20.4	21.24	21.78	22.42	22.58	23.3	23.4	23.85
24.48	25.4	25.59	25.83	25.85	26.22	26.42	26.83
26.86	27.12	27.38	27.58	27.58	27.91	28.22	28.4
28.82	29.11	29.33	29.46	29.6	29.7	30.14	30.15
30.28	30.35	30.39	30.4	30.7	30.83	31.1	31.2
31.54	31.84	32.55	33.2	33.3	33.35	33.4	33.54
33.84	34.15	34.26	35.08	36.15	36.54	38.42	39.47

✓ **1.3.** Para 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 o valor dos seguintes somatórios:

i) $\sum_{i=1}^{5} x_i$ ii) $\sum_{i=3}^{5} x_i$ iii) $\sum_{i=1}^{5} x_3$

iv) $\sum_{j=1}^{5} 2x_j$ v) $\sum_{i=2}^{5} x_i - 4$ vi) $\sum_{i=1}^{5} x_i^2$

vii) $(\sum_{i=1}^{5} x_i)^2$ viii) $\sum_{i=1}^{5} x_i y_i$ ix) $\left(\sum_{i=1}^{5} x_i\right)\left(\sum_{i=1}^{5} y_i\right)$.

✓ **1.4.** 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.

a) Qual foi a produção total (peso em kg) daquele período?

b) Exprima o valor da produção total daquele período em toneladas.

c) 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?

d) 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 no caso de se decidir:

   i) retirar aquele registo;

   ii) substituir aquele registo por um valor que o viticultor considerou mais verosímil (450kg);

   iii) por um valor escolhido por si. Justifique a sua escolha.

e) Se em média, por dia, foram recolhidos 515 kg de uvas, qual o peso médio de uvas recolhidas numa hora de trabalho (1 dia = 8 h de trabalho).

f) Complete a seguinte afirmação:
Se x designa o valor de uma variável registado em kg/dia, e x' designa o valor da mesma variável registado em g/hora, então tem-se $x' = \ldots\ldots x$ (1 dia = 8 h de trabalho).

✓ **1.5.** Considere dois conjuntos de dados: o primeiro contendo o registo efectuado durante 50 dias, do número de casos de intoxicação ocorridos, em cada dia, 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
Nº de dias	13	15	8	6	5	2	1

Misturas de café	A	B	C	D	E
Nº de respostas	190	210	180	205	215

a) Indique a variável considerada em cada um dos casos e classifique-a.

b) Determine uma medida de localização adequada a cada um dos conjuntos de dados.

c) Indique o valor do mínimo e do máximo de cada conjunto de dados, caso existam.

✓ **1.6.** 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

a) Diga qual a variável em estudo e classifique-a.

b) Apresente os dados numa tabela de frequências relativas e faça a representação gráfica adequada. Comente.

c) Descreva a amostra indicando medidas de localização central e de dispersão.

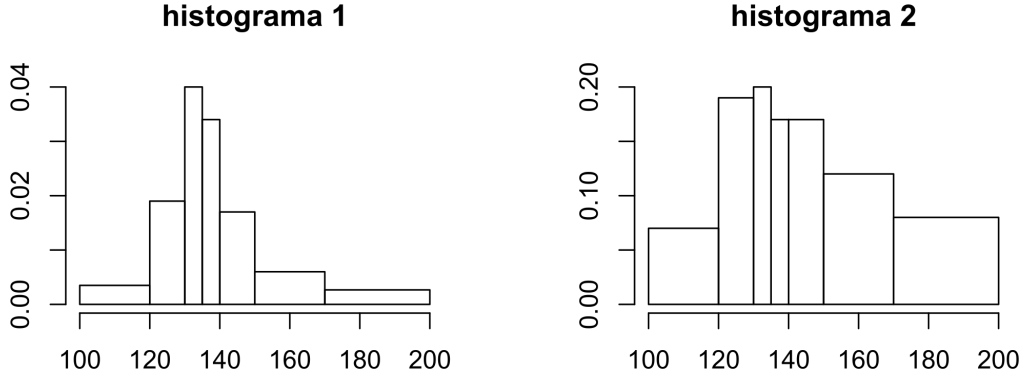
d) Qual é a percentagem de passeios em que no máximo se observaram 2 golfinhos?

1.7. 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[
nºde árvores	7	19	20	17	17	12	8

a) Elabore uma tabela de frequências relativas e frequências relativas acumuladas, dos valores apresentados.

b) Indique, justificando, qual dos histogramas apresentado a seguir se pode considerar o mais adequado para descrever o agrupamento de dados apresentado?



⊗ c) Determine a mediana, aproximada, da altura das árvores observadas.

1.8. De um total de N números há uma proporção p de 1's e uma proporção $q = 1 - p$ de 0's.

a) Calcule a média do conjunto dos N números.

b) Supondo N grande prove que o desvio padrão é aproximadamente $\sqrt{pq}$.

1.9. 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 estão disponíveis no ficheiro "serras.csv".

a) Agrupe os dados em classes. Faça a sua representação gráfica.

b) Comente a distribuição das altitudes das serras.

c) Averigue se há candidatos a "outlier" no conjunto dos dados.

1.10. 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

a) Construa um histograma para os dados da precipitação e comente-o.

b) Obtenha a caixa-de-bigodes dos dados e comente-a.

c) Calcule a precipitação média e mediana diária em Lisboa, naquele mês. Compare os valores obtidos da média e da mediana e comente, tendo em atenção que ambos são indicadores de localização.

d) Complete as seguintes afirmações:

i) Se x designa o valor da precipitação em mm por dia e x' designa o mesmo valor da precipitação expressa em cm por dia então tem-se $x' = \ldots\ldots x$.

ii) Se x designa o valor da precipitação em mm por dia e x' designa o mesmo valor da precipitação expressa em cm por hora, então tem-se $x' = \ldots\ldots x$.

e) Introduza os dados no Python e responda às questões anteriores utilizando esta linguagem de programação.

1.11. 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".

a) Construa uma tabela de frequências e faça a representação gráfica do número de pêras em cada pereira daquele pomar.

b) Qual a produção média e a produção mediana de cada pereira? Calcule ainda o desvio padrão da produção de cada pereira. Comente.

c) Construa a caixa de bigodes dos dados apresentados.

1.12. 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 $\bar{x}_A = 235$ g e $s_A^2 = 254$ g^2; recolheu 20 douradas no viveiro **B** e obteve $\bar{x}_B = 245$ g e $s_B^2 = 267$ g^2.

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

b) Dê a resposta à alínea anterior se os dados fossem registados em kg.

1.13. Pretende-se analisar o consumo diário de energia de um agregado familiar, durante os meses de Inverno. Para isso, durante 18 dias, registaram-se valores da temperatura média diária, x, em ^{o}C e do consumo diário de energia, y, em kWh. Considere as seguintes intruções Python assim como o resultado

```
>>> y = [4.3, 4.3, 5.3, ..., 5.6, 4.5, 5.9, 5.8, 4.6]
>>> hist_energia =  plt.hist(y, bins=[4,5,5.5,6,7])
>>> print('hist_energia[0]:', hist_energia[0])
>>> print('hist_energia[1]:', hist_energia[1])
```

Resultado:

```
hist_energia[0]: [7. 3. 6. 2.]
hist_energia[1]: [4.  5.  5.5 6.  7. ]
```

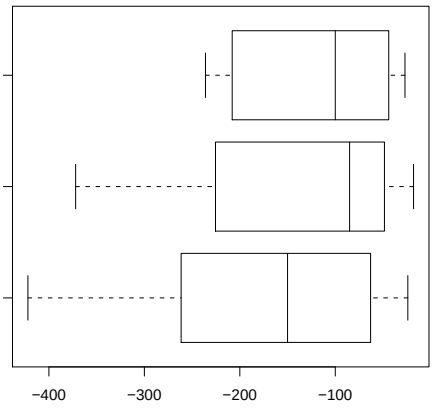
a) Elabore a tabela de frequências absolutas e relativas associada à divisão em classes apresentada no *output*.

b) Esboce o histograma associado à tabela da alínea anterior.

⊗ c) Calcule os valores aproximados da média e da mediana do consumo diário de energia.

1.14. Numa experiência medem-se fluxos de calor de meia em meia hora, das 7 h às 18 h (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 estão disponíveis no ficheiro "FluxoCalor.csv".

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



✓ a) Associe cada diagrama ao respectivo dia. Justifique.

✓ b) Sem fazer contas, pela observação dos dados e do diagrama, parece-lhe que a média correspondente ao diagrama do topo será inferior ou superior a -100. Justifique.

c) Considere agora o conjunto das observações nos 3 dias.

i) Calcule os indicadores de localização e dispersão destes dados.

ii) Desenhe o *boxplot* dos dados e compare com os que lhe são fornecidos.

iii) Construa uma tabela de frequências para dados agrupados em classes de amplitude 50.

iv) Use a tabela da alínea anterior para calcular valores aproximados da média e da mediana das observações nos três dias. Comente os resultados.

1.15. (*Exame 16.01.2012*) Registaram-se os atrasos nas chegadas, em minutos, em voos europeus num dado dia do mês de Julho de 2009 no aeroporto da Portela. Os dados recolhidos em 100 voos foram organizados na seguinte tabela:

Atraso	]0;10]	]10;20]	]20;30]	]30;40]	]40;50]	]50;60]
Nº de voos	24	23	33	9	7	4

a) Identifique e classifique, justificando, a variável em estudo.

⊗ b) Determine valores aproximados da média, desvio padrão e coeficiente de variação da variável em estudo.

⊗ c) Determine valores aproximados dos quartis da variável em estudo. Poderão existir *outliers* na amostra observada? Justifique.

1.16. (*Teste 08.11.2017*) Os seguintes dados, que se apresentam classificados na tabela de frequências relativas, f_i, apresentada abaixo, referem-se ao peso das bagagens individuais numa amostra de 100 passageiros que embarcaram no aeroporto de Lisboa num dado voo.

Peso da bagagem (kg)	[0, 10[	[10, 15[	[15, 20[	[20, 30[
f_i	0.1	A	0.3	B

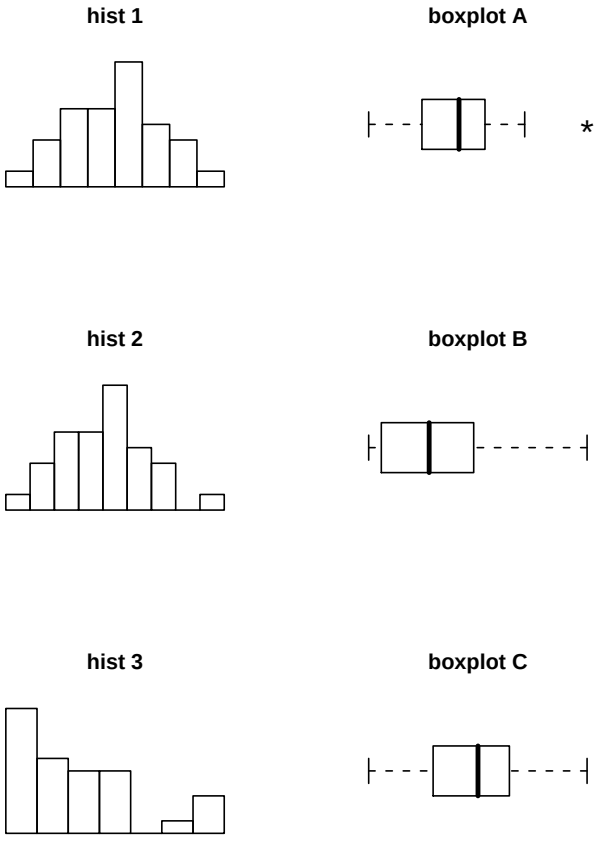
✓ a) Qual a variável em estudo? Classifique-a, justificando.

⊗ b) Sabendo que 18% dos passageiros levava bagagem com peso igual ou inferior a 12 kg, calcule os valores de A e B. Interprete o valor de A.

✓ c) Represente graficamente o histograma relativo à distribuição do peso das bagagens individuais apresentada. [Considere $A = 0.2$.]

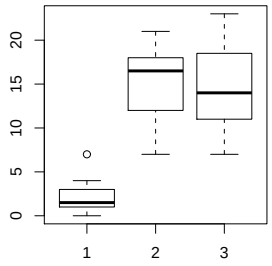
⊗ d) Calcule, aproximadamente, o peso médio das bagagens individuais.

✓ **1.17.** 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.



1.18. (*Exame 25.01.2010*) 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 e diagrama seguintes.

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



✓ a) 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.

✓ **1.19.** (*Teste 11.11.2015*) Considere os seguintes valores referentes ao registo do teor diário de vitamina C no sumo de melões (em mg/dose de sumo) na época da colheita. Considere os resultados das instruções Python:

```
y=[13.6,14.1,14.6,14.8,15.1,15.6,15.8,16.1,16.2,16.4,16.6,17.1,
   17.9,18.9,20.8] # teor de vitamina C

print(len(y))     print(sum(y))      print(sum(np.array(y)**2))
15                243.6              4005.82

print(stat.quantiles(y))
[14.8, 16.1, 17.1]
```

a) Classifique, justificando, a variável em estudo.

b) Calcule o coeficiente de variação dos valores observados de vitamina C.

c) Faça a representação gráfica do *boxplot* dos dados, apresentando os cálculos necessários.

d) Construa um histograma considerando os dados classificados em 5 classes. Comente-o tendo em conta o *boxplot* obtido na alínea anterior.

e) Numa segunda medição do teor diário de vitamina C constatou-se que o último valor registado, 20.8, estaria incorrecto.

Actualize os valores dos resultados apresentados acima se for decidido:

i) substituí-lo pelo penúltimo valor registado;

ii) eliminá-lo.

✓ **1.20.** (*Teste 7.11.2018*) Foi obtida uma amostra de 20 maçãs da variedade Fuji produzidas no ISA. Os pesos (em gramas) das 20 maçãs foram introduzidos numa lista em Python. Considere os seguintes resultados:

```
>>> x = [ ]  # lista com os pesos observados
>>> x.sort()
>>> print(x)
[79, 120, 122, 126, 126, 130, 130, 132, 150, 155, 158, 161, 170,
 173, 174, 174, 176, 180, 190, 191]

>>> print(sum(x))            >>> print(sum(np.array(x)**2))
3017                         471249
```

a) Calcule a variância e o desvio padrão das observações.

b) Calcule e compare a média e a mediana do peso das maçãs observadas. A assimetria sugerida por esses valores poderá reflectir a existência de uma "cauda" das observações "mais pesada" à direita ou à esquerda?

c) Esboce o histograma das observações considerando as classes]75,125],]125,150],]150,175], e]175,200].

d) O ISA procedeu à venda deste variedade de maçãs pelo preço de 1 €por kg. Determine a média e o desvio padrão da nova variável "preço de uma maçã" em euros.

1.21. (*Exame 9.1.2019*) Um fruticultor colheu uma amostra de 40 frutos do seu ginjal e avaliou o teor de sólidos solúveis (em ^{o}Brix) nos frutos. Os dados observados foram introduzidos no Python, executaram-se alguns comandos e obteve-se o seguinte *output*:

```
>>> Brix = [ lista ]
>>> Brix.sort()
>>> print(Brix)
[11.8, 16.0, 16.9, 17.0, 17.7, 17.8, 18.6, 18.7, 18.7, 18.7,
 18.8, 18.9, 19.5, 19.6, 20.1, 20.1, 20.2, 20.3, 20.7, 20.8,
 20.8, 21.0, 21.0, 21.2, 21.2, 21.4, 21.7, 21.8, 21.9, 22.0,
 22.4, 22.5, 23.1, 23.2, 23.7, 24.7, 25.5, 25.5, 25.7, 26.7]
```

```
>>> print(sum(Brix))          >>> print(sum(np.array(Brix*Brix)))
827.9                         17466.09

>>> print(stat.quantiles(Brix))
[18.725, 20.8, A]

>>> print(stat.variance(Brix))
B

>>> print(stat.stdev(Brix)/stat.mean(Brix)*100)
C
```

a) Complete os valores A, B e C, em falta no *output*.

b) Qual a designação da característica numérica cujo valor é C?

c) Faça a representação gráfica do *boxplot* dos dados observados, apresentando os cálculos necessários. Comente-o.

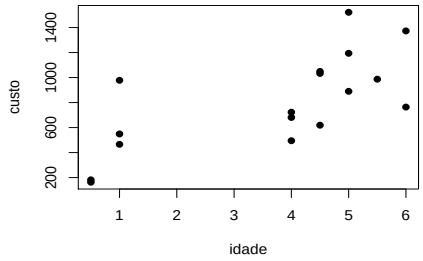
d) 1^{o}Brix é igual a 1g de sólidos solúveis/100 g de solução. Pretende-se registar os valores na seguinte escala "g de sólidos solúveis/1000 g de solução". Indique que valores tomariam agora A, B e C.

✓ **1.22.** (*Exame 14.01.2013*) Pensa-se que o custo de manutenção (em euros) de tractores aumenta com a idade (em anos) do tractor. Para modelar a relação entre o custo e a idade registaram-se observações de idade de 17 tractores e respectivo custo de manutenção. Os dados foram introduzidos nas listas `idade` e `custo` em Python.

Apresentam-se abaixo alguns resultados obtidos.

```
>>> custo = [lista]
>>> custo.sort()
>>> print(custo)
[163, 182, 466, 495, 549, 619, 681, 723, 764, 878, 890,
987, 1033, 1049, 1194, 1373, 1522]

>>> idade = [lista]
>>> idade_df = pd.DataFrame(idade) # converter para data frame
>>> idade_df.describe()
count    17.000000
mean      3.824000
std       1.802266
min       0.500000
25%       3.000000
50%       4.500000
```

```
75%       5.000000
max       6.000000

>>> custo_df=pd.DataFrame(custo)
>>> print(custo_df.describe())
count    17.000000
mean    798.117647
std     377.463058
min     163.000000
25%          A
50%          B
75%    1033.000000
max    1522.000000

>>> stat.correlation(idade, custo)
0.7831099

>>> plt.scatter(idade, custo)
>>> plt.xlabel("idade")
>>> plt.ylabel("custo")
>>> plt.show()
```



a) Classifique a variável "custo de manutenção". Justifique. Face ao objetivo do estudo trata-se da variável resposta ou da variável independente?

b) No *output* acima faltam dois valores (A e B). Calcule-os.

c) Esboce a caixa-de-bigodes do custo de manutenção. Comente.

d) Determine a variância da idade dos tractores se os valores observados fossem convertidos para meses.

e) (**) Poder-se-á admitir a existência de uma relação linear entre as variáveis? Justifique. Independentemente da sua resposta determine a equação da recta de regressão que modela a relação entre as variáveis em estudo.

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

g) (**) Qual a variação anual média do custo de manutenção estimada pela recta de regressão quando os tractores têm idade entre 0.5 e 6 anos?

h) (**) Determine os coeficientes da recta de regressão se a idade fosse converida para meses. Justifique.

(**) – Alínea a resolver na secção "Estatística Descritiva Bivariada".

Estatística Descritiva Bivariada

1.23. A tabela seguinte mostra os valores do índice de preços ao consumidor (IPC) em Portugal nos últimos anos, considerando 2010 como ano base.

Ano(x)	2010	2011	2012	2013	2014	2015	2016	2017
IPC(y)	100	103.65	106.52	106.81	106.51	107.03	107.68	109.16

a) Calcule $cov(x,y)$, a média e a variância de x e de y.

b) Se aos anos de observação, x, se tivesse aplicado a transformação $2(x-2009)$, i.e., se se considerasse os anos representados por 2, 4, ..., 16, qual seria o valor de $cov(x',y)$, com $x' = 2(x-2009)$?

c) 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.

d) Independentemente das respostas anteriores, determine a equação da recta de regressão de y sobre x.

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

f) Qual a variação anual média dos preços, estimada pela regressão, no período 2010-2017?

g) Mantendo-se a actual tendência, qual prevê que seja o IPC em 2018?

h) Quais seriam os valores dos coeficientes da recta de regressão se o índice de preços ao consumidor tivesse como base o ano de 2011, i.e. $y'_i = 100y_i/y_2$?

✓ **1.24.** Para $n = 20$ pares de observações (x_i, y_i), seja $y = -5.6 + 0.7\,x$, a equação da recta de regressão dos mínimos quadrados de y sobre x.

a) Comente as seguintes afirmações:

i) O coeficiente de correlação entre y e x é igual ao simétrico do coeficiente de correlação entre x e y.

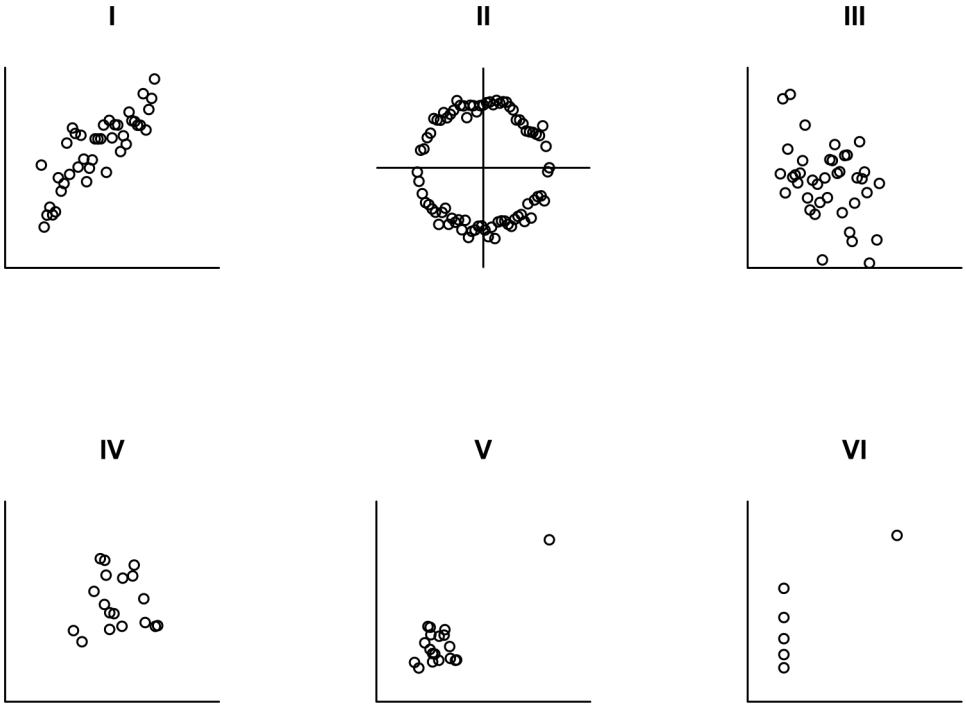
ii) O coeficiente de correlação entre y e x é positivo porque o declive da recta é positivo.

iii) Em média, quando x aumenta y não aumenta, pois o declive da recta é menor do que 1.

b) Sendo $\sum_{i=1}^{20} x_i = 200$ determine $\sum_{i=1}^{20} y_i$.

✓ **1.25.** Diga 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



✓ **1.26.** 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

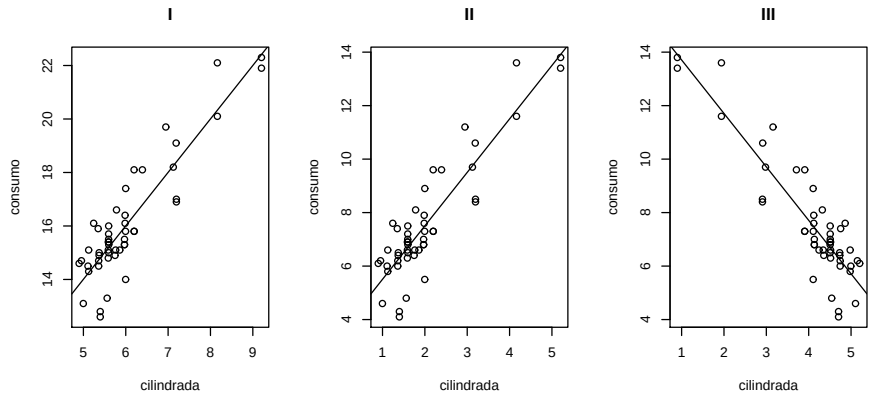
$$y = 3.5 + 2x$$

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:

a) Determine a média e o desvio padrão dos consumos de gasolina dos automóveis observados.

b) 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.

1.27. (*Exame 7.01.2010*) 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:

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

a) Classifique, justificando, a variável x - "humidade relativa ambiente".

b) Poder-se-á admitir a existência de uma relação linear entre as variáveis? Justifique.

c) 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.

d) 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?

1.28. A medição directa do calor específico de ramos de macieira é difícil de efectuar. Um investigador propõe predizer 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.csv") 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

a) Desenhe o diagrama de extremos e quartis para os valores do calor específico observados. Comente a distribuição dos dados.

b) 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.

c) Qual o valor que se prevê para o calor específico quando a percentagem de água é de 60? Qual o resíduo associado a essa observação? Justifique.

d) 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?

1.29. (*Exame 26.01.2015*) A nuvem de pontos de $n = 22$ pares de observações (x_i, y_i), $i = 1, \ldots, 22$, tem centro $(\bar{x}, \bar{y}) = (2.5; 9)$. Foi ajustada uma recta de regressão de y sobre x pelo método dos mínimos quadrados. O resíduo associado ao ponto $(5; 7.5)$ é $e_i = 3.5$.

a) Mostre que a equação da recta de regressão é $y = 14 - 2x$.

b) Sabendo que $\sum_{i=1}^{22}(x_i - \bar{x})^2 = 46.25$ e $s_y^2 = 10.25$, calcule

i) s_x^2;

ii) $cov(x, y)$;

iii) o coeficiente de determinação da recta de regressão.

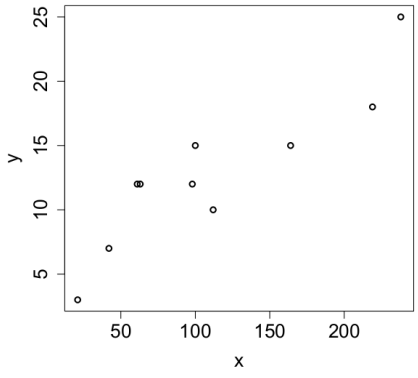
c) Considere a transformação afim $x' = 10 + 0.5x$. Determine os coeficientes da recta de regressão de y sobre x'.

✓ **1.30.** 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 em Python apresentando-se abaixo os resultados obtidos:

```
>>> x = [219,238,42,98,100,164,21,112,61,63]
>>> y = [18,25,7,12,15,15,3,10,12,12]

>>> stat.mean(x)    >>> stat.mean(y)
111.8               12.9

>>> stat.variance(x)       >>> stat.variance(y)
5394.622                   36.1

>>> stat.covariance(x,y)
396.7556
```



a) Parece-lhe adequada a existência de uma relação linear entre x e y? Justifique.

b) 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.

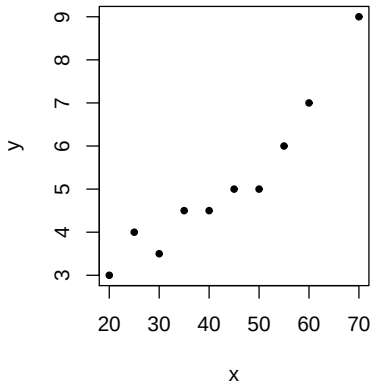
c) Interprete, no contexto do problema, o significado do coeficiente de regressão de y sobre x.

d) 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 25 anos.

e) 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?

✓ **1.31.** (*Teste 23.11.2010*) Num estudo em que se pretende avaliar a influência da velocidade do vento na quantidade de água evaporada por dia na albufeira de uma barragem, obtiveram-se os seguintes dados que foram introduzidos no Python.

```
x = [20, 50,30,55,70,45,60,25,40,35]  #vel. vento (km/h)
y = [3,5,3.5,6,9,5,7,4,4.5,4.5]       #agua evaporada (centenas de litro)
stat.mean(x)
43
stat.mean(y)
5.15
stat.variance(x)
256.6667
stat.variance(y)
3.169444
stat.covariance(x,y)
27
```



Com base nos resultados apresentados, responda às seguintes questões.

a) Parece-lhe adequada a existência de uma relação linear entre x e y? Justifique.

b) 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.

c) 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 registados em m/s. Qual será a precisão desta recta? Justifique. (**Verifique que** 1 km/h = 0.2778 m/s).

d) Determine $cov(x', y)$, sendo x' a variável em m/s.

1.32. 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    : 3.0011   Var    : 3.0314   Var    : 1.8047   Var    :3188.0763

> 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) i) Sendo (x_i, y_i), $i = 1, \ldots, n$, pares de observações das variáveis (x, y), mostre que

$$s^2_{(x-y)} = s^2_x + s^2_y - 2cov(x, y).$$

ii) Determine o coeficiente de correlação entre as variáveis *NLdir* e *NLesq*. Comente o valor obtido.

c) Ajustou-se uma recta de regressão de área foliar (*Area*) sobre comprimento da nervura principal (*NP*), tendo-se obtido a equação $Area = -137.951 + 28.927NP$.

i) Qual a variação esperada na área foliar associada a um aumento de 1 mm no comprimento da nervura principal, estimada pela regressão?

ii) 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 às observações recolhidas nesta folha?

iii) Determine a precisão da recta de regressão e interprete o valor obtido.

1.33. (*Exame 31.01.2012*) Realizou-se um estudo para averiguar a percentagem (P) de resíduos sólidos eliminados por um sistema de filtragem em função da taxa (T) de fluxo de efluente. No quadro abaixo encontram-se alguns resultados observados:

Taxa de fluxo de efluente (T)	1	4	5	6	8	10	12
Percentagem de resíduos sólidos (P)	24	19	18	•	14	•	10

a) Identifique a variável dependente e a variável independente.

b) Considere os resultados obtidos em Python, apresentados abaixo, para responder às seguintes perguntas:

i) Parece-lhe admissível a existência de uma relação linear entre as variáveis? Porquê?

ii) Escreva a equação da recta de regressão dos mínimos quadrados ajustada aos dados. Interprete, no contexto do problema, o significado do coeficiente de regressão.

```
>>> T=[1, 4, 5, 6, 8, 10, 12]
>>> P=[24, 19, 18, 17.5, 14, 12, 10]
>>> print(stat.correlation(T,P)**2)
0.9882260724725246
>>> print(stat.linear_regression(T,P))
LinearRegression(slope=-1.25938566552901, intercept=24.63310580204778)
```

1.34. 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:

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

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

a) Determine a recta de regressão dos mínimos quadrados da evolução da produção de trigo em função do tempo. Indique a sua precisão.

b) Se se decidisse identificar os anos por $1, \ldots, 9$, respectivamente, qual seria a precisão da recta de regressão que se obteria considerando esta transformação? Justifique convenientemente.

1.35. (*Exame 12.01.2015*) Numa estação meteorológica medem-se os valores diários de precipitação (em mm) com dois instrumentos diferentes (A e B). O técnico responsável da estação escolheu ao acaso 18 dias em que existem registos dos dois instrumentos, introduziu os dados em duas listas, `precA` e `precB` no Python e obteve os seguintes resultados:

```
stat.mean(precA)        stat.mean(precB)       stat.covariance(precA,precB)
12.09286                13.72222               0.9956992

stat.variance(precA)        stat.variance(precB)
194.3647                    297.523
```

O instrumento B mede a precipitação de forma mais correcta mas avaria-se com frequência. Com o objectivo de prever o valor da precipitação medida pelo instrumento B, nos dias em que só existem dados do instrumento A, o técnico responsável decidiu realizar uma regressão linear com os dados.

a) Parece-lhe adequada a escolha de uma recta para modelar a relação entre as observações diárias efectuadas por cada um dos instrumentos? Justifique.

b) Indique qual deverá ser a variável resposta e qual a variável preditora nesta recta de regressão.

c) Obtenha a recta de regressão dos mínimos quadrados indicando a sua precisão. Interprete essa precisão.

d) Num dia em que a precipitação medida pelo instrumento A foi 19.25 mm, quanto se prevê que fosse a quantidade medida pelo instrumento B.

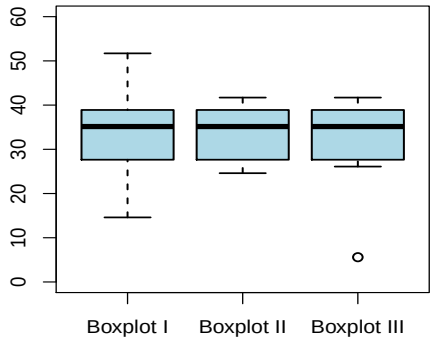
e) Suponha que os dados eram registados em dm. Qual a variação esperada nos registos do instrumento B, quando a precipitação diária registada em A aumenta 1 dm? Justifique.

✓ **1.36.** (*Exame 25.01.2016*) Num estudo realizado em pereiras, determinou-se colorimetricamente o teor de clorofila, x, em 12 folhas e obtiveram-se os seguintes resultados (1 unidade=10 mg/m^2):

min	1º Quartil	2º Quartil	3º Quartil	max	$\bar{x}$	s_x^2
24.6	27.825	35.150	38.500	41.7	33.96	37.6481

a) Classifique, justificando, a variável em estudo.

b) Qual o *boxplot* que melhor corresponde aos dados apresentados? Justifique.



c) O teor de azoto foliar (em g/kg de peso seco) relaciona-se linearmente com o teor de clorofila medido colorimetricamente segundo a seguinte equação de recta de regressão dos mínimos quadrados $y = 2.3195 + 0.5224x$, em que x é o teor de clorofila medido colorimetricamente e y o teor de azoto foliar.

i) Sabendo que 96% da variabilidade de y é explicada pela regressão, determine o coeficiente de correlação entre x e y. Interprete-o.

ii) Qual o resíduo associado ao par (32.6, 18.1)?

iii) Determine a média e a variância do teor de azoto foliar.

iv) Suponha que se decide expressar o teor de azoto foliar em g/100 g de peso seco. Considerando esta unidade determine:

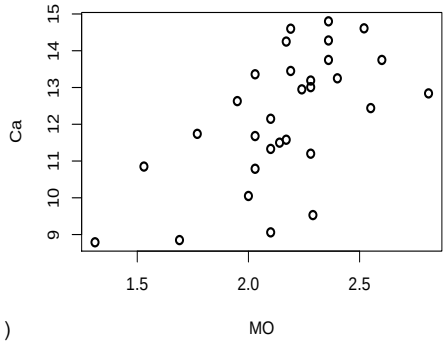
A. o coeficiente de correlação entre o teor de clorofila e o teor de azoto foliar;

B. o declive da recta de regressão.

1.37. (*Exame 10.01.2017*) No programa de desenvolvimento de uma dada região com solos de natureza calcária, pretende fazer-se um estudo para o estabelecimento adequado de pastagens. Recolheram-se amostras de solo em vários locais e mediu-se o valor de várias características importantes no estudo da fertilidade do solo, produtividade e resistência à erosão. De entre elas vamos considerar a matéria orgância (MO) e o cálcio de troca (Ca).

Considere os valores de medições de MO (em %) e de Ca (em cmol(+)/kg), registados em 30 locais. Os dados foram introduzidos no Python e executaram-se alguns comandos, cujos resultados se encontram abaixo:

```
>>> Ca = [lista]
>>> MO = [lista]
>>> MO.sort()
>>> print(MO)
[ 1.31 1.53 1.69 1.77 ...   2.55 2.60 2.81]

>>> print(stat.quantiles(MO))
[2.03,  2.18,   2.36]

>>> hist_MO = plt.hist(MO, bins=list(np.arange(1.2, 3.3, 0.4)))
>>> print(hist_MO[0])
[2.   4.   20.  3.  1.]

>>> print(hist_MO[1])
[1.2   1.6   2.0   2.4   3.2]

>>> print(stat.covariance(MO, Ca))
0.6088

>>> print(stat.linear_regression(MO, Ca))
LinearRegression(slope=3.47, intercept=4.71)
```



a) Utilizando os resultados apresentados no *output*, construa a tabela de frequências absolutas e relativas para os valores observados de MO, considerando como limites das classes 1.2, 1.6, 2.0, 2.4, 2.8 e 3.2 e esboce o histograma associado.

b) Esboce o *boxplot* para os dados de MO, apresentando os procedimentos que necessitou de utilizar.

⊗ c) Calcule valores aproximados para a média e para a variância de MO.

d) Pretende-se verificar se existe alguma relação entre Ca e MO.

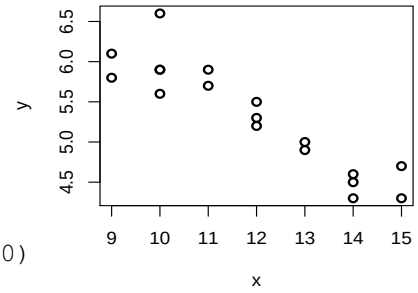
i) Poder-se-á admitir a existência de uma relação linear entre as variáveis? Justifique.

ii) Num dado local observou-se $(2.00, 10.05)$ para o par (MO, Ca). Independentemente da resposta à alínea anterior determine o resíduo associado àquele par resultante de ajustar a recta de regressão dos mínimos quadrados aos dados.

iii) Calcule o declive da recta de regressão se Ca for expresso em cmol(+)/(100g).

✓ **1.38.** (*Exame 26.01.2018*) Pretende-se modelar a relação entre o consumo diário de energia de um agregado familiar e a temperatura média diária, durante o Inverno. Para isso, durante 18 dias, registaram-se valores da temperatura média diária, x, em ^oC e do consumo diário, y, em kWh. Os dados foram introduzidos no Python. Para este estudo considere o *output* seguinte:

```
>>> y = [4.3, 4.3, 5.3, ..., 5.6, 4.5, 5.9, 5.8, 4.6]

>>> x = [15, 14, 12, ..., 14, 10,  9, 14]


>>> print(stat.covariance(x,y)**2)
0.842407

>>> print(stat.linear_regression(x, y))
LinearRegression(slope=-0.3092, intercept=8.9980)
```



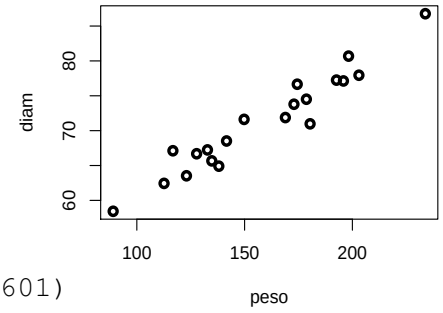
a) Parece-lhe admissível a existência de uma relação linear entre as variáveis consideradas? Justifique.

b) Independentemente da resposta à alínea anterior, escreva a equação da recta de regressão dos mínimos quadrados. Interprete, no contexto do problema, o valor do coeficiente de regressão.

c) Determine o coeficiente de regressão no caso do consumo diário ser registado em W.

✓ **1.39.** (*Teste 7.11.2018*) As maçãs Fuji são pagas aos agricultores em função do seu calibre (diâmetro, em mm). Nas centrais fruteiras, os calibradores electrónicos recorrem ao peso dos frutos para estimar os seus diâmetros. Para estabelecer a relação entre pesos (em gramas) e diâmetros (em mm), pesaram-se e mediram-se 20 maçãs obtidas no mercado. Introduzidos os dados no Python e obtiveram-se os seguintes resultados:

```
>>> peso = [...]
>>> diam = [...]
>>> print(stat.correlation(peso,diam)**2)
0.9298733

>>> print(stat.linear_regression(peso, diam))
LinearRegression(slope=0.1815, intercept=42.4601)
```



a) Escreva a equação da recta de regressão ajustada. Qual a precisão dessa recta? Comente.

b) Parece-lhe adequada a utilização do modelo linear para descrever a relação entre o diâmetro e o peso dos frutos analisados? Justifique.

c) Sabendo que a média e o desvio padrão do peso dos frutos são 158.25 g e 37.09 g, respectivamente, determine a média e o desvio padrão do diâmetro dos frutos observados.

d) De acordo com o modelo ajustado, qual é a variação esperada no diâmetro dos frutos associada ao aumento de 1 grama no peso?

e) Determine a recta de regressão dos mínimos quadrados do diâmetro sobre o peso, caso os pesos fossem registados em kg. Qual a precisão desta recta?

✓ **1.40.** (*Teste 13.11.2019*) Em 23 parcelas de inventário florestal de montado de sobro pesou-se a cortiça extraída e reportou-se o seu valor, em kg de matéria seca por ha. Também se registou o diâmetro máximo das árvores na parcela (cm).

a) Considere os dados referentes ao diâmetro máximo das árvores (cm) na parcela, que foram introduzidos na lista `diam.max`, em Python, e tenha em conta os cálculos efectuados para responder às questões abaixo:

```
>>> diam_max = [...]
>>> diam_max.sort()
>>> print(diam_max)  #Nota: os valores representados por * foram omitidos
[30.80,   *,   *,   50.45,   52.20,   54.43,  56.02,    *,    *,    58.57,
 59.45,  60.16,   61.75,  62.39,   *,    *,    67.30,   74.96,  76.08,  *,
 90.72,  111.86,  127.32]

>>> stat.quantiles(diam_max)
[54.43,    B,    74.96]

>>> hist_diam = plt.hist(diam_max, bins=[30,50,60,80,100,130])
>>> print(hist_diam[0])
[3.   8.   8.  2.  C]
```

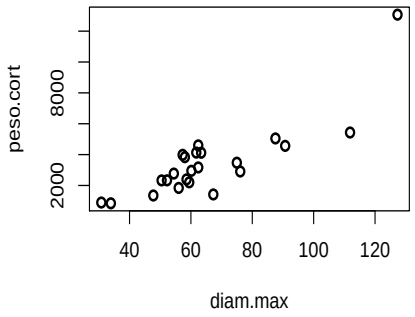
i) Determine os valores B e C, em falta no *output*.

ii) Esboce o *boxplot* dos dados apresentando os cálculos necessários.

iii) Esboce o histograma dos dados considerando as classes indicadas na instrução `plt.hist`.

b) Pretende-se verificar se é possivel prever o peso da cortiça, em kg por ha, como função do diâmetro máximo das árvores na parcela, em cm. Para os dados de cada uma das 23 parcelas da questão anterior efectuaram-se os seguintes cálculos para as variáveis `peso_cort` e `diam_max`

```
>>> stat.variance(diam_max)       >>> stat.variance(peso_cort)
486.5658                          6073647

>>> print(stat.linear_regression(diam_max, peso_cort))
LinearRegression(slope=94.08, intercept=-2690.36)
```



i) Escreva a equação da recta de regressão dos mínimos quadrados do peso da cortiça sobre o diâmetro máximo das árvores na parcela e interprete o significado do declive.

ii) Considera ter sido adequada a escolha de uma relação linear entre o peso da cortiça e o diâmetro máximo das árvores na parcela? Justifique convenientemente.

iii) O peso de cortiça é geralmente expresso em arrobas (1 arroba(@) = 15 kg). Qual seria o declive da recta indicada em a) se o peso da cortiça fosse registado em arrobas? Justifique convenientemente.

1.41. 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

Estes dados podem ser disponibilizados em Python através da instrução

```
ans_df = pd.read_csv('https://query.data.world/s/6p2ntncvkzj5mnvbpkaswfilryvnrk')
print(ans_df)
```

a) Calcule as médias e as variâncias de cada uma das oito variáveis. Comente.

b) Calcule os coeficientes de correlação entre as variáveis x e as variáveis y de cada um dos quatro pares de variáveis. Comente.

c) Calcule as rectas de regressão de y sobre x para cada um dos quatro pares de variáveis (x_i,y_i), (i=1,...,4).

d) Construa as quatro nuvens de pontos correspondentes aos pares de variáveis utilizados nas duas alíneas anteriores. Comente, à luz dos resultados das alíneas anteriores.

e) Construa os gráficos dos resíduos para cada conjunto de pares de observações e comente-os.

Exercícios de Revisão de Estatística Descritiva

R1.1. Diga **justificando** se são verdadeiras ou falsas as afirmações que se seguem:

a) A amplitude interquartil é metade da amplitude total.

b) A média está sempre entre o primeiro e o terceiro quartil.

c) A mediana está sempre entre o primeiro e o terceiro quartil.

d) O desvio padrão é sempre igual à amplitude interquartil.

e) O desvio padrão é menor do que a média dos desvios relativos à média.

R1.2. Considere n pares de observações (x_i, y_i). Seja $z_i = \dfrac{x_i - \overline{x}}{s_x}$, $i = 1, \cdots, n$.

a) Mostre que as observações z_i têm média nula e variância unitária.

b) Determinou-se a recta de regressão dos mínimos quadrados de y sobre x tendo-se obtido $y = 3 + 1.5x$. Sabendo que $\overline{y} = 10.5$ e $s_x^2 = 0.25$, determine a equação da recta de regressão de y sobre z.

R1.3. Considere n pares de observações (x_i, y_i).

a) Sendo $\overline{x}$ a média das observações x_i, mostre que a média de $z_i = x_i/k - m$, (k e m números reais, $k \neq 0$ e $i = 1, \ldots, n$) é $\overline{z} = \overline{x}/k - m$.

b) Sejam $y = b_0 + b_1 x$ a recta de regressão dos mínimos quadrados de y em x e $\hat{y}_i$ os valores estimados pela recta, correspondentes aos valores observados x_i. Mostre que o coeficiente de determinação, R^2, é igual ao quadrado do coeficiente de correlação, r, isto é,

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

R1.4. 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

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

a) Pretende-se ajustar uma recta de regressão aos dados obtidos. Parece-lhe admissível tal ajustamento? Justifique convenientemente.

b) 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.

c) 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.

R1.5. 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 $(y = b_0 + b_1 x)$, calculada a partir dos dados das Estatísticas Agrícolas do INE, num dado ano, tem os seguintes coeficientes: $b_0 = 13.1$; $b_1 = 0.32$.

a) Comente as seguintes afirmações:

i) Em média, quando a superfície territorial aumenta, não aumenta a superfície florestal, pois b_1 é menor do que 1;

ii) O coeficiente de correlação entre a superfície florestal e a superfície territorial tem de ser positivo, porque o coeficiente b_0 é positivo.

b) Sendo a superfície territorial total do Continente 8892.7 milhares de hectares, diga qual a superfície florestal total.

R1.6. 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: $\sum_{i=1}^{22} x_i^2 = 590754 \qquad \sum_{i=1}^{22} y_i^2 = 16288 \qquad \sum_{i=1}^{22} x_i y_i = 97547$

a) Calcule a distância média entre os assentos e os cotovelos e a altura média dos indivíduos observados.

b) Calcule a mediana da variável altura.

c) Estime um modelo de regressão linear simples da distância (y) em função da altura das pessoas. Indique a precisão da recta e comente-a.

d) 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.

e) Qual o aumento esperado para a distância entre o assento e o cotovelo por cada aumento unitário na altura de uma pessoa?

R1.7. Dados n pares de observações (x,y), seja $y = b_0 + b_1 x$ a recta de mínimos quadrados ajustada.

a) Defina coeficiente de correlação, $r_{x,y}$ e indique uma sua propriedade.

b) Sendo $s_x^2 = 5.1$; $b_1 = -3$; $\overline{x} = 3$; $\overline{y} = 2.8$ e $r^2 = 0.92$ determine s_y e a equação da recta de regressão.

c) Prove que o declive da recta é invariante quando se efectua uma mesma transformação de escala a ambas as variáveis.

R1.8. (*Exame 17.01.2011*) 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 cm$\times$dia.

a) Averigue 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.

b) Poder-se-á admitir a existência de uma relação linear entre as variáveis? Justifique.

c) 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?

d) 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.

e) 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).

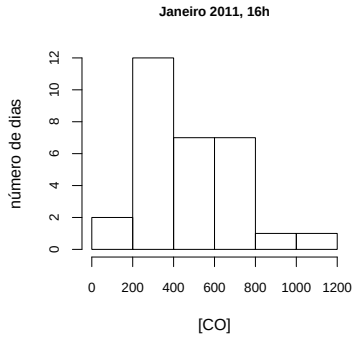
R1.9. Para oito pares de observações $\{(x_i, y_i)\}_{i=1}^{8}$ determinou-se a recta de regressão dos mínimos quadrados, $y = 2.45 - 1.2x$, cuja precisão é 0.9604. Responda, **justificando convenientemente**, se são verdadeiras ou falsas as afirmações nas seguintes alíneas.

a) Sabendo que $\sum_{i=1}^{8} x_i = 15$ então $\overline{y} = 0.2$.

b) O coeficiente de correlação é $r = 0.98$.

c) Se $s_x^2 = 0.5522$ então $s_y^2 = 0.828$.

R1.10. (*Exame 27.01.2014*) O gráfico mostra o histograma dos 30 registos diários disponíveis da concentração de CO, μg/m^3, medida às 16h, em Janeiro de 2011, numa avenida de Lisboa.



a) Identifique e classifique, justificando, a variável em estudo.

b) Elabore a tabela de frequências absolutas, relativas e relativas acumuladas associada ao histograma.

c) Determine valores aproximados para os seguintes indicadores:
i) mediana ii) média.

d) Em cada dia de registo da concentração de CO registou-se também o número de automóveis que passaram nessa avenida entre as 15h30 e as 16h. Com os dados disponíveis obteve-se a seguinte recta de regressão dos mínimos quadrados da concentração de CO (y) sobre o número de automóveis (x): $y = 468.25 + 0.021\,x$.

i) Calcule um valor aproximado do número total de automóveis registado.

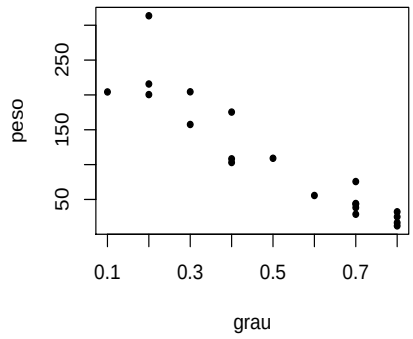
ii) Interprete, no contexto do problema, o valor do coeficiente de regressão.

iii) Num dos dias registou-se $x = 282$, ao qual se verificou estar associado um resíduo igual a -12.37, para o modelo de regressão dado. Qual o valor observado para a concentração de CO naquele dia?

R1.11. (*Teste 12.11.2014*) Pretende-se estudar a influência da infecção de um certo fungo no peso de alfaces em estufa. Para cada uma de 20 alfaces obteve-se o peso em gramas e registou-se o grau de infecção numa escala de 0 a 1 (sendo 0 – inexistência de infecção e 1 – infecção muito elevada). Estes valores foram introduzidos em Python. A seguir apresentam-se alguns dos comandos e respectivos resultados:

```
>>> peso = [...]
>>> peso.sort()
>>> print(peso)
[12.1, 16.7,  25.1,  28.8,  32.3,
 38.3, 43.3,  44.2,  55.7,  75.7,
103.0, 108.3, 109.1, 157.6, 175.4,
200.5, 204.3, 204.5, 215.7, 313.6]

>>> stat.mean(peso)      >>> stat.mean(grau)
108.21                   0.515
```

```
>>> stat.variance(peso)
7439.075

>>> stat.variance(grau)
0.05818421

>>> stat.covariance(peso,grau)
[1] -19.30121
```



a) Calcule os indicadores necessários para construir o *boxplot* do peso das alfaces. Faça a sua representação gráfica.

b) Será de admitir a existência de uma relação linear entre as variáveis? Justifique.

c) Independentemente da resposta dada na alínea anterior, pretende-se ajustar uma recta de regressão aos dados.

   i) Qual é a variável resposta?

   ii) Obtenha a equação da recta de regressão dos mínimos quadrados para as variáveis em estudo. Calcule a precisão desta recta e interprete-a.

d) Se os valores do peso das alfaces forem registados em kg determine justificando:

   i) A variância da variável peso das alfaces.

   ii) O declive da recta de regressão das variáveis expressas nas novas unidades.

R1.12. (*Teste 9.11.2016*) O teor de fenóis na casca das batatas varia com a duração do ciclo da batateira. Para determinar a data de colheita óptima, que permitirá maior período de conservação, avaliou-se o teor de fenóis em diferentes dias do ciclo cultural da batateira. Considere o *output* em Python com os valores registados do teor de fenóis (em μg/g de casca fresca) assim como alguns cálculos.

```
>>> fenois = [...]
>>> fenois.sort()
>>> print(fenois)
[23.022, 23.236, 24.003, 24.501, 24.793, 25.321, 26.013, 26.568, 29.372]

>>> print(stat.quantile(fenois))
[24.003,   A,     26.013]

>>> print(stat.mean(fenois))
25.20322
```

a) Classifique, justificando, a variável "teor de fenóis".

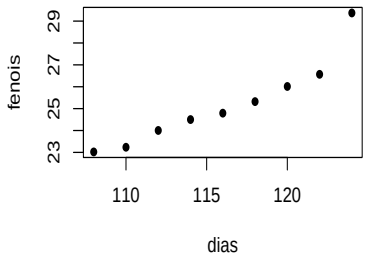
b) Determine o valor de A, em falta no *output*.

c) Esboce a caixa-de-bigodes dos dados, apresentando os cálculos necessários.

d) Sem efectuar cálculos altere uma única observação da lista `fenois` de modo que, para a nova lista de dados, os resultados apresentados em `stat.quantiles(fenois)` se mantenham e a média aumente. A caixa-de-bigodes do novo conjunto de dados alterou-se? Justifique.

e) Considere mais alguns resultados, obtidos em Python, sobre as listas `fenois` e `dias`:

```
>>> print(stat.mean(dias))
116
>>> print(stat.mean(fenois))
25.20322
```




```
>>> print(stdev(dias))
5.477226
>>> print(stdev(fenois))
1.957057
```

```
>>> print(stat.covariance(dias,fenois))
10.059
```

i) Poder-se-á admitir a existência de uma relação linear entre as variáveis? Justifique.

ii) Independentemente da resposta na alínea anterior, determine a equação da recta de regressão dos mínimos quadrados que modela a relação entre o teor de fenóis acumulados na casca dos tubérculos e a duração do ciclo, em dias.

iii) Sabe-se que, na variedade de batata em estudo, o teor mínimo de fenóis, à colheita, recomendado é 28 μg/g de casca. Qual deve ser a menor duração do ciclo que o agricultor deve respeitar para prever atingir aquele valor de teor de fenóis?

iv) Determine, justificando, o declive da recta de regressão se o teor de fenóis for registado em mg/g de casca fresca (1mg=$10^3\mu$g).

R1.13. (*Teste 22.11.2013*) Considere n observações de uma variável x. Designe por $x = \{x_1, \ldots, x_n\}$ essas n observações e por $\overline{x}$ a média dessas observações. Designe por $x' = \{x_1, \ldots, x_n, \overline{x}\}$ o novo conjunto das $n+1$ observações.

a) Mostre que as médias de x e de x' são iguais.

b) Estabeleça a relação existente entre s_x^2 e $s_{x'}^2$.

R1.14. (*Teste 23.11.2010*) Considere os dois quadros seguintes. O primeiro sistematiza os resultados de um estudo sobre a opinião dos alunos acerca da qualidade das refeições que lhes são servidas numa dada cantina e o segundo refere-se ao número de avarias em 200 elevadores de um dado fabricante no período de 5 anos.

Qualidade da refeição	Fraca	Normal	Boa	Muito boa
Nº de alunos	3	18	22	7

Nº de avarias	0	1	2	3	4	5
Nº de elevadores	84	52	31	17	10	6

a) Indique qual a variável considerada em cada um dos quadros e classifique-a, justificando.

b) Determine uma medida de localização adequada a cada um dos conjuntos de dados.

c) Para os dados do segundo quadro calcule a variância e o coeficiente de variação.

R1.15. As colunas do ficheiro "Cor.csv") 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.

R1.16. (*Exame 9.1.2020*) O grau de verdura da vegetação herbácea de um Montado foi medido, em 62 datas, com recurso a um espectroradiómetro de campo e a uma máquina fotográfica digital vulgar. Com estes dois métodos foram obtidos dois índices (adimensionais), NDVI e GCC, respectivamente.

a) Os valores de NDVI foram introduzidos em Python e obtidos os seguintes resultados:

```
>>> hist_NDVI = plt.hist(NDVI, bins=[0.1,0.2,0.3,0.5,0.8])

>>> print(hist_NDVI[0])
[11. 14. 18. 19.]

>>> print(hist_NDVI[1])
[0.1  0.2  0.3  0.5  0.8]
```

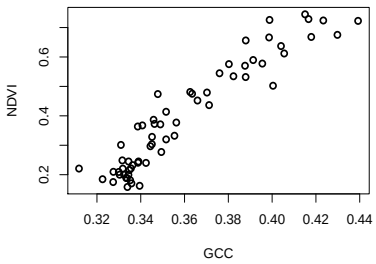
i) Construa a tabela de frequências absolutas, relativas e relativas acumuladas para os dados de NDVI, considerando as classes apresentadas no *output*.

ii) Determine valores aproximados da média, variância e do primeiro quartil dos valores obtidos para NDVI.

iii) Esboce o histograma dos dados considerando as classes apresentadas no *output*.

b) Pretende-se averiguar se a utilização do espectroradiómetro, um aparelho muito caro, pode ser substituída pela utilização de uma máquina fotográfica digital. Os valores de NDVI e GCC observados nas 62 datas foram introduzidos em duas listas em Python. Considere os seguintes resultados:

```
>>> stat.correlation(GCC,NDVI)
0.945354

>>> stat.mean(GCC)          >>> stat.mean(NDVI)
0.36                        0.39

>>> stat.variance(GCC)      >>> stat.variance(NDVI)
0.00099                     0.03422
```



i) Considera admissível a existência de uma relação linear entre GCC e NDVI? Justifique convenientemente.

ii) Independentemente da resposta dada em i) escreva a equação da recta de regressão dos mínimos quadrados de NDVI sobre GCC. Determine a sua precisão e interprete-a.

iii) Qual o valor previsto para NDVI quando se obtem o valor de GCC igual a 0.4?

Soluções ou resoluções dos Exercícios de Estatistica Descritiva

1.1. a) Qualitativa nominal

b) Quantitativa contínua

c) Quantitativa discreta

d) Qualitativa ordinal

e) Quantitativa contínua

f) Quantitativa discreta

1.2. Comandos Python para construção das tabelas de frequência e respetivas representações gráficas.

a)

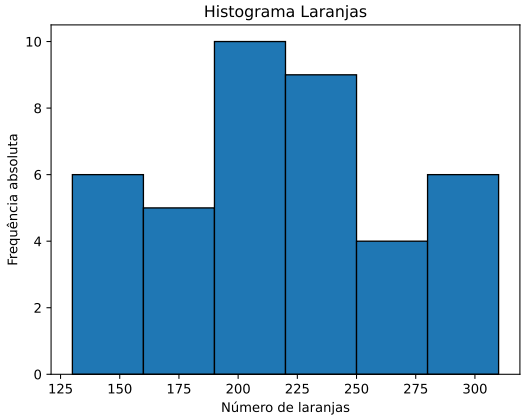
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# nematodes
nematodes_df = pd.read_csv('nematodes.txt')
print(nematodes_df.head())
ni = nematodes_df['numero'].value_counts().sort_index() # Freq.abs.
fi = ni / ni.sum()  # Freq. relativa
Ni = ni.cumsum() # Frequência absoluta acumulada
Fi = Ni / ni.sum() # Frequência relativa acumulada
# Tabela de frequências
tabela_frequencias = pd.DataFrame({'ni': ni, 'fi': fi, 'Fi': Fi,
                                   'Ni': Ni})
print(tabela_frequencias)

        ni        fi        Fi  Ni
numero
0        3  0.050000  0.050000   3
1       12  0.200000  0.250000  15
2       17  0.283333  0.533333  32
3       13  0.216667  0.750000  45
4        9  0.150000  0.900000  54
5        6  0.100000  1.000000  60

# desenha o grafico de barras
plt.bar(ni.index, ni.values, color='skyblue',width=0.1)
plt.xlabel('Número de nematodes')
plt.ylabel('Frequência absoluta')
plt.title('Diagrama de barras')
```

```
plt.show()
```

b)
```
# laranjas
laranjas_df = pd.read_csv('laranjas.txt')
print(laranjas_df.head())
# Regra de Sturges
num_sturges = round(1 + np.log2(len(laranjas_df['numero'])))
# Amplitude aproximada de cada classe
amplitude_classe = (laranjas_df['numero'].max() -
                    laranjas_df['numero'].min()) / num_sturges
print(amplitude_classe)
# Histograma com classes definidas pelo utilizador
hist_laranjas = plt.hist(laranjas_df['numero'],
                         bins=[130, 160, 190, 220, 250, 280, 310],
                         edgecolor='black')
plt.title("Histograma Laranjas")
plt.xlabel("Número de laranjas")
plt.ylabel("Frequência absoluta")
plt.show()
# Frequências
ni = hist_laranjas[0]  # Frequência absoluta
fi = ni / sum(ni)  # Frequência relativa
Ni = np.cumsum(ni)  # Frequência absoluta acumulada
Fi = Ni / sum(ni)  # Frequência relativa acumulada
nclass = len(ni) # Número de classes
classes = hist_laranjas[1]
esq = ["]" + str(classes[i]) + "," for i in range(nclass)]
dir = [str(classes[i+1]) + "]" for i in range(nclass)]

# Tabela de frequências
tabela_frequencias = pd.DataFrame({'esq':esq, 'dir':dir,'ni': ni,
```

```
                                        'fi': fi, 'Fi': Fi, 'Ni': Ni})
print(tabela_frequencias)
```

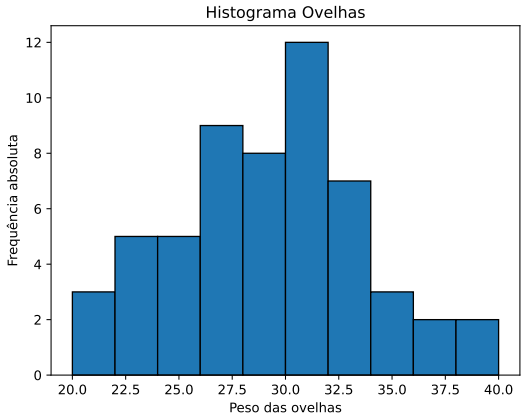
```
       esq     dir    ni     fi     Fi    Ni
0  ]130.0,  160.0]   6.0  0.150  0.150   6.0
1  ]160.0,  190.0]   5.0  0.125  0.275  11.0
2  ]190.0,  220.0]  10.0  0.250  0.525  21.0
3  ]220.0,  250.0]   9.0  0.225  0.750  30.0
4  ]250.0,  280.0]   4.0  0.100  0.850  34.0
5  ]280.0,  310.0]   6.0  0.150  1.000  40.0
```



c)
```
# ovelhas
ovelhas_df = pd.read_csv('ovelhas.txt')
print(ovelhas_df.head())
# Regra de Sturges
num_sturges = round(1 + np.log2(len(ovelhas_df['peso'])))
amplitude_classe = (ovelhas_df['peso'].max() -
ovelhas_df['peso'].min()) / num_sturges
print(amplitude_classe)
hist_ovelhas = plt.hist(ovelhas_df['peso'], bins=list(range(20, 41, 2)),
                        edgecolor='black')
plt.title("Histograma Ovelhas")
plt.xlabel("Peso das ovelhas")
plt.ylabel("Frequência absoluta")
plt.show()
# Frequências
ni = hist_ovelhas[0]  # Frequência absoluta
fi = ni / sum(ni)  # Frequência relativa
Ni = np.cumsum(ni)  # Frequência absoluta acumulada
```

```
Fi = Ni / sum(ni)  # Frequência relativa acumulada
nclass = len(ni) # Número de classes
classes = hist_ovelhas[1]
esq = ["]" + str(classes[i]) + "," for i in range(nclass)]
dir = [str(classes[i+1]) + "]" for i in range(nclass)]
# Tabela de frequências
tabela_frequencias = pd.DataFrame({'esq':esq, 'dir':dir, 'ni': ni,
 'fi': fi, 'Fi': Fi, 'Ni': Ni})
print(tabela_frequencias)
```

```
      esq    dir    ni        fi        Fi    Ni
0  ]20.0,  22.0]   3.0  0.053571  0.053571   3.0
1  ]22.0,  24.0]   5.0  0.089286  0.142857   8.0
2  ]24.0,  26.0]   5.0  0.089286  0.232143  13.0
3  ]26.0,  28.0]   9.0  0.160714  0.392857  22.0
4  ]28.0,  30.0]   8.0  0.142857  0.535714  30.0
5  ]30.0,  32.0]  12.0  0.214286  0.750000  42.0
6  ]32.0,  34.0]   7.0  0.125000  0.875000  49.0
7  ]34.0,  36.0]   3.0  0.053571  0.928571  52.0
8  ]36.0,  38.0]   2.0  0.035714  0.964286  54.0
9  ]38.0,  40.0]   2.0  0.035714  1.000000  56.0
```



1.3. i) 38 ii) 31 iii)30 iv) 76 v) 32 vi) 390 vi) 1444 viii) 98 ix) 304

1.4. a) 7725 kg

b) 7.725 t

d) i) 550.357 kg

ii) 543.6667 kg

e) 64.375 kg=64375 g.

f) $x' = 125x$

1.5. a) 1º conjunto (quadro da esquerda): variável - nº de casos de intoxicação em cada dia - quantitativa discreta

2º conjunto (quadro 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 $mo =$ mistura tipo E.

c) Só é possível indicar o mínimo e o máximo para o 1º conjunto:
mínimo - 0 casos/dia; máximos 6 casos/dia

1.6. a) variável - nº de golfinhos em cada passeio - variável discreta porque toma valores em $\mathbf{N}_0$.

c) média $\bar{x} = 2.28$, $mo = 2$ e $\tilde{x} = 2$ golfinhos/passeio.
Medidas de dispersão: variância $s^2 = 1.73$ (golfinhos/passeio)^2 e desvio padrão $s = 1.31$ golfinhos/passeio.

d) Pretende-se $F(2) = \frac{17+45+84}{235} = 0.6213$, frequência relativa acumulada, logo é 62%

1.7. b) As classes dadas têm amplitudes diferentes e como a área de cada rectângulo deverá representar a frequência relativa associada à respectiva classe, isto é, $A_i = f_i$ e $A_i = h_i \times alt_i \Rightarrow alt_i = \dfrac{f_i}{h_i}$, h_i amplitude da classe i e alt_i altura do rectângulo, com base na classe i.

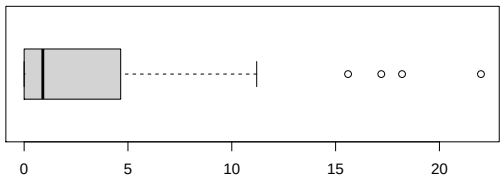
Tem-se, por exemplo $A_1 = f_1 = 0.07$, $A_2 = f_2 = 0.19$, $A_3 = f_3 = 0.2$, etc., donde
$alt_1 = \dfrac{0.07}{20} = 0.0035$; $alt_2 = \dfrac{0.19}{10} = 0.019$; $alt_3 = \dfrac{0.2}{5} = 0.04$, vemos que esta última altura é quase o dobro da segunda, portanto só poderá ser o **histograma 1**.

c) 1º calcular classe mediana que é $[135;140[$.
$me \approx 136.1765$ dm.

1.8. a) $\bar{x} = \dfrac{Np \times 1 + N(1-p) \times 0}{N} = p$

b) $s = \sqrt{s^2} = \sqrt{\dfrac{Np \times 1^2 + N(1-p) \times 0^2}{N} - p^2} = \sqrt{p - p^2} = \sqrt{pq}$

1.10. b)



c) $\bar{x} = 4.116$ mm; mediana=0.9 mm ... forte assimetria, média muito superior à mediana pois é muito influenciada por valores extremos à direita

d) i) $x' = 10^{-1}x$ ii) $x' = x/240$

1.12. a) Para as 35 douradas tem-se $\overline{x} = 240.71$ g e $s^2 = 279.0042$ g^2 (quando se faz arredondamentos em cálculos intermédios pode obter-se resultados um pouco diferentes – por ex. $s^2 = 281.13$ g^2)

1.13. a)
```
hist_energia =  plt.hist(y, bins=[4,5,5.5,6,7])
ni = hist_energia[0]
fi = ni / sum(ni)
nclass = len(ni)
classes = hist_energia[1]
esq = ["]" + str(classes[i]) + "," for i in range(nclass)]
dir = [str(classes[i+1]) + "]" for i in range(nclass)]
alt = [fi[i]/(classes[i+1]-classes[i]) for i in range(nclass)]
print(pd.DataFrame({'esq':esq, 'dir':dir, 'ni': ni,
 'fi':fi, 'altura_i':alt}))
```

```
     esq   dir  ni        fi  altura_i
0  ]4.0,  5.0]   7  0.388889  0.388889
1  ]5.0,  5.5]   3  0.166667  0.333333
2  ]5.5,  6.0]   6  0.333333  0.666667
3  ]6.0,  7.0]   2  0.111111  0.111111
```

c) $\overline{x} \simeq 5.2639$ kWh ; mediana $\simeq 5.3235$ kWh

1.14. 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.

1.15. a) A variável em estudo é o "atraso nas chegadas", em minutos. Variável quantitativa contínua.

c) $Q_1 \simeq 10 + 10\frac{0.25-0.24}{0.23} = 10.43$; $Q_2 \simeq 20 + 10\frac{0.5-0.47}{0.33} = 20.91$ e $Q_3 \simeq 20 + 10\frac{0.75-0.47}{0.33} = 28.48$
$AIQ \simeq Q_3 - Q_1 = 28.48 - 10.43 = 18.05$,
barreiras inferiores e superiores, aproximadamente,
$Q_1 - 1.5 \times AIQ = -16.645$ e $Q_3 + 1.5 \times AIQ = 55.55$
Então, não existirão valores muito pequenos que possam ser *outliers* (a barreira inferior, aproximada, é negativa), mas podem existir valores elevados que sejam candidatos a *outliers*, visto a barreira superior, embora aproximada, estar contida na última classe observada.
(ver resolução detalhada na colectânea de exames)

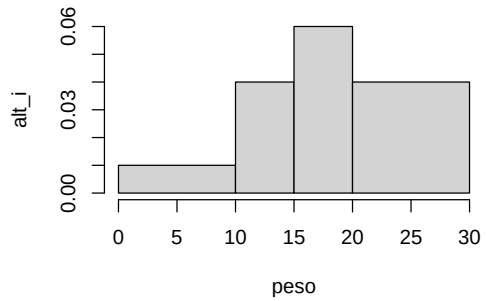
1.16. a) "peso das bagagens" (em kg); é quantitativa contínua .

b) 12 kg é o quantil de ordem 0.18, i.e. $Q^*_{0.18} = 12$; este quantil pertence ao intervalo $[10, 15[$, então
$12 = 10 + (15-10)\dfrac{0.18-0.10}{A} \Longrightarrow A = 0.2.$
$\sum_i^4 f_i = 1 \Rightarrow B = 1 - 0.1 - 0.2 - 0.3 = 0.4.$

c) Como as classes da distribuição apresentada do peso das bagagens têm amplitudes diferentes, a altura de cada rectângulo $i\,(i = 1,2,3,4)$ no histograma deve ser $alt_i = f_i / amplitude_i$.

Temos então as alturas alt_i assim dadas:

$$alt_1 = \frac{0.1}{10} = 0.01; \quad alt_2 = \frac{0.2}{5} = 0.04; \quad alt_3 = \frac{0.3}{5} = 0.06; \quad alt_4 = \frac{0.4}{10} = 0.04$$



d) $\overline{x} \simeq 18.25$ kg.

(ver resolução detalhada na colectânea de exames)

1.17. O histograma 1 corresponde ao *boxplot* C; o histograma 2 corresponde ao *boxplot* A e o histograma 3 corresponde ao *boxplot* B.

1.18. a) Spray A $--> $ *boxplot* 3; Spray B $--> $ *boxplot* 2; Spray C $--> $ *boxplot* 1

b) **Medidas de localização**

Sprays	média ($\bar{x}$)	mediana ($\tilde{x}$)	1º quartil (Q_1)	3º quartil (Q_3)
Spray A	14.5	14	11.5	17.8
Spray B	15.3	16.5	12.5	17.5
Spray C	2.08	1.5	1	3

O spray B é o que mata, em média, mais insectos. O spray A apresenta mais simetria, o que pode também ser visto no *boxplot* 3 e é também sugerido pelos valores próximos da média e da mediana, $\bar{x}_A \simeq \tilde{x}_A$

Medidas de dispersão

Sprays	Ampl. total	Ampl. interquartis (AIQ)	variância (s_x^2)	desvio padrão (s_x)
Spray A	16	6.3	22.27	4.72
Spray B	14	5	18.24	4.27
Spray C	7	2	3.9	1.98

Quanto à dispersão verifica-se que o spray A apresenta menor variância, os resultados são mais semelhantes entre si. Apresenta por isso maior previsibilidade nos resultados.

c) a totalidade das observações é $n = 36$. Para calcular a média necessitamos de
$\sum_{i=1}^{36} x_i = 174 + 184 + 25$ donde $\bar{x} = 10.63$ mortes
Para calcular a variância do total devemos usar a expressão

$$s_x^2 = \frac{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n(n-1)} = \frac{36(2768+3022+95) - (174+184+25)^2}{35 \times 36} = 51.72\,\text{mortes}^2$$

c) Podemos construir a tabela (que fica para completar):

classes	n_i	f_i	F_i
$[0;5]$	11	11/36	11/36
$]5;10]$	5	5/36	16/36
$]10;15]$	9		
$]15;20]$	8		
$]20;25]$	3		

e com base nela, para estes dados agrupados temos:

$\bar{x} \approx 10.69$ mortes

$s_x^2 \approx 44.65$ mortes^2

mediana $\equiv Q_{0.5}^* \approx 11.11$ mortes

1.19. a) A variável em estudo é o "teor de vitamina C no sumo de melões"(mg/dose); é quantitativa contínua porque assume valores numéricos em $\mathbf{R}$, mais precisamente em $\mathbf{R}^+$.

b) Seja y o teor diário de vitamina C.

O coeficiente de variação é $CV_y = \frac{s_y}{\bar{y}} \times 100\%$ $\quad CV_y = 11.61\%$

c) Para representar o *boxplot* dos dados necessitamos dos seguintes valores (expressos em mg/dose de sumo), obtidos directamente do *output*:

mínimo: 13.6 ; máximo: 20.8; mediana: $\tilde{y} = Q_2$=16.1

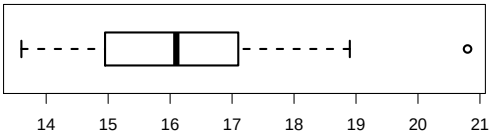
1º quartil: Q_1= 14.8 e 3º quartil: Q_3=17.1

Barreiras superior e inferior:

$BS = Q_3 + 1.5(Q_3 - Q_1) = 20.55$ observa-se que há um valor na amostra, 20.8, que é superior a BS; logo é um candidato a *outlier*.

$BI = Q_1 - 1.5(Q_3 - Q_1) = 11.35$ todos os valores da amostra são superiores à barreira inferior, portanto não há nenhum candidato a *outlier*.

O *boxplot* é:



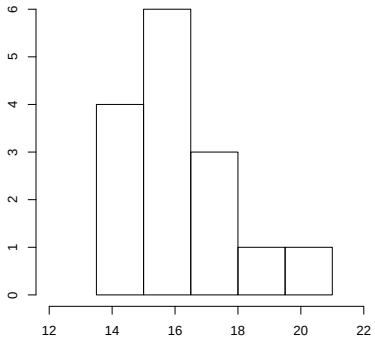
d) Vamos considerar a amplitude, h, constante; para construir o histograma com 5 classes, tem-se

$h \approx \frac{max(x_i) - min(x_i)}{5} = 1.44$; usemos $h = 1.5$ mg/dose de sumo

e vamos construir a tabela das frequências absolutas necessária para representarmos o histograma (note-se que, como usamos classes com a mesma amplitude, podemos marcar as frequências absolutas)

classes	n_i
$[13.5; 15[$	4
$[15; 16.5[$	6
$[16.5; 18[$	3
$[18; 19.5[$	1
$[19.5; 21[$	1

n_i– frequência absoluta.



Verificamos que, embora a divisão em classes e possivelmente a escolha das classes não permita mostrar que há um valor que se destaca do conjunto dos dados, o candidato a *outlier* marcado no *boxplot*, ainda assim o histograma apresenta uma cauda mais prolongada do lado direito.

e) i) As duas últimas observações passam a ser: 18.9 e 18.9.

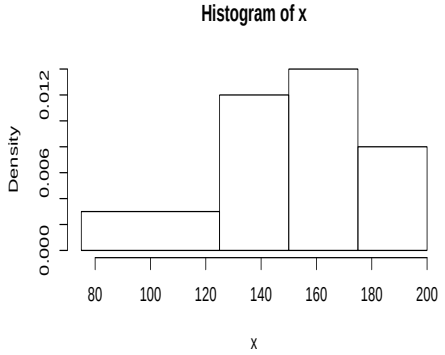
```
>>> len(y)     >>> sum(y)      >>> sum(np.array(y)**2)
15             241.7          3930.39

>>> stat.quantiles(y)
[14.8,    16.1,    17.1]
```

os quantis não sofrem alteração pois continuamos com a amostra com a mesma dimensão e como o máximo depois de alterado mantém a mesma posição, i.e. continua a ser a observação $x_{(15)}$, nenhum quartil é alterado.

1.20. a) $s_x^2 = 849.1868\ \text{g}^2$ e $s_x = 29.14$ g.

b) $\bar{x} = 150.85$ g e $\tilde{x} = 156.5$ g. A média é inferior à mediana, manifestando a existência de valores inferiores mais extremos.

c) Temos classes com amplitudes diferentes

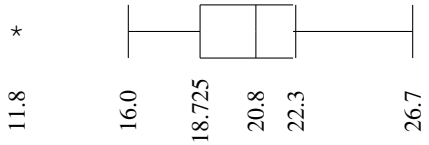


d) Seja $y = 10^{-3}x$ a variável "preço de uma maçã" em euros.
$\bar{y} = 0.15085$ euros e $s_y = 0.02914$ euros.
(soluções na colectânea de exames)

1.21. a) $A = 22.3\ ^{\text{o}}\text{Brix}$, $B = 8.4776\ ^{\text{o}}\text{Brix}^2$, $C = 14.0676\%$.

b) É o *coeficiente de variação*.

c) Barreira inferior é 13.363 e a barreira superior é 27.6625.



O diagrama sugere que a distribuição das observações é razoavelmente simétrica, sendo o valor mínimo um *outlier*.

d) Designando os novos valores respectivamente por A', B' e C', obtêm-se $A' = 10 \times A$, $B' = 10^2 \times B$ e $C' = C$.

1.22. a) Variável quantitativa contínua pois pode tomar qualquer valor numérico num intervalo $\mathbb{R}_0^+$. No contexto do problema é a variável resposta, varia em função da idade.

b) A= 549, B= 764

c) não há candidatos a *outlier*; os extremos dos bigodes são min e max.

d) s_x^2=467.735 meses^2

e) $r = 0.783$ relativamente próximo de 1, que corresponde à relação linear perfeita, e nuvem de pontos sugere a possibilidade de traçar uma recta por entre os pontos, podemos assumir relação linear

$y = 170.914 + 164.013x$

f) $r^2 = 0.613$ apenas 61.3% da variabilidade do custo pode ser explicada pelo modelo

g) $b_1 = 164.013$ €/ano

h) $b_1' = 1/12 b_1 = 13.66775$; $b_0' = b_0$

(Resolução detalhada na colectânea de exames)

1.23. a) $\sum_{i=1}^8 x_i = 16108$; $\sum_{i=1}^8 y_i = 847.36$; $\sum_{i=1}^8 x_i^2 = 32433500$;
$\sum_{i=1}^8 y_i^2 = 89808.9$; $\sum_{i=1}^8 x_i y_i = 1706202$
$cov(x,y) = 6.10714$; $\bar{x} = 2013.5$; $\bar{y} = 105.92$; $s_x^2 = 6$; $s_y^2 = 8.075257$

b) $cov(x^{'}, y) = 12.21429$

d) $y = -1943.535 + 1.018x$

e) $r^2 = 0.76978$, i.e., cerca de 77% da variabilidade do IPC é explicada pela recta de regressão.

f) A variação anual média estimada é dada por $b_1 = 1.018$

g) $\hat{y}_{2018} = 110.789$

h) $b_1^{'} = 0.982$; $b_0^{'} = -1875.094$

1.24. a) i) A afirmação é falsa.
Por definição tem-se $r_{y,x} = \dfrac{cov(y,x)}{s_y s_x} = \dfrac{cov(x,y)}{s_x s_y} = r_{x,y}$, com $s_x > 0, s_y > 0$
ii) A afirmação é verdadeira. A relação $r = b_1 s_x/s_y$, $(s_x > 0, s_y > 0)$ estabelece que r e b_1 têm o mesmo sinal. Logo $b_1 > 0 \Rightarrow r > 0$.
iii) A afirmação é falsa, pois b_1 é positivo. O coeficiente b_1 representa a variação esperada para y quando x aumenta de uma unidade.

b) Uma consequência da recta dos mínimos quadrados é: $\bar{y} = b_0 + b_1 \bar{x}$, o que é equivalente a $\sum y_i = n b_0 + b_1 \sum x_i = 20 \times (-5.6) + 0.7 \times 200 = 28$.

1.25. Os valores que se aproximam mais dos coeficientes de correlação dos dados são:

para a nuvem I $->$ b); para a nuvem II $->$ a); para a nuvem III $->$ c);

para a nuvem IV $->$ a); para a nuvem V $->$ b); para a nuvem VI $->$ b).

1.26. 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 litros por 100 km.

c) Gráfico II.

1.29. b) i) $s_x^2 = 2.2024$;
ii) $cov(x,y) = -4.4048$
iii) $R^2 = r^2 = 0.8595$

b) Seja $y = b_0' + b_1' x'$ a recta de regressão de sobre y sobre x'. $b_1' = \dfrac{b_1}{0.5} = -4$ e $b_0' = b_0 - 20b_1 = 54$.

1.30. 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, $b_1 = 0.0735$, 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 este previsão, pelo que, a grande diferença entre o valor ajustado e o valor real (25 anos) é justificável.

e) $y = 4.677 + 2.205x'$.

1.31. Ver resolução detalhada na colectânea de exames

1.32. a) 10.65 mm.

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

1.34. a) $y = 7043.93 - 3.383\ x$; a precisão da recta é $R^2 = 0.485$.

b) A mesma precisão.

1.35. a) Por consulta do *output* do R tem-se o coeficiente de correlação, $r = 0.9956992$.

b) A variável resposta (y) é `precB`; a variável preditora (x) é `precA`.

c) Equação da recta de regressão é $precB = b_0 + b_1\ precA$;
vamos referi-la como $y = b_0 + b_1\ x$.
Resposta: $y = -1.17512 + 1.231912\ x$
Precisão da recta $R^2 = r^2 = 0.9956992^2 = 0.9914$.

d) 22.5391 mm

e) Sejam x' e y' os dados em dm; então temos $x' = 0.01x$ e $y' = 0.01y$

A variação esperada nos registos do instrumento B, quando a precipitação diária registada em A aumenta 1 dm é dada pelo declive, b'_1, da recta de regressão de y' sobre x';
Resposta: $b'_1 = b_1$, i.e., quando a precipitação diária registada em A aumenta 1 dm a precipitação diária registada em B aumenta em média 1.23192 dm.

(ver resolução detalhada na colectânea de exames)

1.37. Ver resolução detalhada na colectânea de exames

1.38. Ver resolução detalhada na colectânea de exames

1.39. a) $y = 42.4601 + 0.1815\ x$. Precisão: $R^2 = 0.9299$

b) Sim, $r = 0.9643$ e a nuvem de pontos apresenta-se próximo de uma recta.

c) $\overline{y} = 71.18$ mm; $s_y = 6.98$ mm

d) Quando o peso aumenta 1 g espera-se que o diâmetro aumente 0.1815 mm.

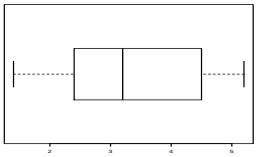
e) $y = 42.46 + 181.5\ x'$, com $x' = 10^{-3}\ x$, (x' em kg). A precisão da recta $R^2 \equiv r^2$ mantém-se dado que a variável x sofreu uma transformação afim.

1.40 Ver solução com algum detalhe na colectânea de exames

R1.2. b) Seja $y = b'_0 + b'_1 z$ recta de y sobre z
$b'_0 = \overline{y} - b'_1 \overline{z}$ e como $\overline{z} = 0 \Rightarrow b'_0 = \overline{y} = 10.5$
$b'_1 = \frac{cov(z,y)}{s_z^2}$, como $z = \frac{x}{s_x} - \frac{\overline{x}}{s_x}$ $\quad cov(z,x) = \frac{1}{s_x} cov(x,y)$ e $s_z^2 = 1$, então $\quad b'_1 = \frac{1}{s_x} cov(x,y) = 0.75$

R1.8. 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 $y_i = b_0 + b_1\ x_i + e_i$, onde e_i é o resíduo, então para $x_i = 10$ tem-se $\quad y_i = 0.713 + 0.2702 \times 10 - 0.21538 = 3.1996$ cm.

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

Sendo assim, consideremos b'_1 e b'_0 os coeficientes da recta de regressão de y' em x.

$b'_1 = \frac{cov(x,y')}{s_x^2} = \frac{0.1 cov(x,y)}{s_x^2} = 0.1\ b_1$

$b'_0 = \overline{y'} - b'_1 \overline{x} = 0.1\overline{y} - 0.1\ b_1\ \overline{x} = 0.1(\overline{y} - b_1\overline{x}) = 0.1\ b_0$.

R1.10. Ver resolução detalhada na colectânea de exames

R1.11. a) mínimo: 12.1 gramas ; máximo: 313.6 gramas

mediana: $n = 20$ é par, então $\tilde{x} = Q_2 = (x_{(10)} + x_{(11)})/2 = (75.7 + 103)/2 = 89.35$ gramas;

$1^{\underline{o}}$ quartil: $Q_1 \equiv Q^*_{0.25}$, tem-se $n \times 0.25 = 5$ (inteiro), logo $Q_1 = (x_{(5)} + x_{(6)})/2 = (32.3 + 38.3)/2 = 35.3$ gramas;

$3^{\underline{o}}$ quartil: $Q_3 \equiv Q^*_{0.75}$, tem-se $n \times 0.75 = 15$ (inteiro), logo $Q_3 = (x_{(15)} + x_{(16)})/2 = (175.4 + 200.5)/2 = 187.95$ gramas.

Não há candidatos a *outliers*

b) $r = -0.927732$.

c) i) A variável resposta é "o peso das alfaces"

ii) $y = 279.04884 - 331.7259\ x$; $R^2 = r^2 = 0.86068$

d) i) $s^2_{y'} = 10^{-6} s^2_y = 10^{-6} \times 7439.075\ \text{kg}^2$

ii) $b'_1 = \frac{cov(x,y')}{s_x^2} = \frac{10^{-3} cov(x,y)}{s_x^2} = 10^{-3} \times b_1$

R1.12. Ver resolução detalhada na colectânea de exames

R1.13. a) $\overline{x'} = \frac{x_1 + x_2 + \cdots + x_n + \overline{x}}{n+1} = \frac{n\overline{x} + \overline{x}}{n+1} = \frac{(n+1)\overline{x}}{n+1} = \overline{x}$.

b) $s^2_{x'} = \frac{\sum_{i=1}^{n+1} (x'_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 + (\overline{x} - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = \frac{(n-1)\, s_x^2}{n}$.

R1.14. Ver solução com algum detalhe na colectânea de exames

R1.16. Ver solução com algum detalhe na colectânea de exames