



UNIVERSIDADE
DE LISBOA



INSTITUTO
SUPERIOR DE
AGRONOMIA

ESTATÍSTICA

Caderno de Exercícios com algumas soluções

2026/27

Parte A - Estatística Descritiva

Nota: Os dados referidos em alguns exercícios encontram-se em ficheiros que estão comprimidos (*zipados*) no ficheiro <https://fenix.isa.ulisboa.pt/downloadFile/563022967902869/DadosEstatDescr.zip>.

- 1.1.** a) 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.

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

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

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

- b) Utilize a linguagem Python para repetir a construção da tabela de frequências e a representação gráfica. Os dados encontram-se nos ficheiros “nematodes.txt”, “laranjas.txt” e “ovelhas.txt”.

- 1.2.** Reescreva cada expressão sem usar o símbolo do somatório.

- | | | |
|-----------------------------|------------------------------|--|
| 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) $(\sum_{i=1}^5 x_i)(\sum_{i=1}^5 y_i)$ |

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

- a) Qual foi a produção total (peso em kg) daquele período?
- b) 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?
- c) 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) substituir aquele registo por um valor escolhido por si. Justifique a sua escolha.

1.4. Considere dois conjuntos de dados: o primeiro contendo o registo efetuado 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.

N ^o de casos	0	1	2	3	4	5	6	Misturas de café	A	B	C	D	E
N ^o de dias	13	15	8	6	5	2	1	N ^o 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.

1.5. Num estudo, efetuado 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:

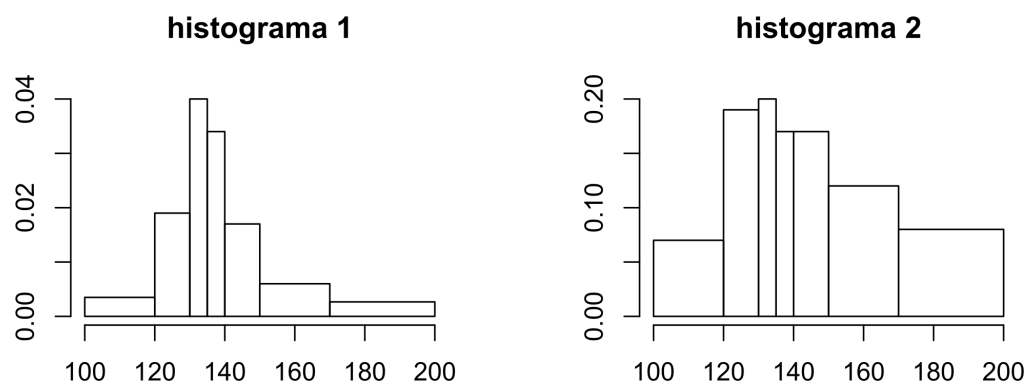
n ^o de golfinhos	0	1	2	3	4	5	6	8
n ^o de passeios	17	45	84	52	23	11	2	1

- a) Identifique e classifique a variável em estudo.
- 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?

1.6. 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[
n ^o de árvores	7	19	20	17	17	12	8

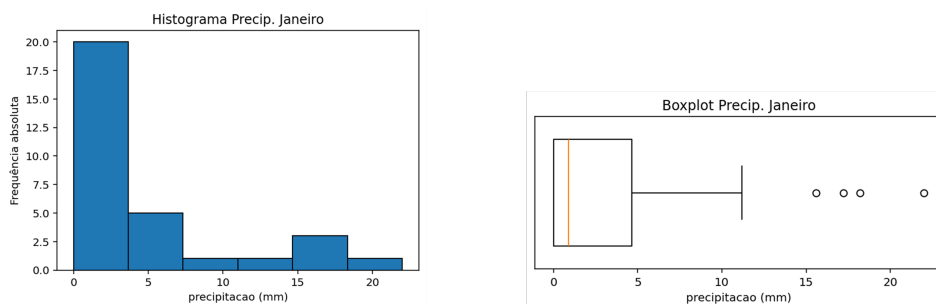
- a) Elabore uma tabela de frequências relativas e frequências relativas acumuladas, dos valores apresentados.
- b) Indique, justificando, qual dos histogramas apresentado a seguir se pode considerar o mais adequado para descrever o agrupamento de dados apresentado?



1.7. 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) Relacione o histograma com o *boxplot* dos dados de precipitação



b) As precipitações média e mediana diária em Lisboa, naquele mês, foram $\bar{x} = 4.12$ mm e $\tilde{x} = 0.9$ mm. Comente a diferença entre estes indicadores de localização, relacionando com as representações gráficas da alínea anterior.

1.8. 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 do consumo diário de energia, y , em kWh. Considere as seguintes instruções Python assim como o resultado

```
>>> y = [4.3, 4.3, 5.3, ..., 5.6, 4.5, 5.9, 5.8, 4.6]
>>> import matplotlib.pyplot as plt
```

```
>>> 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])
```

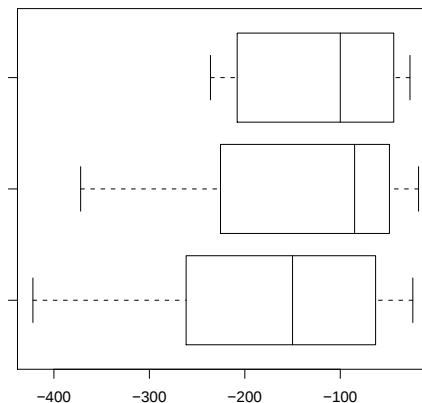
Resultado:

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

- Elabore a tabela de frequências absolutas e relativas associada à divisão em classes apresentada no *output*.
- Esboce o histograma associado à tabela da alínea anterior.

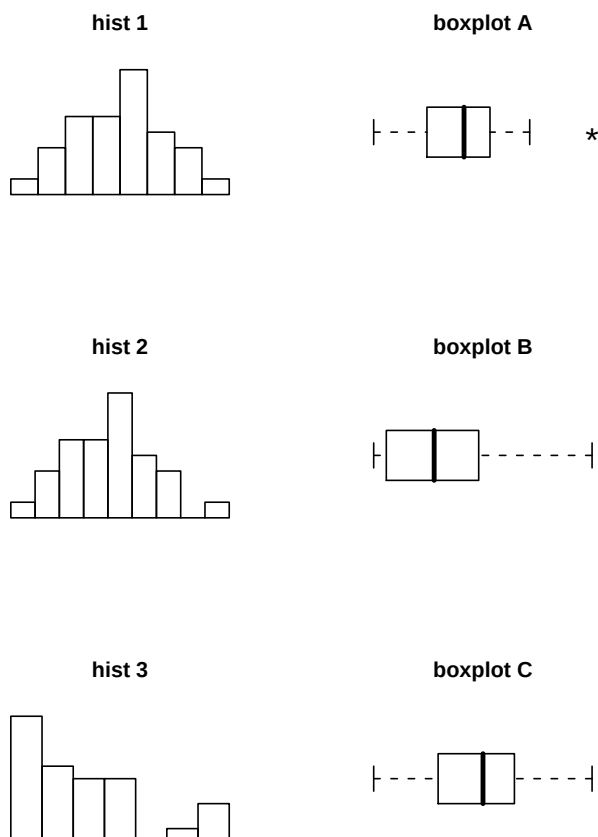
1.9. Numa experiência medem-se fluxos de calor de meia em meia hora, das 7 h às 18 h (inclusivé), durante três dias consecutivos. Os resultados obtidos (em W m^{-2}) são indicados na tabela em baixo. Ao lado da tabela estão as caixas-de-bigodes dos três dias, sem qualquer ordem aparente.

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



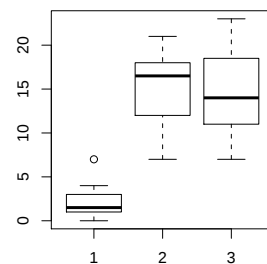
- Associe cada diagrama ao respetivo dia. Justifique.
- Sem fazer contas, pela observação dos dados e do diagrama, parece-lhe que a média correspondente ao diagrama do topo será inferior ou superior a -100. Justifique.
- Associe cada diagrama a cada valor de **mediana**: -150, -100, -85.
- Associe cada diagrama a cada valor de **média**: -173.826, -135.0, -116.826.
- Associe cada diagrama a cada valor de **amplitude total**: 209, 354, 398.
- Associe cada diagrama a cada valor de **amplitude inter-quartil**: 37, 59, 89.

- 1.10.** 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.



- 1.11.** Num estudo realizado para avaliar o efeito de três *sprays*, *A*, *B* e *C*, em insetos, organizaram-se 3 grupos de 12 recipientes cada, nos quais se colocou o mesmo número de insetos a que se aplicaram aqueles inseticidas. Indicadores relativos ao nº de insetos 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



- Associe cada *boxplot* a cada *spray*.
- Calcule os valor das barreiras de *outliers* (inferior e superior) no primeiro diagrama. Qual é o valor do *outlier*?

- c) Compare os três conjuntos de dados quanto à localização, dispersão e simetria.
- d) Para a totalidade das observações calcule a média, a variância e a amplitude total.

1.12. 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(np.mean(y))      print(np.var(y))
15                 16.24                  3.31707

print(statistics.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.

1.13. 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(np.mean(x))      >>> print(np.var(x))
150.85                    806.7275
```

- a) Calcule o coeficiente de variação 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.

1.14. Pensa-se que o custo de manutenção (em euros) de tratores aumenta com a idade (em anos) do trator. Para modelar a relação entre o custo e a idade registaram-se observações de idade de 17 tratores e respetivo custo de manutenção. Os dados foram introduzidos nas listas `idade` e `custo` em Python. Apresentam-se abaixo alguns resultados obtidos.

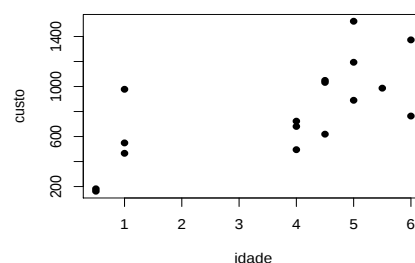
```
>>> custo = [...]
>>> custo.sort()
>>> print(custo)
[163, 182, 466, 495, 549, 619, 681, 723, 764, 878, 890,
987, 1033, 1049, 1194, 1373, 1522]

>>> idade = [...]
>>> 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

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

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



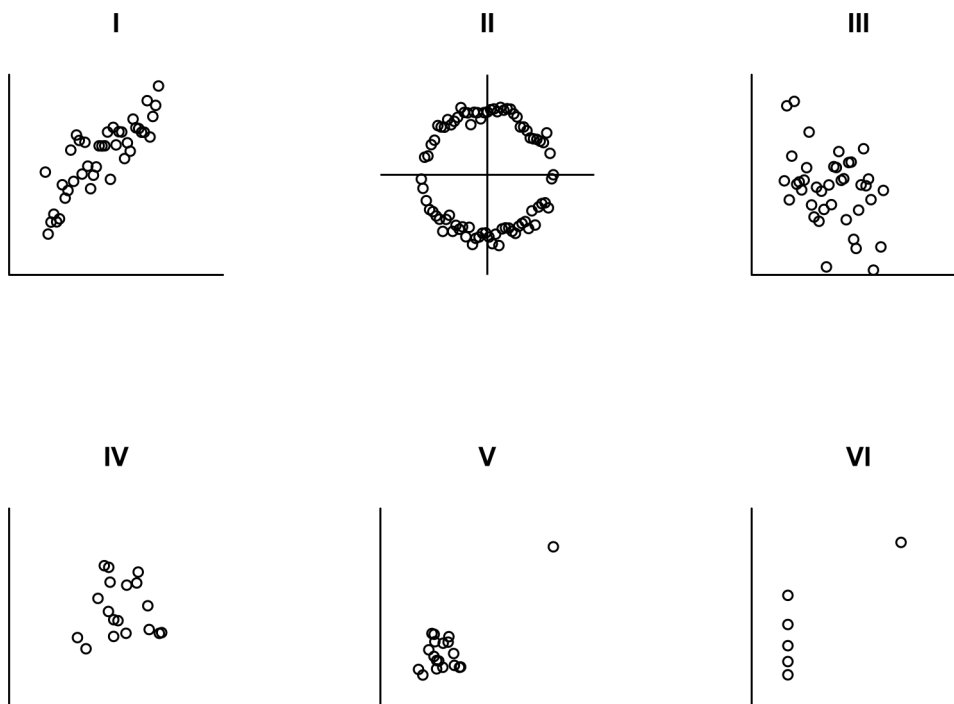
- a) Classifique a variável “custo de manutenção”. Justifique. Face ao objetivo do estudo trata-se da variável resposta ou da variável independente?
- b) No *output* acima faltam dois valores (A e B). Calcule-os.
- c) Esboce a caixa-de-bigodes do custo de manutenção. Comente.
- d) Determine a variância da idade dos tratores nas unidades originais e em meses².
- e) Poder-se-á admitir a existência de uma relação linear entre o custo de manutenção e a idade dos tratores? Justifique. Independentemente da sua resposta determine a equação da reta de regressão que modela a relação entre as variáveis em estudo.
- f) Calcule a precisão da reta e interprete o seu significado.
- g) Qual a variação anual média do custo de manutenção, estimada pela reta de regressão, quando os tratores têm idade entre 0.5 e 6 anos?
- h) Determine os coeficientes da reta de regressão se a idade fosse convertida para meses. Justifique.

1.15. Para $n = 20$ pares de observações (x_i, y_i) , seja $y = -5.6 + 0.7x$, a equação da reta 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 reta é positivo.
 - iii) Em média, quando x aumenta y não aumenta, pois o declive da reta é menor do que 1.
- b) Sendo $\sum_{i=1}^{20} x_i = 200$ determine $\sum_{i=1}^{20} y_i$.

1.16. Qual dos valores abaixo indicados pode ser o coeficiente de correlação dos dados descritos nas seguintes nuvens de pontos:

0 0.8 -0.5 2.0

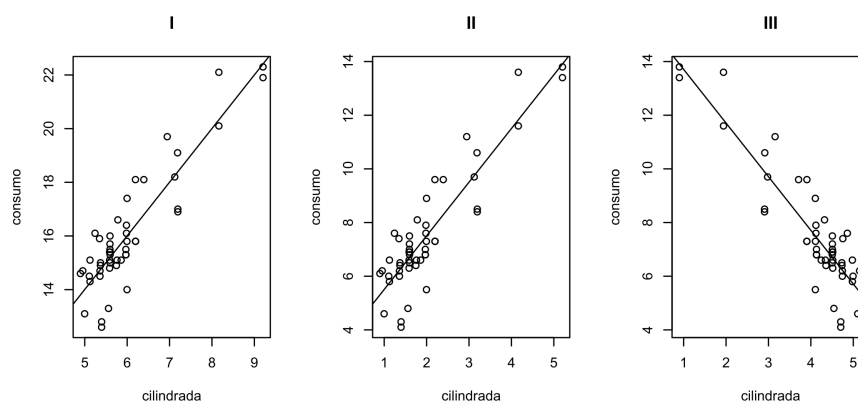


1.17. Num estudo sobre o consumo de gasolina de vários modelos de automóveis ligeiros de passageiros e a cilindrada do respetivo motor, foi estabelecida a seguinte equação da reta de regressão dos mínimos quadrados

$$y = 3.5 + 2x$$

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 reta 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), respetivamente, responda às seguintes questões:

- Determine a média e o desvio padrão dos consumos de gasolina dos automóveis observados.
- Qual é a variação esperada para o consumo de gasolina quando se aumenta a cilindrada de 1000 cm^3 ?
- Qual dos seguintes gráficos pode corresponder à nuvem de pontos e reta de regressão dos dados?



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.

1.18. A medição direta do calor específico de ramos de macieira é difícil de efetuar. Um investigador propõe prever o calor específico de ramos individuais a partir de medições (muito mais simples de efetuar) da percentagem de água no ramo, em vez de medir diretamente 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

- Desenhe o diagrama de extremos e quartis para os valores do calor específico observados. Comente a distribuição dos dados.
- Parece-lhe adequada a existência de uma relação linear entre x e y ? Porquê? Independentemente da sua resposta ajuste aos dados a reta de regressão dos mínimos quadrados.
- 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.
- 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 reta e os da reta que obteve em b). Haverá alteração na precisão da regressão?

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

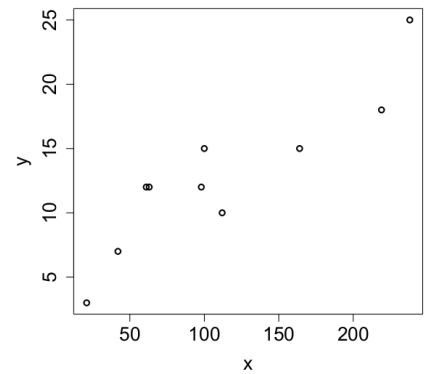
	urso	hipopótamo	canguru	leopardo	leão	macaco	rato	porco	cão	gato
x_i	219	238	42	98	100	164	21	112	61	63
y_i	18	25	7	12	15	15	3	10	12	12

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]
>>> import statistics
>>> statistics.mean(x)    >>> statistics.mean(y)
111.8                     12.9
```

```
>>> statistics.variance(x)
5394.622
>>> statistics.variance(y)
36.1

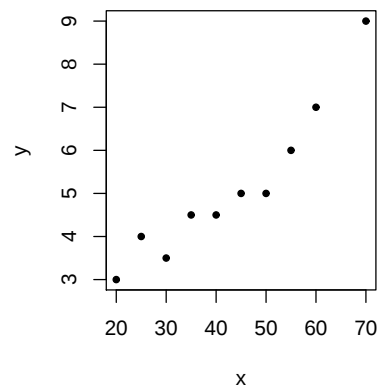
>>> statistics.covariance(x,y)
396.7556
```



- Parece-lhe adequada a existência de uma relação linear entre x e y ? Justifique.
- Independentemente da resposta à alínea anterior determine a reta de regressão dos mínimos quadrados de y sobre x . Calcule a precisão da reta e interprete o seu significado.
- Interprete, no contexto do problema, o significado do coeficiente de regressão de y sobre x .
- O período de gestação de uma girafa é de 425 dias. Se usasse a reta 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.

1.20. 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)
statistics.mean(x)
43
statistics.mean(y)
5.15
statistics.variance(x)
256.6667
statistics.variance(y)
3.169444
statistics.covariance(x,y)
27
```

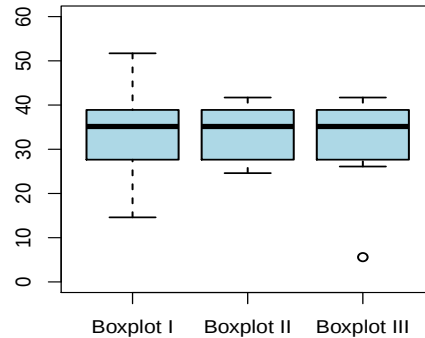


Com base nos resultados apresentados, responda às seguintes questões.

- Parece-lhe adequada a existência de uma relação linear entre x e y ? Justifique.
 - Independentemente da resposta à alínea anterior determine a reta de regressão dos mínimos quadrados de y sobre x . Calcule a precisão da reta e interprete o seu significado.
- 1.21.** 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/m²):

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.



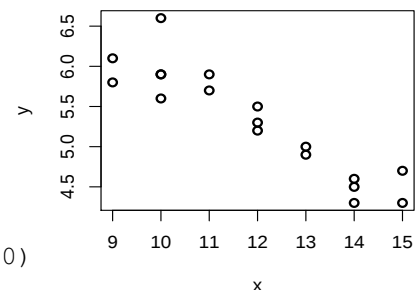
- c) O teor de azoto foliar (em g/kg de peso seco) relaciona-se linearmente com o teor de clorofila medido colorimetricamente segundo a seguinte equação de reta de regressão dos mínimos quadrados $y = 2.3195 + 0.5224x$, em que x é o teor de clorofila medido colorimetricamente e y o teor de azoto foliar.
- Sabendo que 96% da variabilidade de y é explicada pela regressão, determine o coeficiente de correlação entre x e y . Interprete-o.
 - Qual o resíduo associado ao par (32.6, 18.1)?
 - Determine a média e a variância do teor de azoto foliar.

1.22. 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 $^{\circ}\text{C}$ 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(statistics.correlation(x,y)**2)
0.842407

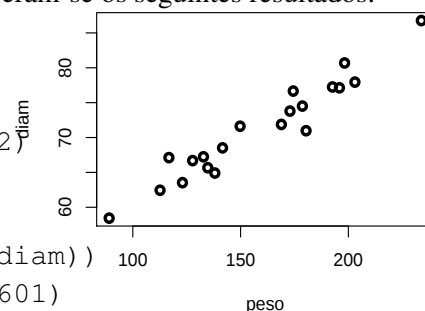
>>> print(statistics.linear_regression(x, y))
LinearRegression(slope=-0.3092, intercept=8.9980)
```



- Parece-lhe admissível a existência de uma relação linear entre as variáveis consideradas? Justifique.
- Independentemente da resposta à alínea anterior, escreva a equação da reta de regressão dos mínimos quadrados. Interprete, no contexto do problema, o valor do coeficiente de regressão.
- Qual é o resíduo associado à observação (15,4.3)?

1.23. 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(statistics.correlation(peso, diam) ** 2)
0.9298733
```



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

- Escreva a equação da reta de regressão ajustada. Qual a precisão dessa reta? Comente.
- Parece-lhe adequada a utilização do modelo linear para descrever a relação entre o diâmetro e o peso dos frutos analisados? Justifique.
- Sabendo que a média e o desvio padrão do peso dos frutos são 158.25 g e 37.09 g, respetivamente, determine a média e o desvio padrão do diâmetro dos frutos observados.
- De acordo com o modelo ajustado, qual é a variação esperada no diâmetro dos frutos associada ao aumento de 1 grama no peso?
- Determine a reta de regressão dos mínimos quadrados do diâmetro sobre o peso, caso os pesos fossem registados em kg. Qual a precisão desta reta?

1.24. 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).

- 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 efetuados 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]
```

```
>>> import statistics
>>> import matplotlib.pyplot as plt

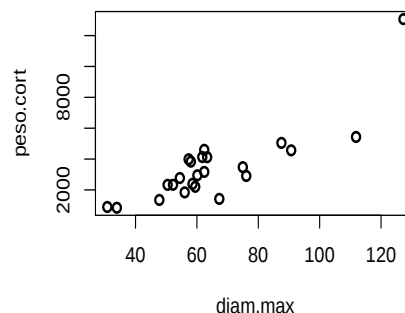
>>> statistics.quantiles(diam_max)
[54.43,      B,      74.96]

>>> 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 possível 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 efetuaram-se os seguintes cálculos para as variáveis `peso_cort` e `diam_max`

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

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



- i) Escreva a equação da reta de regressão dos mínimos quadrados do peso da cortiça sobre o diâmetro máximo das árvores na parcela e interprete o significado do declive.
- ii) Se excluísse a observação correspondente ao ponto mais acima e à direita na nuvem de pontos, qual a alteração que prevê para o declive da reta ajustada?
- iii) 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.
- iv) O peso de cortiça é geralmente expresso em arrobas (1 arroba(@) = 15 kg). Qual seria o declive da reta indicada em a) se o peso da cortiça fosse registado em arrobas? Justifique convenientemente.

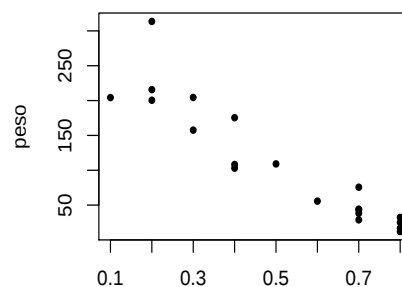
1.25. Pretende-se estudar a influência da infecção de um certo fungo no peso de alfaces em estufa. Para cada uma de 20 alfaces obteve-se o peso em gramas e registou-se o grau de infecção numa escala de 0 a 1 (sendo 0 – inexistência de infecção e 1 – infecção muito elevada). Estes valores foram introduzidos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

>>> statistics.mean(peso)      >>> statistics.mean(grau)
108.21                        0.515

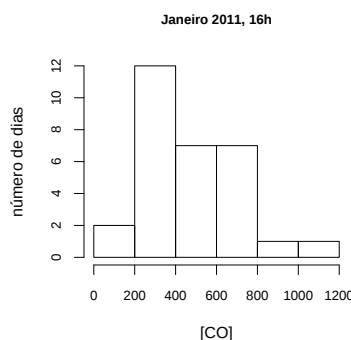
>>> statistics.variance(peso)   >>> statistics.variance(grau)
7439.075                      0.05818421

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



- Calcule os indicadores necessários para construir o *boxplot* do peso das alfaces. Faça a sua representação gráfica.
- Será de admitir a existência de uma relação linear entre as variáveis? Justifique.
- Independentemente da resposta dada na alínea anterior, pretende-se ajustar uma reta de regressão aos dados.
 - Qual é a variável resposta?
 - Obtenha a equação da reta de regressão dos mínimos quadrados para as variáveis em estudo. Calcule a precisão desta reta e interprete-a.

- 1.26.** O gráfico mostra o histograma dos 30 registos diários disponíveis da concentração de CO, $\mu\text{g}/\text{m}^3$, 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) 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 reta de regressão dos mínimos quadrados da concentração de CO (y) sobre o número de automóveis (x): $y = 468.25 + 0.021x$.
- 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?
- 1.27.** 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, respetivamente.

- a) Os valores de NDVI foram introduzidos em Python e foram 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) Esboce o histograma dos dados considerando as classes apresentadas no *output*.
- b) Pretende-se averiguar se a utilização do espectrorradiómetro, 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:

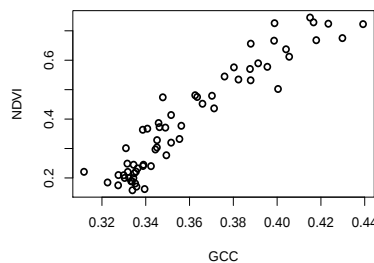
```
>>> statistics.correlation(GCC, NDVI)
0.945354
```

```
>>> statistics.mean(GCC)
0.36
```

```
>>> statistics.mean(NDVI)
0.39
```

```
>>> statistics.variance(GCC)
0.00099
```

```
>>> statistics.variance(NDVI)
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 reta 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 obtém o valor de GCC igual a 0.4?
- 1.28.** Para oito pares de observações $\{(x_i, y_i)\}_{i=1}^8$ determinou-se a reta 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 $\bar{y} = 0.2$.
 - b) O coeficiente de correlação é $r = 0.98$.
 - c) Se $s_x^2 = 0.5522$ então $s_y^2 = 0.828$.
- 1.29.** Considere n pares de observações (x_i, y_i) . Seja $z_i = \frac{x_i - \bar{x}}{s_x}$, $i = 1, \dots, n$. Mostre que as observações z_i têm média nula e variância unitária.
- 1.30.** Considere n observações de uma variável x . Designe por $x = \{x_1, \dots, x_n\}$ essas n observações e por $\bar{x}$ a média dessas observações. Designe por $x' = \{x_1, \dots, x_n, \bar{x}\}$ 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 $s_{x'}^2$.

1.31. 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 retas de regressão de y sobre x para cada um dos quatro pares de variáveis (x_i, y_i) , $(i=1, \dots, 4)$.
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.

Soluções ou resoluções dos Exercícios de Estatística Descritiva

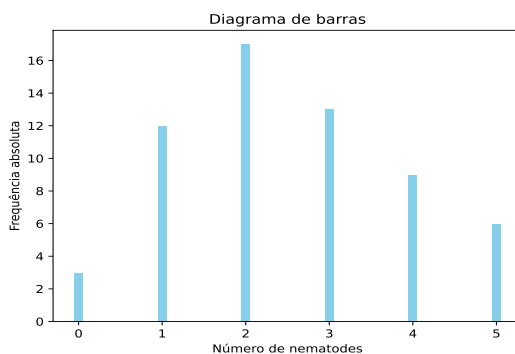
1.1. Comandos Python para construção das tabelas de frequência e respectivas representações gráficas.

```
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')

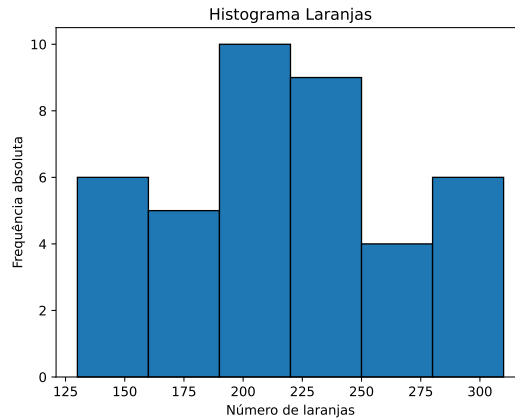
    plt.title("Histograma Laranjas")
    plt.xlabel("Número de laranjas")
    plt.ylabel("Frequência absoluta")
    plt.show()
    # Frequências
    ni = hist_laranjas[0] # Frequência absoluta
    fi = ni / sum(ni) # Frequência relativa
    Ni = np.cumsum(ni) # Frequência absoluta acumulada
    Fi = Ni / 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

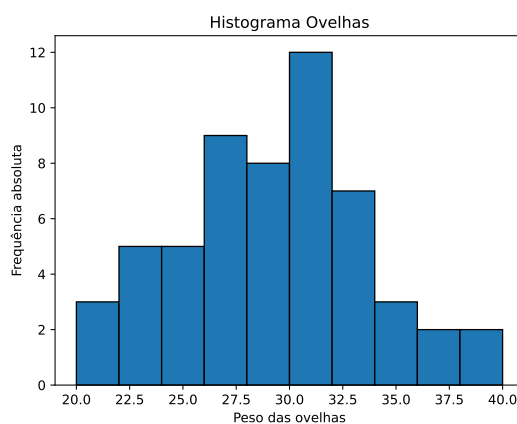


```
c) # ovelhas
ovelhas_df = pd.read_csv('ovelhas.txt')
print(ovelhas_df.head())
# Regra de Sturges
num_sturges = round(1 + np.log2(len(ovelhas_df['peso'])))
amplitude_classe = (ovelhas_df['peso'].max() -
ovelhas_df['peso'].min()) / num_sturges
print(amplitude_classe)
hist_ovelhas = plt.hist(ovelhas_df['peso'], 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



- 1.4.** 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.
- 1.5.** a) variável - nº de golfinhos em cada passeio - variável discreta porque toma valores em $\mathbb{N}_0$.
c) média $\bar{x} = 2.28$, $mo = 2$ e $\tilde{x} = 2$ golfinhos/passeio.
Medidas de dispersão: variância $s^2 = 1.73$ e desvio padrão $s = 1.31$.
d) Pretende-se $F(2) = \frac{17+45+84}{235} = 0.6213$, frequência relativa acumulada, logo é 62%
- 1.6.** b) As classes dadas têm amplitudes diferentes e como a área de cada rectângulo deverá representar a frequência relativa associada à respetiva classe, isto é, $A_i = f_i$ e $A_i = h_i \times alt_i \Rightarrow alt_i = \frac{f_i}{h_i}$, h_i amplitude da classe i e alt_i altura do rectângulo, com base na classe i .

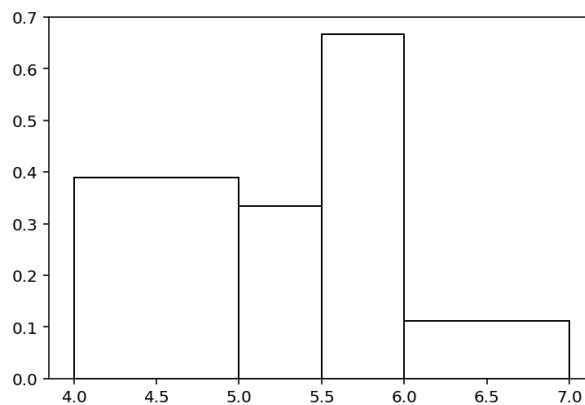
Tem-se, por exemplo $A_1 = f_1 = 0.07$, $A_2 = f_2 = 0.19$, $A_3 = f_3 = 0.2$, etc., donde
 $alt_1 = \frac{0.07}{20} = 0.0035$; $alt_2 = \frac{0.19}{10} = 0.019$; $alt_3 = \frac{0.2}{5} = 0.04$, vemos que esta última altura
é quase o dobro da segunda, portanto só poderá ser o **histograma 1**.

1.8. 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

b) Histograma



1.9. 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.

1.11. a) Spray A $\rightarrow$ *boxplot* 3; Spray B $\rightarrow$ *boxplot* 2; Spray C $\rightarrow$ *boxplot* 1

b) BI= -2, BS= 6. O máximo é *outlier* pois $\max = 7 > BS$.

c)

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 insetos. 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.

- d) a totalidade das observações é $n = 36$. Para calcular a média necessitamos de

$$\sum_{i=1}^{36} x_i = 174 + 184 + 25 \quad \text{donde} \quad \bar{x} = 10.63 \text{ mortes}$$

Para calcular a variância do total devemos usar a expressão

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

- 1.12.** 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 $\mathbb{R}$, mais precisamente em $\mathbb{R}^+$.

- b) Seja y o teor diário de vitamina C.

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

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

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

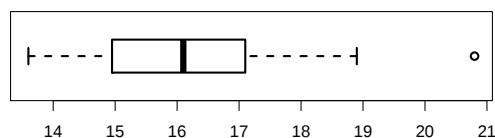
1º quartil: $Q_1 = 14.8$ e 3º quartil: $Q_3 = 17.1$

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* inferior.

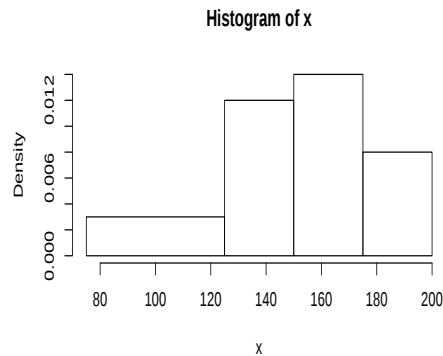
O *boxplot* é:



1.13. a) $CV = 0.18828$.

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, ou seja uma causa mais pesada à esquerda.

c) Como as classes têm amplitudes diferentes, a altura de cada retângulo é o quociente entre a frequência relativa e a amplitude da respetiva classe.



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.

1.14. 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) $s_x^2 = 467.735$ meses²

e) $r = 0.783$ relativamente próximo de 1, que corresponde à relação linear perfeita, e nuvem de pontos sugere a possibilidade de traçar uma reta 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) $b'_1 = 1/12b_1 = 13.66775$; $b'_0 = b_0$

1.15. a) i) A afirmação é falsa.

Por definição tem-se $r_{y,x} = \frac{\text{cov}(y,x)}{s_y s_x} = \frac{\text{cov}(x,y)}{s_x s_y} = r_{x,y}$, 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 reta dos mínimos quadrados é: $\bar{y} = b_0 + b_1 \bar{x}$, o que é equivalente a $\sum y_i = nb_0 + b_1 \sum x_i = 20 \times (-5.6) + 0.7 \times 200 = 28$.

1.16. Os valores possíveis para coeficientes de correlação dos dados são:

para a nuvem I – > 0.8 ; para a nuvem II – > 0 ; para a nuvem III – > -0.5 ;

para a nuvem IV – > 0 ; para a nuvem V – > 0.8 ; para a nuvem VI – > 0.8 .

1.17. 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.

d) A nuvem de pontos mostra uma tendência linear entre as variáveis; o coeficiente de correlação é elevado, $r = \sqrt{0.803} = 0.89$, logo o modelo linear é adequado.

1.19. 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 reta 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 reta 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 reta 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 reta não permite efetuar esta previsão, pelo que, a grande diferença entre o valor ajustado e o valor real (25 anos) é justificável.

1.20. a) O coeficiente de correlação entre as variáveis x e y , $r = \frac{\text{cov}(x,y)}{(s_x s_y)} = 0.9466$, é próximo de 1 ($|r| = 1$ significa a existência de relação linear perfeita entre os valores observados), no entanto a nuvem de pontos sugere a existência de tendência curvilínea. Assim, um modelo curvilíneo poderá ser mais adequado do que o modelo linear.

b) Sendo a equação da reta $y = b_0 + b_1x$, o declive b_1 é

$$b_1 = \frac{\text{cov}(x,y)}{s_x^2} = \frac{27}{256.6667} = 0.1052$$

e a ordenada na origem é

$$b_0 = \bar{y} - b_1\bar{x} = 5.15 - 0.1052 \times 43 = 0.6266.$$

A equação da reta é então $y = 0.6266 + 0.1052x$.

A precisão da reta é $R^2 = r^2 = (0.9466)^2 = 0.8961$ e significa que 89.61% da variabilidade da quantidade de água evaporada é explicada pela reta de regressão.

- 1.23.** a) $y = 42.4601 + 0.1815x$. Precisão: $R^2 = 0.9299$
 b) Sim, $r = 0.9643$ e a nuvem de pontos apresenta-se próximo de uma reta.
 c) $\bar{y} = 71.18$ mm; $s_y = 6.98$ mm
 d) Quando o peso aumenta 1 g espera-se que o diâmetro aumente 0.1815 mm.
- 1.25.** 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;
1º 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;
3º 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) $y = 279.04884 - 331.7259x$; $R^2 = r^2 = 0.86068$
- 1.30.** a) $\bar{x'} = \frac{x_1 + x_2 + \dots + x_n + \bar{x}}{n+1} = \frac{n\bar{x} + \bar{x}}{n+1} = \frac{(n+1)\bar{x}}{n+1} = \bar{x}$.
 b) $s_{x'}^2 = \frac{\sum_{i=1}^{n+1} (x'_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + (\bar{x} - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{(n-1)s_x^2}{n}$.
- 1.31.** a)
- ```
> import pandas as pd
> import matplotlib.pyplot as plt
> import numpy as np
> import statistics as stat

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

> print(np.mean(ans_df, axis=0))
x1 9.000000
x2 9.000000
x3 9.000000
x4 9.000000
y1 7.500909
y2 7.500909
y3 7.500000
y4 7.500909
```

```
> print(np.var(ans_df,axis=0))
x1 10.000000
x2 10.000000
x3 10.000000
x4 10.000000
y1 3.752063
y2 3.752390
y3 3.747836
y4 3.748408
```

b)

```
> print(stat.correlation(ans_df['x1'],ans_df['y1']))
0.81642051634484
```

```
> print(stat.correlation(ans_df['x2'],ans_df['y2']))
0.8162365060002428
```

```
> print(stat.correlation(ans_df['x3'],ans_df['y3']))
0.8162867394895981
```

```
> print(stat.correlation(ans_df['x4'],ans_df['y4']))
0.8165214368885029
```

c) > print(stat.linear\_regression(ans\_df['x1'],ans\_df['y1']))  
LinearRegression(slope=0.5000909090909091, intercept=3.0000909090909085)

```
> print(stat.linear_regression(ans_df['x2'],ans_df['y2']))
LinearRegression(slope=0.5, intercept=3.000909090909091)
```

```
> print(stat.linear_regression(ans_df['x3'],ans_df['y3']))
LinearRegression(slope=0.49972727272727274, intercept=3.0024545454545457)
```

```
> print(stat.linear_regression(ans_df['x4'],ans_df['y4']))
LinearRegression(slope=0.4999090909090909, intercept=3.0017272727272726)
```

d)

```
fig, axes = plt.subplots(2, 2, figsize=(10.5, 7))
ax = axes[0,0]
ax.scatter(ans_df['x1'],ans_df['y1'])
ax.set_xlabel("x1")
ax.set_ylabel("y1")
```

```

ax = axes[0,1]
ax.scatter(ans_df['x2'],ans_df['y2'])
ax.set_xlabel("x2")
ax.set_ylabel("y2")

ax = axes[1,0]
ax.scatter(ans_df['x3'],ans_df['y3'])
ax.set_xlabel("x3")
ax.set_ylabel("y3")

ax = axes[1,1]
ax.scatter(ans_df['x4'],ans_df['y4'])
ax.set_xlabel("x4")
ax.set_ylabel("y4")

Adjust layout to prevent labels from overlapping
plt.tight_layout()

Display the figure
plt.show()

```

