Parte C - Introdução à Inferência Estatística

Inferência sobre Parâmetros Populacionais

3.1. (*Exame 12.01.2009*) O peso médio dos indivíduos duma certa espécie de bivalves é 31 g e o respetivo desvio padrão é 2.4 g. Recolhe-se uma amostra aleatória de 100 indivíduos desta espécie.

a) Qual a probabilidade, aproximada, de a média da amostra ser inferior a 30 g?

b) Qual a probabilidade, aproximada, de o peso total da amostra ser superior a 3150 g?

3.2. Mediu-se o comprimento (em cm) de 25 coelhos adultos de uma dada raça escolhidos aleatoriamente. A média da amostra foi 56.2 cm e o desvio padrão 4 cm. Admita que a população dos comprimentos dos coelhos adultos daquela raça é bem modelada por uma distribuição normal.

a) Indique estimativas para a média e o desvio padrão da população.

b) Usando as estimativas da alínea anterior como os parâmetros da população, determine a probabilidade de o comprimento de um coelho adulto daquela raça, escolhido ao acaso, ser superior a 57 cm.

c) Chama-se erro padrão da média amostral, $\overline{X}$, a uma estimativa do desvio padrão da v.a. média amostral. Com base nos resultados obtidos, determine o erro padrão da média amostral.

3.3. Depois de fabricado e embalado, a atividade de um certo adubo segue aproximadamente uma distribuição normal com $\mu = 120$ dias e $\sigma = 40$ dias.

Pretende-se enviar um lote de embalagens do referido adubo de modo que a atividade média amostral ($\overline{X}$) seja superior a 118 dias com probabilidade de 0.95. Qual o tamanho do lote a enviar?

3.4. Seja $(X_1, ..., X_n)$ uma amostra aleatória de tamanho n proveniente de uma população X com distribuição $Uniforme(0,1)$. Exprima, em função de n, a probabilidade, aproximada, de $\overline{X}$ ser superior 0.9.

3.5. Uma amostra aleatória de dimensão 50, $(X_1, ..., X_{50})$, é extraída de uma população de Poisson com $\lambda = 10$. Recorra à distribuição normal para calcular um valor aproximado de $P(\overline{X} > 11)$.

3.6. (*Exame 31.01.2011*) Considere a amostra aleatória $X_1, X_2, \ldots, X_{100}$, retirada de uma população X para a qual se tem $E[X] = 0.5$ e $Var[X] = 0.0625$. Considere a variável aleatória $S = \sum_{i=1}^{100} X_i$.

a) Defina "amostra aleatória".

b) Determine o valor esperado e a variância de S. Justifique convenientemente.

c) Determine, aproximadamente, $P[S < 45]$.

3.7. (*Exame 17.01.2011*) Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de dimensão n, proveniente de uma população X com valor médio μ e variância σ^2. Seja $\overline{X}$ a média da amostra aleatória.

Responda, **justificando convenientemente**, se são verdadeiras ou falsas as afirmações nas seguintes alíneas.

a) $\overline{X} \frown \mathcal{N}(\mu, \sigma/\sqrt{n})$.

b) $E[\mu] = \overline{X}$.

c) Se $X \frown Bernoulli(p)$ então $n\overline{X} \frown \mathcal{B}(n,p)$.

3.8. Seja X uma população com distribuição normal, de média μ e desvio padrão σ=2. Uma amostra aleatória de dimensão 25 foi extraída desta população, tendo-se obtido $\overline{x} = 78.3$.

a) Calcule o intervalo de confiança a 99% para μ.

b) Qual o erro máximo cometido (a 99% de confiança) ao estimar μ por $\overline{x}$ =78.3?

c) Qual deverá ser a dimensão da amostra para que o erro máximo cometido, a 99% de confiança, ao estimar μ por $\overline{x}$, não exceda 0.1?

3.9. Para avaliar a tensão máxima suportada por uma barra de aço testaram-se n barras tendo-se obtido, nas unidades adequadas, $\overline{x}_n = 20$ e para extremo superior do intervalo de confiança a 95% para o verdadeiro valor médio da tensão obteve-se 21.7. Sabendo que se admite que a tensão suportada por uma barra de aço é uma v.a. normal com desvio padrão $\sigma = 3$, determine o extremo inferior do intervalo de confiança e a dimensão n da amostra.

3.10. Uma empresa de peixe congelado está a ser investigada com o objetivo de se verificar se cada embalagem pesa de facto 1 kg em média. Numa amostra aleatória de 100 embalagens de peixe registou-se o peso (kg) de cada embalagem, x_i $(i = 1, \ldots, 100)$, tendo-se obtido os seguintes valores:

$$\sum_{i=1}^{100} x_i = 95.9 \text{ kg} \qquad \sum_{i=1}^{100} x_i^2 = 93.12 \text{ kg}^2$$

a) Indique uma estimativa para a média e para a variância do peso de uma embalagem de peixe.

b) O que pode dizer sobre o resultado da investigação? Justifique convenientemente, especificando as hipóteses que necessitou de considerar.

c) Determine, explicitando as hipóteses necessárias, um intervalo a 95% de confiança para a variância do peso de uma embalagem de peixe.

3.11. O grau de acidez do azeite produzido em certa região supõe-se ter distribuição normal. Uma amostra da produção de dimensão 25 conduziu a uma acidez média de 1 grau e a um desvio padrão de 0.33 graus.

a) Face a estes valores, alguém sugeriu o intervalo]0.815; 1.185[para o valor esperado da acidez. Qual o nível de confiança associado a este intervalo?

b) Para o nível de confiança determinado na alínea anterior, qual o erro máximo cometido ao estimar o verdadeiro valor médio da acidez por 1 grau?

3.12. Numa amostra de 16 elementos, que se supõe ter sido retirada de uma população com distribuição normal, o desvio padrão obtido foi de 5.2.

a) Foi calculado um intervalo de confiança para a média populacional tendo-se obtido $]24.2297; 29.7703[$. Indique justificando:

i) Qual a média da amostra;

ii) Qual o grau de confiança do intervalo calculado.

b) Qual o intervalo de confiança a 95% para a variância populacional?

3.13. (*Exame 31.01.2012*) Um silvicultor sabe que a altura das árvores para madeira é importante para os compradores. Afirmou a um comprador que a altura média das suas árvores era 28 metros. O comprador fechou negócio, mas aquando do abate escolheu ao acaso 65 árvores e mediu as alturas. Obteve $\sum_{i=1}^{65} x_i^2 = 43959.32\ m^2$, onde x_i designa a altura de cada árvore selecionada.
O comprador obteve o seguinte intervalo de confiança para a altura média, em metros:

$$]25.1538; 26.5462[.$$

a) Calcule uma estimativa para a altura média das árvores daquela floresta.

b) Determine a confiança do intervalo dado acima.

c) Considera que o comprador tem razão para reclamar? Justifique.

d) Pretende-se construir um intervalo de confiança para a variância da altura das árvores.

i) Indique o estimador da variância usado na construção deste intervalo, apresentando a sua expressão.

ii) Com base nos resultados fornecidos construa o intervalo de confiança a 95% para a variância das alturas, indicando o(s) pressuposto(s) necessário(s).

3.14. Suponha que o rendimento de um pé de tomateiro expresso em kg é uma variável aleatória com distribuição normal de valor médio 1 kg. Numa parte da produção foi utilizado um novo fertilizante. Observada uma amostra de 10 pés de tomateiro da parte da produção em que foi utilizado o novo fertilizante obtiveram-se os seguintes resultados:

1.375 1.223 1.773 1.752 0.779 1.407 1.068 1.633 1.201 1.042

expressos em kg. Que decisão se deverá tomar perante estes resultados, face ao novo fertilizante?

3.15. (*Exame 26.01.2018*) Para analisar a qualidade de um determinado tipo de azeite virgem, proveniente de uma região, R1, foram selecionadas aleatoriamente 20 garrafas e determinada a acidez do azeite de cada uma delas, expressa em concentração de ácido oleico (g/100g). Os dados obtidos foram introduzidos em Python, apresentando-se de seguida alguns dos comandos e respetivos resultados.

```
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> acidez=[0.635, 0.822, 0.833, ... , 0.862, 0.694, 0.665]
```

```
>>> print(stat.mean(acidez))
  0.76925
>>> print(stat.variance(acidez))
  0.01504357
```

```
>>> print(shapiro(acidez))
  ShapiroResult(statistic=0.95549,
  pvalue=0.4581)
```

a) De acordo com a legislação em vigor, o azeite virgem só pode ser classificado como Extra se tiver uma acidez média inferior a 0.8. Será que o azeite virgem analisado pode ter essa classificação? Justifique convenientemente a sua resposta.

b) Numa outra região, R2, analisou-se uma amostra de 20 garrafas de azeite virgem do mesmo tipo. Com os valores observados construiu-se o seguinte intervalo de confiança a 95% para a diferença da acidez média do azeite virgem daquele tipo entre as regiões R1 e R2.

$$]-0.0357\,;0.0667[$$

Justificando convenientemente, responda às seguintes questões:

i) Que pressupostos foi necessário verificar para construir o intervalo de confiança dado?

ii) Determine o valor da acidez média observada na região R2.

iii) Compare a acidez média deste tipo de azeite nas duas regiões R1 e R2.

3.16. (*Exame 28.01.2013*) Pretende-se comparar dois antibióticos A e B usados para o tratamento de um certo tipo de infeção em bovinos. Escolhem-se ao acaso dois grupos de bovinos doentes: a um grupo é administrado o antibiótico A e ao outro o antibiótico B, registando-se o tempo (h) até os sintomas desaparecerem. Os resultados obtidos pela aplicação dos dois antibióticos foram introduzidos em Python. Face aos resultados obtidos poder-se-á concluir que o antibiótico A é mais lento do que o antibiótico B no desaparecimento dos sintomas da infeção? Explicite e valide os pressupostos necessários à resolução do problema.

Anexo

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> A=[49.5, 48, 52.5, 46.5, 52, 48.5, 40, 50.5, 46, 46.5]
>>> B=[47.5, 48.5, 40.5, 43.5, 49, 43, 42, 45, 48.5, 40]
>>> n=10
```

```
>>> print(stat.mean(A))
  48.0
```

```
>>> print(stat.mean(B))
      44.75
```

```
>>> print(stat.variance(A))
  13.055555555555555
```

```
>>> print(stat.variance(B))
  11.847222222222221
```

```
>>> D=np.array(A)-np.array(B)
>>> print(stat.variance(D))
  19.291666666666668

>>> print(shapiro(A))
  ShapiroResult(statistic=0.9189298830189923, pvalue=0.3481161319651563)

>>> print(shapiro(B))
  ShapiroResult(statistic=0.9009386633181248, pvalue=0.22435057485226023)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.9482147705872899, pvalue=0.6474277848783038)

>>> F_calc=stat.variance(A)/stat.variance(B)
>>> print(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)))
  0.8873403061455494

>>> F_calc=stat.variance(A)/stat.variance(B)
>>> print(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)))
  0.8873403061455494

>>> print(ttest_rel(A,B,alternative='two-sided'))
  TtestResult(statistic=2.3399064056156256, pvalue=0.04402702582690393, df=9)

>>> print(ttest_rel(A,B,alternative='less'))
  TtestResult(statistic=2.3399064056156256, pvalue=0.977986487086548, df=9)

>>> print(ttest_rel(A,B,alternative='greater'))
  TtestResult(statistic=2.3399064056156256, pvalue=0.022013512913451966, df=9)

>>> print(ttest_ind(A,B,alternative='two-sided',equal_var=True))
  TtestResult(statistic=2.0594889418536337, pvalue=0.05421204503444033, df=18.0)

>>> print(ttest_ind(A,B,alternative='less',equal_var=True))
  TtestResult(statistic=2.0594889418536337, pvalue=0.9728939774827798, df=18.0)

>>> print(ttest_ind(A,B,alternative='greater',equal_var=True))
  TtestResult(statistic=2.0594889418536337, pvalue=0.027106022517220166, df=18.0)
```

3.17. (*Exame 25.01.2016*) Um investigador está interessado em avaliar, numa dada região, a produção por macieira de duas variedades de maçã: Gala e Golden. Para essa finalidade, foram registadas as produções, em kg, de 10 macieiras de cada uma das variedades de maçã em estudo. Os dados foram introduzidos em Python. No ANEXO apresentam-se resultados de comandos, alguns inadequados. Responda, de forma completa, às seguintes questões utilizando os resultados do ANEXO.

a) Indique uma estimativa para o valor médio da produção da variedade Gala. Qual é o erro máximo, a 95% de confiança, que se comete ao estimar μ pela estimativa indicada?

b) Conjetura-se que as variabilidades das produções das duas variedades são iguais. Será esta conjetura compatível com os dados obtidos?

c) Com base nos valores obtidos poder-se-á concluir que a produção média da variedade Golden é significativamente superior à produção média da variedade Gala?

ANEXO

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> gala=[84,82,90,86,80,91,85,79,81,82.]
>>> golden=[95,102,85,93,104,89,98,99,107,106]
>>> n=10

>>> print(stat.mean(gala))       >>> print(stat.mean(golden))
  84.0                            97.8

>>> print(stat.variance(gala))   >>> print(stat.variance(golden))
  16.444444444444443             53.51111111111111

>>> D=np.array(gala)-np.array(golden)
>>> print(stat.variance(D))
  121.28888888888889

>>> print(shapiro(gala))
  ShapiroResult(statistic=0.9240743148918314, pvalue=0.3921921814902618)

>>> print(shapiro(golden))
  ShapiroResult(statistic=0.9572942210777938, pvalue=0.7546349097741636)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.9013872973491893, pvalue=0.22690085777222685)

>>> F_calc=stat.variance(gala)/stat.variance(golden)
>>> print(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)))
  0.09365153303876877

>>> macas_res=ttest_rel(gala,golden,alternative='two-sided')
>>> ci = macas_res.confidence_interval(confidence_level=0.95)
>>> print(macas_res)
  TtestResult(statistic=-3.9624936350722733, pvalue=0.0032918235989336905, df=9)
>>> print(ci)
  ConfidenceInterval(low=-21.67831394127077, high=-5.921686058729232)

>>> macas_res=ttest_ind(gala,golden,alternative='two-sided',equal_var=True)
>>> ci = macas_res.confidence_interval(confidence_level=0.95)
```

```
>>> print(macas_res)
  TtestResult(statistic=-5.217566360800018, pvalue=5.811449361572697e-05, df=18.0)
>>> print(ci)
  ConfidenceInterval(low=-19.356752353578063, high=-8.243247646421931)

>>> print(ttest_ind(gala,golden,alternative='less',equal_var=True))
  TtestResult(statistic=-5.217566360800018, pvalue=2.9057246807863483e-05, df=18.0)

>>> print(ttest_ind(gala,golden,alternative='greater',equal_var=True))
  TtestResult(statistic=-5.217566360800018, pvalue=0.9999709427531922, df=18.0)
```

3.18. (*Exame 11.01.2016*) Os rolamentos produzidos por uma empresa devem ter diâmetro entre 150 e 160 mm (inclusive). Uma amostra de 35 rolamentos foi selecionada aleatoriamente e os valores observados foram introduzidos num interpretador de Python. Utilize os resultados apresentados no **Anexo** e responda, de forma completa, às seguintes questões. (Note que alguns resultados podem não ser necessários ou adequados).

a) Indique uma estimativa para o valor médio e para a variância do diâmetro dos rolamentos. Explicite os estimadores associados.

b) Determine um intervalo de confiança a 95% para o valor médio dos diâmetros das peças produzidas?

c) Poder-se-á admitir que mais de 60% dos rolamentos produzidos pela empresa satisfazem o exigido? Justifique.

ANEXO

```
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import ttest_1samp
>>> diametro=[135, 154, 159, 155, 167, 159, 158, 159, 154, 155,
          152, 169, 154, 158, 140, 149, 145, 157, 151,154,168,153,151,
          160, 155, 155, 143,157, 139, 159, 139, 159, 162, 151, 150]

>>> diametro.sort()
>>> print(diametro)
[135, 139, 139, 140, 143, 145, 149, 150, 151, 151, 151, 152, 153, 154, 154, 154,
154, 155, 155, 155, 155, 157, 157, 158, 158, 159, 159, 159, 159, 159, 160, 162,
167, 168, 169]

>>> print(stat.mean(diametro))     >>> print(stat.variance(diametro))
  153.85714285714286                 63.2436974789916

>>> print(shapiro(diametro))
  ShapiroResult(statistic=0.947622279715614, pvalue=0.09576992372651963)

>>> diametro_res=ttest_1samp(diametro, popmean=0)
>>> print(diametro_res.confidence_interval(confidence_level=0.95))
```

```
  ConfidenceInterval(low=151.1253324010069, high=156.58895331327884)

>>> diametro_res=ttest_1samp(diametro, popmean=0, alternative='less')
>>> print(diametro_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=-inf, high=156.1301393099192)
```

3.19. Num estudo sobre o número de folhas por planta de tabaco, obtiveram-se os seguintes dados:

nº de folhas	17	18	19	20	21	22	23	24
nº de plantas	3	22	44	42	22	10	6	1

a) Indique a unidade estatística e a variável em estudo.

b) Determine a média e a mediana da amostra. Compare-as e comente.

c) Com base nesta amostra poder-se-á afirmar que 90% das plantas de tabaco têm 21 ou menos folhas? Justifique.

3.20. Duas marcas de chocolate (A e B) produzem chocolate rotulado "70% de cacau". Selecionaram-se aleatoriamente 15 chocolates de cada uma das marcas, que foram analisados quanto ao teor de cacau. Os resultados foram introduzidos num interpretador de Python. Utilize, sempre que possível, os resultados apresentados abaixo para responder às seguintes questões (alguns cálculos apresentados são desnecessários ou inadequados).

a) Conjetura-se que as variabilidades do teor de cacau dos chocolates das duas marcas são diferentes. Será esta conjetura compatível com os dados obtidos? Explicite e valide os pressupostos necessários à resolução do problema.

b) Indique estimativas para o valor esperado do teor de cacau dos chocolates de cada uma das marcas.

c) Os dados recolhidos evidenciam que o teor médio de cacau do chocolate da marca A é significativamente superior ao da marca B? Explicite e valide os pressupostos necessários à resolução do problema.

d) O produtor de chocolate da marca A afirma que o teor médio de cacau do seu chocolate é superior a 70%. O que pode dizer sobre a afirmação do produtor? Justifique convenientemente.

Especifique as hipóteses nula e alternativa de um teste de hipóteses que permita averiguar a validade desta afirmação.

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> A=[72.8, 69.6, 70.0, 73.3, 69.8, 71.1, 71.0, 71.0, 71.0, 76.3, 71.3,
   75.4, 72.1, 70.4, 68.6]
```

```
>>> B=[69.8, 71.3, 67.3, 69.4, 67.7, 67.9, 70.7, 70.5, 70.2, 71.4, 71.6,
   73.1, 70.6, 68.5, 70.2]
>>> n=len(A)

>>> print(stat.mean(A))        >>> print(stat.mean(B))
  71.58                         70.01333333333334

>>> print(stat.variance(A))    >>> print(stat.variance(B))
  4.483142857142861             2.616952380952376

>>> D=np.array(A)-np.array(B)
>>> print(stat.variance(D))
  3.643809523809521

>>> print(shapiro(A))
  ShapiroResult(statistic=0.9065568565492009, pvalue=0.11991143895340922)

>>> print(shapiro(B))
  ShapiroResult(statistic=0.9612671948024654, pvalue=0.7144130336570444)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.9753850616872665, pvalue=0.9282542044943882)

>>> F_calc=stat.variance(A)/stat.variance(B)
>>> print(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)))
  0.325353975385434

>>> print(ttest_rel(A,B,alternative='less'))
  TtestResult(statistic=3.1786623612285476, pvalue=0.9966505812435795, df=14)

>>> print(ttest_rel(A,B,alternative='greater'))
  TtestResult(statistic=3.1786623612285476, pvalue=0.0033494187564204747, df=14)

>>> print(ttest_ind(A,B,alternative='less',equal_var=True))
  TtestResult(statistic=2.277142168091996, pvalue=0.9846919024547445, df=28.0)

>>> print(ttest_ind(A,B,alternative='greater',equal_var=True))
  TtestResult(statistic=2.277142168091996, pvalue=0.015308097545255545, df=28.0)
```

3.21. (*Exame 10.01.2018*) Com o objetivo de estudar o teor de gordura no leite de duas raças de vacas leiteiras (*Ayrshire*, raça A e *Holstein-Fresian*, raça B), foram escolhidas aleatoriamente 20 vacas de cada raça e foi determinada a percentagem de gordura no respetivo leite.

Os dados obtidos foram introduzidos em Python e foram efetuados cálculos que se encontram no **Anexo**. Utilize os resultados que considere adequados para responder às seguintes alíneas.

a) Indique uma estimativa pontual para a variância da percentagem de gordura no leite da raça A. Indique, explicitando-o, o estimador associado.

b) Comente a veracidade da seguinte afirmação: "Não existe diferença significativa entre as variâncias das percentagens de gordura no leite das raças A e B". Justifique convenientemente.

c) Será que a percentagem média de gordura no leite da raça A é superior à da raça B? Justifique convenientemente a sua resposta.

Anexo

```
import numpy as np
import statistics as stat
from scipy.stats import shapiro
from scipy.stats import f
from scipy.stats import ttest_ind
from scipy.stats import ttest_rel

A=[3.53, 3.71, ..., 4.37, 4.41, 4.44]
B=[3.30,3.38 ,..., 3.94 ,3.95 ,4.25]
n=20

print(stat.mean(A))        print(stat.mean(B))
  4.0600                     3.6605

print(stat.variance(A))    print(stat.variance(B))
  0.0680                     0.0542

print(shapiro(A))
  ShapiroResult(statistic=0.9569, pvalue=0.4836)

print(shapiro(B))
  ShapiroResult(statistic=0.945, pvalue=0.2974)

print(shapiro(np.array(A)-np.array(B)))
  ShapiroResult(statistic=0.9241, pvalue=0.1191)

F_calc=stat.variance(A)/stat.variance(B)
print(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)))
  0.6253

print(ttest_rel(A,B,alternative='less'))
  TtestResult(statistic=18.4124, pvalue=1, df=19)

print(ttest_rel(A,B,alternative='greater'))
  TtestResult(statistic=18.4124, pvalue=7.122e-14, df=19)
```

```
print(ttest_ind(A,B,alternative='less',equal_var=True))
  TtestResult(statistic=5.109, pvalue=1, df=38.0)

print(ttest_ind(A,B,alternative='greater',equal_var=True))
  TtestResult(statistic=5.109, pvalue=4.72e-06, df=38.0)
```

3.22. A fim de investigar os efeitos de ambientes nitrosos e de ambientes fosfatados no desenvolvimento de colónias de bactérias, contaminam-se 10 plaquetas envolvidas em cada um daqueles ambientes com as bactérias em estudo, e deixa-se incubar durante 24 horas. Após esse tempo, procede-se à contagem do número de colónias de bactérias em cada plaqueta, tendo-se obtido os seguintes resultados:

Ambiente nitroso	60	47	12	29	51	46	49	74	63	101
Ambiente fosfatado	8	46	21	13	58	33	20	46	31	38

Investigue a hipótese de o tipo de ambiente não influir no desenvolvimento das colónias de bactérias. Responda de forma completa à questão.

3.23. Numa Estação Florestal estudam-se problemas de intercepção de precipitação. Nesse sentido em 12 dias de chuva (suponha que se trata de uma amostra aleatória de dias com precipitação) são colocados dois udómetros para medir a quantidade de precipitação: um numa zona desarborizada e outro sob as copas das árvores. As leituras da quantidade de água em cada udómetro (medidas em cm de altura) deram os seguintes valores (cada coluna corresponde a um dia):

zona descoberta	5.87	1.30	2.34	2.82	5.89	9.09	1.93	9.27	4.65	4.35	5.00	8.43
sob coberto	4.96	1.14	2.24	2.26	4.75	7.83	1.86	8.85	4.17	3.65	4.08	7.99

a) Estime a precipitação média na zona desarborizada e na zona sob coberto.

b) Será admíssivel supor que a 95% de confiança as precipitações médias nos dois casos são iguais? Explicite as hipóteses necessárias à resolução desta questão.

c) Tendo em conta que, pela própria natureza do problema, o nível de precipitação sob coberto não deverá ser superior ao nível da correspondente precipitação a descoberto, indique um procedimento estatístico mais adequado para avaliar se existem diferenças significativas nos dois casos.

3.24. Pretende-se testar se a proporção de ulmeiros afetados pela grafiose é idêntica em duas zonas A e B. Na zona A foi recolhida uma amostra aleatória de 30 ulmeiros e verificou-se que 20 estavam afetados pela grafiose. Na zona B recolheu-se uma amostra de 35 ulmeiros e verificou-se que 27 estavam afetados pela grafiose. Que conclusão se pode tirar ao nível de significância de 0.05?

3.25. Diga, justificando, se cada uma das afirmações seguintes é verdadeira ou falsa:

a) Num teste de hipóteses $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ em que se obteve um *p-value* $= 0.034$, não se rejeita a hipótese de a média populacional ser nula com um nível de significância de 0.01.

b) Se $]2.15; 3.24[$ é um intervalo de confiança a 95% para σ^2 então a probabilidade de σ^2 pertencer a este intervalo é igual a 0.95.

3.26. É desencadeado um programa de controlo da poluição de um rio em que são efetuadas medições, antes de lançar a campanha antipoluição e um ano após. As medições são combinações de vários índices; quanto maior for o valor resultante maior é a poluição. Obtiveram-se os seguintes resultados:

Ponto de controlo	1	2	3	4	5	6	7	8	9	10
Antes da campanha	68	88	101	82	96	74	65	74	52	99
Um ano após	67	87	90	76	98	69	68	65	59	70

Será que a campanha antipoluição reduziu de facto a poluição? Explicite e verifique todas as hipóteses necessárias à resolução do problema, justificando convenientemente.

3.27. (*Exame 31.01.2011*) Para comparar dois tipos de máquina de ceifar (segadeiras) quanto à sua eficiência, foram selecionadas ao acaso 9 searas tendo sido cada uma dividida em dois lotes. Em cada seara uma das segadeiras foi atribuída ao acaso a um dos lotes, ficando a outra para o outro lote.

A eficiência é avaliada num intervalo de valores de 0 (eficiência mínima) a 10 (eficência máxima). Os valores registados relativos à eficiência de cada segadeira foram introduzidos num interpretador de Python. Responda às seguintes questões utilizando os resultados de alguns dos comandos apresentados abaixo.

a) Indique uma estimativa da eficiência média de cada uma das segadeiras.

b) Calculou-se um intervalo a 95% de confiança para a variância da eficiência da segadeira 1, tendo-se obtido $]0.5144; 4.1381[$. Qual o valor observado para a variância da segadeira 1?

c) Será que a segadeira 1 é mais eficiente do que a segadeira 2? Justifique convenientemente indicando e validando as condições necessárias à resolução.

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> # seg1 designa os valores observados de eficiência da segadeira 1
>>> # seg2 designa os valores observados de eficiência da segadeira 2
>>> n=9

>>> print(stat.mean(seg1))
  7.400000

>>> print(stat.mean(seg2))
  6.488889

>>> print(stat.variance(seg2))
  0.5861111

>>> D=np.array(seg1)-np.array(seg2)
>>> print(stat.variance(D))
  0.6761111

>>> print(shapiro(seg1))
```

```
  ShapiroResult(statistic=0.8848, pvalue=0.1762)

>>> print(shapiro(seg2))
  ShapiroResult(statistic=0.9105, pvalue=0.3193)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.9648, pvalue=0.8472)

>>> ceifar_res=ttest_rel(seg1,seg2,alternative='two-sided')
>>> ci = ceifar_res.confidence_interval(confidence_level=0.95)
>>> print(ceifar_res)
  TtestResult(statistic=3.3242, pvalue=0.01047, df=8)
>>> print(ci)
  ConfidenceInterval(low=0.2790663, high=1.5431559)

>>> print(ttest_rel(seg1,seg2,alternative='greater')
  TtestResult(statistic=3.3242, pvalue=0.005237, df=8)

>>> print(ttest_rel(seg1,seg2,alternative='less')
  TtestResult(statistic=3.3242, pvalue=0.9948, df=8)

>>> ceifar_res=ttest_ind(seg1,seg2,alternative='two-sided',equal_var=False)
>>> ci = ceifar_res.confidence_interval(confidence_level=0.95)
>>> print(ceifar_res)
  TtestResult(statistic=2.088, pvalue=0.05482, df=14.548)
>>> print(ci)
  ConfidenceInterval(low=-0.02147036, high=1.84369258)

>>> print(ttest_ind(seg1,seg2,alternative='less',equal_var=False))
  TtestResult(statistic=2.088, pvalue=0.9726, df=14.548)

>>> print(ttest_ind(seg1,seg2,alternative='greater',equal_var=False))
  TtestResult(statistic=2.088, pvalue=0.02741, df=14.548)
```

3.28. (*Exame17.01.2011*) Uma máquina de ensacar está regulada para encher sacos com 16 kg de açúcar. Para controlar o seu funcionamento escolheram-se, aleatoriamente, 15 sacos, que foram pesados. Os valores obtidos foram introduzidos em Python. Responda às seguintes questões utilizando, sempre que possível, os resultados dos comandos apresentados.

a) Indique uma estimativa para o peso médio e para a variância do peso de cada saco.

b) Foi calculado um intervalo de confiança para a variância do peso de cada saco, σ^2, tendo-se obtido $]0.01368\ ;\ 0.04931[$.

   i) Qual o estimador que foi utilizado na construção do intervalo?

   ii) Determine o nível de confiança do intervalo dado.

c) Que conclusão pode tirar quanto à regulação da máquina? Justifique convenientemente, indicando e validando os pressupostos necessários à sua resposta.

```
import statistics as stat
from scipy.stats import shapiro

peso=[16.1, 15.8, 15.9, 16.1, 15.8, 16.2, 16.0, 15.9, 16.0, 15.7, 15.8,
      15.7, 16.0, 16.0, 15.8]

print(stat.mean(peso)            print(stat.variance(peso))
  15.92                            0.02314286

print(shapiro(peso))
  ShapiroResult(statistic=0.9377, pvalue=0.3544)
```

3.29. Um enólogo pretende avaliar a acidez total de um vinho. Para isso seleciona aleatoriamente 20 garrafas de vinho na adega e mede a acidez através do método clássico e de um dispositivo de titulação automática. Alguns resultados das análises, em g/l, foram:

garrafa	1	2	3	···	18	19	20
método clássico	4.8	3.4	2.5	···	2.9	5.4	2.1
titulação automática	6.1	5.1	2.1	···	4.5	3.9	1.5

Os dados foram introduzidos em Python. Abaixo apresentam-se resultados de comandos, alguns inadequados. Responda às seguintes questões utilizando os resultados apresentados abaixo.

a) De acordo com a legislação em vigor um vinho de mesa deverá ter uma acidez total superior a 3.5 g/l. Com base nos resultados das análises efetuadas pelo método clássico, o enólogo poderá concluir que o seu vinho cumpre os requisitos de acidez impostos pela legislação? Explicite e valide os pressupostos necessários à resolução do problema.

b) Com base nos valores obtidos poder-se-á concluir que os dois métodos de análise da acidez total do vinho têm resultados significativamente diferentes? Explicite e valide os pressupostos necessários à resolução do problema.

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import ttest_1samp
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> classico=[4.8, 3.4, 2.5, 3.8, 4.3, 3.6, 3.5, 3.5, 4.0, 3.6, 6.3, 3.0,
              3.1, 3.7, 2.8, 5.1, 4.0, 2.9, 5.4, 2.1]
>>> automatico=[6.1, 5.1, 2.1, 4.9, 6.6, 4.0, 4.5, 1.0, 4.7, 3.5, 8.2,
                3.9, 3.6, 4.5, 3.5, 6.3, 4.7, 4.5, 3.9, 1.5]
>>> n=len(classico)

>>> print(stat.mean(classico))          >>> print(stat.mean(automatico))
  3.77                                    4.355
```

```
>>> print(stat.variance(classico))     >>> print(stat.variance(automatico))
  1.040105                             2.887868

>>> D=np.array(classico)-np.array(automatico)
>>> print(stat.variance(D))
  1.326605

>>> print(shapiro(classico))
  ShapiroResult(statistic=0.951, pvalue=0.3827)

>>> print(shapiro(automatico))
  ShapiroResult(statistic=0.9625, pvalue=0.5959)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.9163, pvalue=0.08413)

>>> classico_res=ttest_1samp(classico,popmean=3.5)
>>> print(classico_res)
  TtestResult(statistic=1.183968, pvalue=0.25103, df=19)
>>> print(classico_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=3.29269, high=4.24731)

>>> print(ttest_1samp(classico,popmean=3.5, alternative='greater'))
  TtestResult(statistic=1.183968, pvalue=0.12551, df=19)

>>> print(ttest_1samp(classico,popmean=3.5, alternative='less'))
  TtestResult(statistic=1.183968, pvalue=0.87449, df=19)

>>> vinho_res=ttest_ind(classico, automatico, equal_var=True)
>>> print(vinho_res)
  TtestResult(statistic=-1.320038, pvalue=0.194719, df=38.0)
>>> print(vinho_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=-1.4821486, high=0.3121486)

>>> vinho_res=ttest_rel(classico, automatico)
>>> print(vinho_res)
  TtestResult(statistic=-2.27143, pvalue=0.03493, df=19)
>>> print(vinho_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=-1.1240512776362028, high=-0.04594872)
```

3.30. 10 pessoas asmáticas participaram numa experiência destinada a estudar o efeito de um novo tratamento da função pulmonar. Durante 1 segundo registou-se o volume de ar expirado (em litros) antes e depois da aplicação do tratamento. Os resultados obtidos foram os seguintes:

	1	2	3	4	5	6	7	8	9	10
Antes	1.69	2.77	1.00	1.66	0.85	1.42	2.82	2.58	1.98	2.02
Depois	2.69	2.22	3.07	3.35	2.74	3.61	5.14	2.44	2.25	3.41

Nota: Para responder a esta pergunta utilize, sempre que possível, resultados que lhe são apresentados no Anexo.

a) Determine um intervalo a 99% de confiança para o volume médio de ar expirado por uma pessoa antes da aplicação do tratamento. Indique e valide os pressupostos necessários à sua resolução.

b) Qual é o erro máximo cometido ao usar a média da amostra para estimar o volume médio de ar expirado por pessoa, com o nível de confiança da alínea anterior?

c) Os dados recolhidos dão indicação de que o volume de ar expirado é superior após o tratamento? Justifique, indicando e validando os pressupostos necessários à resolução desta questão.

ANEXO

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> antes=[1.69,2.77,1.00,1.66,.85,1.42,2.82,2.58,1.98,2.02]
>>> depois=[2.69,2.22,3.07,3.35,2.74,3.61,5.14,2.44,2.25,3.41]
>>> n=len(antes)

>>> print(stat.mean(antes))          >>> print(stat.mean(depois))
  1.879                               3.092

>>> print(stat.variance(antes))      >>> print(stat.variance(depois))
  0.480743                            0.756751

>>> D=np.array(antes)-np.array(depois)
>>> print(stat.variance(D))
  1.0561122

>>> print(shapiro(antes))
  ShapiroResult(statistic=0.9402496, pvalue=0.555794)

>>> print(shapiro(depois))
  ShapiroResult(statistic=0.8633146, pvalue=0.083483)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.8971656, pvalue=0.203886)

>>> indep_res=ttest_ind(antes, depois, equal_var=True)
>>> print(indep_res)
```

```
  TtestResult(statistic=-3.44817399, pvalue=0.0028683, df=18.0)
>>> print(indep_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=-1.95206318, high=-0.4739368)

>>> print(ttest_ind(antes, depois, equal_var=True, alternative='less'))
  TtestResult(statistic=-3.44817399, pvalue=0.001434167, df=18.0)

>>> print(ttest_ind(antes, depois, equal_var=True, alternative='greater'))
  TtestResult(statistic=-3.44817399, pvalue=0.9985658, df=18.0)

>>> empar_res=ttest_rel(antes, depois)
>>> print(empar_res)
  TtestResult(statistic=-3.73255113, pvalue=0.004679, df=9)

>>> print(empar_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=-1.9481531, high=-0.4778469)

>>> print(ttest_rel(antes, depois, alternative='less'))
  TtestResult(statistic=-3.73255114, pvalue=0.0023395, df=9)

>>> print(ttest_rel(antes, depois, alternative='greater'))
  TtestResult(statistic=-3.732551138, pvalue=0.997660503, df=9)
```

3.31. Uma associação de defesa de consumidores pretende publicar um estudo comparativo dos preços de dois supermercados. Para isso selecionou 20 produtos e registou os preços de cada um nos supermercados A e B (em cêntimos):

Produto	1	2	3	4	5	6	⋯	15	16	17	18	19	20
Super. A	202	201	560	253	384	332	⋯	624	851	742	501	476	765
Super. B	185	187	516	239	349	295	⋯	613	851	731	546	490	762

Utilize os resultados apresentados abaixo para responder às seguintes questões.

a) Determine uma estimativa para a diferença dos preços médios nos dois supermercados.

b) Com base nos dados recolhidos, o que pode afirmar relativamente aos preços médios praticados pelos dois supermercados? Explicite e valide as hipóteses necessárias à resolução do problema.

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro

>>> A=[202, 201, 560, 253, 384, 332, 549, 722, 153, 676, 804,
   535, 472, 335, 624, 851, 742, 501, 476, 765.]
>>> B=[185, 187, 516, 239, 349, 295, 552, 667, 132, 676, 745,
```

```
   529, 460, 316, 613, 851, 731, 546, 490, 762]

>>> print(stat.mean(A))                >>> print(stat.mean(B))
  506.85                                 492.05

>>> print(round(stat.variance(A),3))   >>> print(round(stat.variance(B),3))
  46116.766                              46454.682

>>> print(round(stat.variance(np.array(A)-np.array(B)),3))
  568.379

>>> print('statistic=%.4f, pvalue=%.4f' %shapiro(A))
  statistic=0.9570, pvalue=0.4864

>>> print('statistic=%.4f, pvalue=%.4f' %shapiro(B))
  statistic=0.9552, pvalue=0.4521

>>> print('statistic=%.4f, pvalue=%.4f' %shapiro(np.array(A)-np.array(B)))
  statistic=0.9514, pvalue=0.3889
```

3.32. (*Exame 27.01.2014*) Uma empresa agrícola pretende adquirir um gerador. Face a dois tipos de geradores (que designamos por ger1 e ger2), de preços semelhantes, quer decidir comparando a produção de energia elétrica, em kWh, de cada um dos geradores. Recolheu uma amostra de 25 observações da quantidade de energia elétrica produzida por cada gerador, em diferentes situações. Os dados foram introduzidos num interpretador de Python. Considere os seguintes resultados referentes ao gerador tipo 1:

```
>>> import statistics as stat
>>> from scipy.stats import shapiro

>>> print(stat.mean(ger1))              >>> print(stat.variance(ger1))
  9.6056                                 18.90528
>>> print(shapiro(ger1))
  ShapiroResult(statistic=0.9514, pvalue=0.2698)
```

a) Indique uma estimativa da variância da produção de energia elétrica do gerador 1. Explicite o estimador que lhe está associado.

b) Poder-se-á considerar que a variância da produção de energia elétrica no gerador 1 é superior a 15$(\text{kWh})^2$? Justifique convenientemente a sua resposta.

c) Para obter um intervalo de confiança a 95% para a diferença entre as produções médias de energia elétrica dos geradores 1 e 2, utilizaram-se as instruções que se apresentam de seguida. Alguns valores estão omissos e foram substituídos por $\star\star\star$:

```
>>> res=ttest_***(ger1, ger2, equal_var=True)
>>> print(res)
  TtestResult(statistic=-0.6632, pvalue=0.5104, df=***)
```

```
>>> print(res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=-3.009280, high=***)

>>> print(stat.mean(ger2))
  10.3520
```

i) Qual a diferença entre as médias observadas para a produção de energia elétrica dos dois geradores? Qual o erro máximo que se cometeria, a 95% de confiança, ao identificar esta diferença com a verdadeira diferença entre as produções médias de energia elétrica dos geradores?

ii) Complete o *output* indicando a informação em falta nos três locais assinalados por $***$.

iii) Que pressupostos deverão ser verificados para construir o intervalo de confiança dado no *output*?

iv) O que pode afirmar, a 95% de confiança, sobre a diferença entre as produções médias de energia elétrica dos geradores?

v) A conclusão da alínea anterior seria alterada se considerasse uma confiança de 99%? Justifique.

3.33. (*Exame 26.01.2015*) Para estudar o peso de uma certa espécie de peixe, recolheram-se dados do peso (g) de 15 peixes fêmea e de 15 peixes macho. Os resultados obtidos foram introduzidos em Python. Utilize, sempre que possível, os resultados apresentados no Anexo para responder às seguintes questões.

a) Pretende-se estudar a variabilidade do peso dos peixes fêmea.

   i) Face aos dados recolhidos indique uma estimativa da variância do peso dos peixes fêmea.

   ii) Qual o estimador associado à estimativa obtida em i)? Explicite-o e mostre que é um estimador centrado da variância populacional.

b) Justificando convenientemente, diga se os dados recolhidos são compatíveis com as seguintes conjeturas:

   i) O peso de um peixe fêmea daquela espécie segue uma distribuição normal.

   ii) O peso médio de um peixe fêmea daquela espécie é 1000 g.

c) Comente, justificando convenientemente, a seguinte afirmação "O peso médio das fêmeas é superior ao dos machos".

ANEXO

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_1samp
>>> from scipy.stats import ttest_ind
```

```
>>> from scipy.stats import ttest_rel

>>> femea=[1007,  965, ...  978, 1020]
>>> macho= [934, 1014, ... 1017, 1023]
>>> n=15

>>> print(stat.mean(femea))          >>> print(stat.mean(macho))
  1011.0000                           983.1333

>>> print(stat.variance(femea))      >>> print(stat.variance(macho))
  679.5714                            765.6952

>>> D=np.array(femea)-np.array(macho)
>>> print(stat.variance(D))
  1629.695

>>> print(shapiro(femea))
  ShapiroResult(statistic=0.9545, pvalue=0.5977)

>>> print(shapiro(macho))
  ShapiroResult(statistic=0.9371, pvalue=0.3477)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.9638, pvalue=0.7581)

>>> F_calc=stat.variance(femea)/stat.variance(macho)
>>> print(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)))
  0.8265

>>> print(ttest_1samp(femea,popmean=1000, alternative='greater'))
  TtestResult(statistic=1.6343, pvalue=0.06224, df=14)

>>> femea_res=ttest_1samp(femea,popmean=1000)
>>> print(femea_res)
  TtestResult(statistic=1.6343, pvalue=0.1245, df=14)
>>> print(femea_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=996.5637, high=1025.4363)

>>> print(ttest_ind(femea, macho, equal_var=True, alternative='greater'))
  TtestResult(statistic=2.8389, pvalue=0.004167, df=28)

>>> print(ttest_rel(femea, macho, alternative='greater'))
  TtestResult(statistic=2.6735, pvalue=0.009088, df=14)
```

3.34. (*Exame 25.01.2019*) Realizou-se um estudo de investigação em modelos animais (ratinhos), onde se pretende conhecer a importância de algumas proteínas do sistema imune na infeção com malária na gravidez. Quantificou-se a expressão relativa dos genes associados a uma proteína, A, em placentas de 6 fêmeas não infetadas (NI) e 6 fêmeas infetadas (INF), tendo-se obtido os seguintes indicadores:

	Média	Variância
NI	1.520	0.612
INF	0.347	0.131

O cientista suspeita que a expressão relativa dos genes da proteína A pode sofrer alterações, em média, com a infeção.

Os resultados da experiência foram analisados com recurso à linguagem Python e estão apresentados no Anexo. Utilize-os, sempre que possível, para responder de forma completa às questões apresentadas a seguir.

a) Utilizando um intervalo de confiança adequado (consultar o Anexo) esclareça a dúvida do cientista. Responda de forma completa à questão, justificando convenientemente a sua resposta.

b) Um parâmetro de interesse neste estudo é a variabilidade da expressão relativa dos genes quando há infeção.

   i) Indique o estimador associado ao estudo desse parâmetro?

   ii) Com base nos dados recolhidos poder-se-á admitir que o valor daquele parâmetro é inferior a 0.3? Justifique convenientemente.

ANEXO

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_1samp
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> NI=[...]
>>> INF=[...]
>>> n=6

>>> F_calc=stat.variance(NI)/stat.variance(INF)
>>> print(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)))
  0.1159

>>> print(shapiro(NI))
  ShapiroResult(statistic=0.92543, pvalue=0.5453)

>>> print(shapiro(INF))
```

```
  ShapiroResult(statistic=0.79852, pvalue=0.05699)

>>> print(shapiro(np.array(NI)-np.array(INF)))
  ShapiroResult(statistic=0.9378, pvalue=0.6414)

>>> indep_res=ttest_ind(NI, INF, equal_var=True)
>>> print(indep_res)
  TtestResult(statistic=3.334, pvalue=0.007566, df=10)
>>> print(indep_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=0.3891588, high=????)

>>> emp_res=ttest_rel(NI, INF)
>>> print(emp_res)
  TtestResult(statistic=2.66, pvalue=0.04488, df=5)
>>> print(emp_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=0.03945195, high=????)
```

3.35. (*Exame 09.01.2020*) Num estudo sobre o efeito da congelação no valor nutricional de legumes e frutos, determinou-se o valor de um índice nutricional por caloria ingerida para cada um de 13 alimentos (legumes ou frutos) frescos e congelados. Os dados obtidos (ilustrados abaixo) foram introduzidos nas listas `fresco` e `congel` do Python.

Legume/Fruto	amora	espinafre	$\cdots$	pêssego
Fresco	0.731	3.903	$\cdots$	0.657
Congelado	0.490	3.490	$\cdots$	0.865

Responda às seguintes questões utilizando, sempre que possível, o *output* apresentado em Anexo.

a) Face à amostra observada obteve-se o intervalo de confiança $]1.226,\ 2.223[$ para o valor esperado do índice nutricional, μ, de um alimento fresco. Responda às seguintes questões apresentando os cálculos efetuados:

   i) Determine o índice nutricional médio na amostra dos alimentos frescos.

   ii) Qual o erro máximo cometido ao usar a média da amostra para estimar μ.

   iii) Determine a confiança deste intervalo, justificando convenientemente.

b) Poder-se-á afirmar que a congelação dos legumes e frutos provoca uma redução do índice nutricional por caloria ingerida? Explicite e valide os pressupostos necessários.

ANEXO

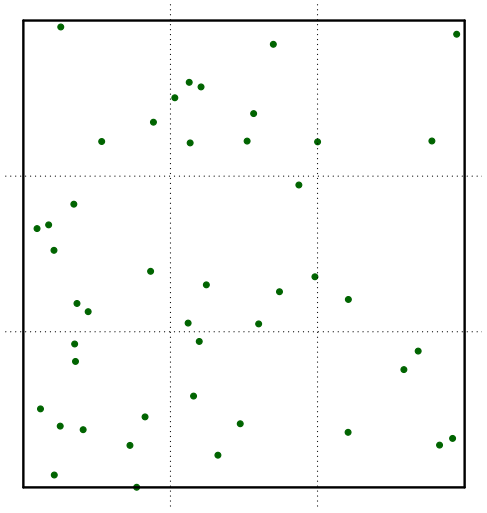
```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> fresco=[0.731, 3.903, ... 0.657]
>>> congel=[0.490, 3.490, ... 0.865]

>>> print(stat.variance(fresco))     >>> print(stat.variance(congel))
  1.017156                            0.8771506

>>> D=np.array(fresco)-np.array(congel)
>>> print(stat.variance(D))
  0.09266823

>>> print(shapiro(fresco))
  ShapiroResult(statistic=0.89088, pvalue=0.1003)

>>> print(shapiro(congel))
  ShapiroResult(statistic=0.92141, pvalue=0.2621)

>>> print(shapiro(D))
  ShapiroResult(statistic=0.89526, pvalue=0.1151)

>>> emp_res=ttest_rel(fresco, congel)
>>> print(emp_res)
  TtestResult(statistic=3.21, pvalue=0.007493, df=12)
>>> print(emp_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=0.0870592, high=0.4549710)

>>> print(ttest_rel(fresco, congel, alternative='greater')
  TtestResult(statistic=3.21, pvalue=0.003746, df=12)

>>> print(ttest_rel(fresco, congel, alternative='less')
  TtestResult(statistic=3.21, pvalue=0.9963, df=12)

>>> indep_res=ttest_ind(fresco, congel, equal_var=True)
>>> print(indep_res)
  TtestResult(statistic=0.70997, pvalue=0.4846, df=24)
>>> print(indep_res.confidence_interval(confidence_level=0.95))
  ConfidenceInterval(low=-0.5168317, high=1.0588619)

>>> print(ttest_ind(fresco, congel, equal_var=True, alternative='greater'))
  TtestResult(statistic=0.70997, pvalue=0.2423, df=24)

>>> print(ttest_ind(fresco, congel, equal_var=True, alternative='less'))
  TtestResult(statistic=0.70997, pvalue=0.7577, df=24)
```

Testes χ^2 de Pearson

Em geral adaptados de **Exercícios de Estatística e Delineamento 2019/20** de Jorge Cadima, disponível em `https://fenix.isa.ulisboa.pt/courses/estdel-5-283463546569515/aulas-praticas`

3.36. Galinhas homozigóticas de penas brancas são cruzadas com galos homozigóticos de penas pretas, produzindo uma primeira geração em que a conjugação dos dois alelos (de penas brancas e de penas pretas) produz penas de côr azul. Na segunda geração (F_2), e segundo a teoria genética, seria de esperar que $1/4$ dos pintos tenha penas brancas, $1/4$ tenha penas pretas e $2/4$ tenha penas azuis. Foi realizada uma experiência nos moldes acima indicados, verificando-se na segunda geração 36 pintos de penas brancas, 32 pintos de penas pretas e 73 de penas azuis. Verifique se estes valores observados são compatíveis com a teoria genética (utilize um nível de significância $\alpha = 0.05$).

3.37. Suspeita-se que uma certa espécie de planta se distribui uniformemente no terreno. Para averiguar esta hipótese marcou-se a posição das plantas encontradas numa região do seu *habitat*, no total de 45. Com base nos dados recolhidos, representados na figura abaixo, o que pode concluir?
Sugestão: divida o terreno em quadrículas de igual área e averigúe se o número de plantas por quadrícula está de acordo com a hipótese de uniformidade.



3.38. Um investigador afirma que numa determinada cultivar de videira, com um esquema de condução padronizado, o número médio de cachos por pé é igual a 4. É sugerido que se modele o número de cachos por pé através duma distribuição de Poisson com parâmetro $\lambda = 4$. A fim de avaliar essa hipótese, foram amostrados 200 pés e contados os cachos em cada pé, tendo sido obtidos os resultados indicados na tabela. Teste a hipótese indicada com base nestes resultados.

No. cachos	0	1	2	3	4	5	6	7	8	¿8
No. de pés	2	20	29	47	54	29	14	4	1	0

3.39. Uma determinada cultura encontra-se atacada pelo fungo F. Foram seleccionadas aleatoriamente 100 plantas e classificadas quanto à intensidade da infestação. A intensidade da infestação é medida da seguinte forma: a planta é dividida em quatro partes iguais e conta-se o número de partes onde se manifesta a infestação. No quadro figuram o número de plantas classificadas em cada uma das classes de intensidade. Teste se é admissível considerar que a distribuição Binomial com $p = 0.455$ é um modelo aceitável para descrever a intensidade do ataque do referido fungo. Utilize os níveis de significância 0.05 e 0.01, comentando os resultados nos dois casos. Discuta o valor de prova (*p-value*) associado ao valor da estatística. Comente.

Grau de ataque	0	1	2	3	4
No. de plantas	15	20	40	18	7

3.40. A porosidade de cascas de ovo pode ser medida em termos do número de poros por mm^2 de superfície. Uma experiência aleatória seleciona 320 bocados de casca e procede-se à contagem do número de poros em $1\,mm^2$ de cada bocado. Os resultados obtidos são dados na tabela de frequências. Será admissível considerar que a variável aleatória que indica o número de poros por milímetro quadrado de casca segue uma distribuição de Poisson com valor esperado 2? Justifique adequadamente. Calcule o valor de prova associado ao valor calculado da estatística.

No. de poros	0	1	2	3	4	5
Frequência	18	74	139	70	17	2

3.41. É formulada a hipótese de que a côr de ervilheiras é determinada por um gene, e o tipo de superfície é determinado por outro gene. Admite-se ainda que a côr amarela é dominante da verde e a superfície lisa é dominante da enrugada. Admitindo a segregação independente dos dois genes (isto é, que o alelo herdado do gene da côr é independente do alelo herdado do gene que controla o tipo de superfície), seria de esperar que do cruzamento de ervilheiras heterozigóticas relativamente à côr da semente e tipo de superfície, se observassem as proporções de $9/16$ de ervilheiras amarelas lisas, $3/16$ amarelas rugosas, $3/16$ verdes lisas e $1/16$ verdes rugosas.

a) Foi realizada uma experiência, cruzando-se ervilheiras heterozigóticas relativamente à côr da semente e tipo de superfície. Os resultados obtidos na descendência são indicados na tabela seguinte.

	Superfície	
Côr	Lisa	Rugosa
Amarelas	556	184
Verdes	193	61

Verifique se os resultados obtidos são compatíveis com as hipóteses genéticas acima formuladas, ao nível de significância $\alpha = 0.05$.

b) Repita a alínea anterior admitindo que a descendência era 30 vezes maior, mas que as proporções observadas em cada célula da tabela de contingências se mantinham iguais. Comente os seus resultados. Que lição geral se pode extrair destes resultados?

3.42. Cruzaram-se duas linhas puras de cobaias, sendo os progenitores masculinos de pelos curtos e côr amarela e os progenitores femininos de pelos longos e côr branca. A característica pelos curtos é dominante relativamente a pelos longos. Quanto à côr, um genótipo híbrido (isto é, com um alelo de côr amarela e outro de côr branca) terá a côr creme. Admitindo a segregação independente de cada gene, a teoria genética prevê que numa segunda geração (F_2) do referido cruzamento, seriam de esperar as seguintes proporções: $6/16$ de cobaias de pelo curto e côr creme; $3/16$ de pelo curto e côr amarela; $3/16$ de pelo curto e côr branca; $2/16$ de pelo longo e côr creme; $1/16$ de pelo longo e côr amarela; e $1/16$ de pelo longo e côr branca.

Uma experiência realizada nestas condições produziu os seguintes resultados:

	Côr		
Pelo	Creme	Amarelo	Branco
Curto	178	93	89
Longo	62	29	31

Verifique se estes resultados são compatíveis com o mecanismo genético referido, ao nível de significância $\alpha = 0.05$.

3.43. Considere a experiência descrita no Exercício 3 dos Exercícios Introdutórios, e referente ao enraizamento de estacas semi-lenhosas de oliveiras, submetidas a quatro diferentes tratamentos. Em cada tratamento, foram ensaiadas 60 estacas. Os resultados obtidos foram os seguintes:

	Resultado			
Tratamento	Morte	Com calo	Enraizamento	Total
Sem incisão/sem boro	26	18	16	60
Com incisão/sem boro	32	22	6	60
Sem incisão/com boro	24	24	12	60
Com incisão/com boro	39	19	2	60
Total	121	83	36	240

a) Teste se é possível admitir que a distribuição das observações pelas três categorias é igual para os quatro tratamentos.

b) Em caso de concluir que diferentes tratamentos estão associados a diferentes distribuições dos resultados, identifique quais as combinações de tratamento/resultado que mais contribuem para essas diferenças. Interprete e comente.

3.44. Uma direção regional necessita identificar, na sua região, três tipos de ocupação de solo agrícola: cultura de sequeiro, cultura de regadio e não cultivado. Para tal, envia equipas ao terreno para obter essa informação, o que acarreta custos elevados. Uma pequena empresa propõe fornecer, por um preço muito menor, um mapa de ocupação do solo, obtido por análise de imagens de satélite. Para avaliar se o mapa da empresa é útil, os técnicos da direção regional escolhem ao acaso 100 parcelas de terreno da região e comparam na tabela abaixo a classificação realizada pelas equipas no terreno (linhas) e a classificação fornecida pelo mapa da empresa (colunas).

		satélite		
		não cultivado	sequeiro	regadio
terreno	não cultivado	16	15	4
	sequeiro	15	22	3
	regadio	0	5	20

Efetue um teste de independência à tabela e comente os seus resultados.

Exercícios de Revisão de Inferência Estatística

R3.1. (*Exame 9.01.2020*) Considere a v.a. X com distribuição uniforme contínua no intervalo $]0,\theta[$, $\theta > 0$. Seja $\overline{X}$ a média da amostra aleatória $(X_1, \cdots, X_n)$ e $T = 2\overline{X}$ um estimador de θ.

a) Defina "amostra aleatória de dimensão n".

b) Mostre que T é um estimador centrado de θ.

c) Determine $Var[T]$.

R3.2. Recolheu-se uma amostra aleatória de dimensão 5 de uma população normal. Determine a probabilidade de o desvio padrão da amostra ser inferior ao desvio padrão da população.

R3.3. (*Exame 12.01.2009*) A quantidade (em ppm) de um poluente no solo de uma certa região é uma v.a. X que se admite ter distribuição normal com valor médio μ e variância σ^2, desconhecidos.

Recolheu-se uma amostra de dimensão 10 tendo-se obtido o desvio padrão de 1.3943 ppm. Com base na amostra construiu-se o seguinte intervalo de confiança para μ

$$]22.447\ ;\ 25.313[$$

a) Indique estimativas do valor médio e da variância da quantidade daquele poluente no solo.

b) Qual o grau de confiança daquele intervalo?

c) Construa um intervalo a 95% de confiança para σ^2.

R3.4. Num certo processo químico é muito importante que uma dada solução tenha um pH de exatamente 8.20. O método utilizado na determinação do pH fornece medições que se admite terem distribuição normal de valor médio igual ao verdadeiro valor do pH da solução e desvio padrão de 0.02.

Para avaliar o pH de uma solução, efetuaram-se 10 medições independentes tendo-se obtido os seguintes valores:

8.18 8.16 8.17 8.22 8.19 8.17 8.15 8.21 8.16 8.18

a) Indique uma estimativa do valor médio do pH da solução.

b) Com base nestas 10 medições, o que pode concluir relativamente à utilização desta solução no referido processo químico?

c) Pretende-se efetuar um novo conjunto de medições para diminuir o erro máximo cometido na estimativa do verdadeiro valor do pH da solução. Mantendo-se todas as condições referidas acima, qual deverá ser o tamanho da amostra para que aquele erro máximo não exceda 0.01, a 95% de confiança?

R3.5. Faz-se uma experiência para saber se dois regimes alimentares A e B produzem o mesmo aumento de peso nos animais, durante um período de tempo fixado. Tomam-se 20 animais e de entre eles 10 ao acaso aos quais é dado o alimento A. Aos outros 10 é dado o alimento B. Os aumentos de peso (expressos em kg) no mesmo intervalo de tempo são os seguintes:

Regime A	-2.0	0.0	4.2	6.3	9.6	4.3	10.2	11.0	12.4	13.1
Regime B	4.0	6.0	8.0	11.3	12.3	14.4	14.5	14.7	14.7	16.0

Diga se existe diferença significativa entre os dois regimes alimentares, justificando convenientemente todas as hipóteses necessárias à resolução do problema.

R3.6. Um estudo pretende comparar um tipo de semente melhorada com o tipo de semente usado anteriormente. A nova semente passará a ser utilizada se, em média, o crescimento das plantas após 20 dias fôr superior ao das obtidas das 'velhas' sementes. São criadas 15 diferentes situações laboratoriais, variando temperatura e humidade. Em cada situação semeia-se uma semente de cada tipo e obtêm-se os seguintes resultados para o crescimento (em cm) das plantas após 20 dias :

Situação	1	2	3	4	5	6	7	8
Sementes melhoradas	3.46	3.48	2.74	2.83	4.00	4.95	2.24	6.92
sementes tradicionais	3.18	3.67	2.92	3.10	4.10	4.86	2.21	6.91

Situação	9	10	11	12	13	14	15
sementes melhoradas	6.57	6.18	8.30	3.44	4.47	7.59	3.87
sementes tradicionais	6.83	6.19	8.05	3.46	4.18	7.43	3.85

Deverá passar a usar-se as sementes melhoradas? Responda justificando e explicitando quaisquer hipóteses adicionais que seja necessário impôr.

R3.7. Num estudo sobre a incidência de certa doença numa população de insetos, um grupo de biólogos registou ao longo de um ano o número de insetos contaminados em cada amostra de 5 insetos, tendo para tal recolhido 100 amostras. Os resultados obtidos foram:

Num. de insetos contaminados	0	1	2	3	4	5
Num. de amostras	9	26	34	22	8	1

a) Construa uma tabela de frequências absolutas, relativas e relativas acumuladas para o número de insetos contaminados por amostra.

b) Determine a média e a mediana do número de insetos contaminados por amostra.

c) Considerando agora a totalidade de insetos observados como uma amostra aleatória da população de insetos:

   i) Determine uma estimativa da verdadeira proporção de insetos contaminados.

   ii) Qual é o erro máximo associado à estimativa obtida na alínea anterior, com uma confiança de 95% ?

R3.8. O dono de uma ervanária produz um chá relativamente ao qual afirma que é 90% eficaz para curar dores de cabeça. Num inquérito feito a 250 pessoas, 198 concordaram que o chá cura as dores de cabeça. Acha que o resultado do inquérito é compatível com a pretensão do produtor ?

R3.9. Um investigador pretende estudar a incidência a nível nacional, de uma doença que ataca os pinheiros. Observações efetuadas através do país resultaram em 1233 casos de pinheiros afetados (a nível nacional) num total de 4250 observações.

a) Estime a percentagem de pinheiros afetados a nível nacional.

b) Determine um intervalo a 95% de confiança para a verdadeira proporção de pinheiros afetados.

R3.10. (*Exame 16.01.2012*) Em várias espécies de animais, altos níveis de testosterona tornam os machos mais atrativos para as fêmeas. No entanto, conjetura-se que altos níveis de testosterona podem enfraquecer o sistema imunitário, i.e., reduzem o número de anticorpos. Para responder a esta questão consideraram-se 13 melros de asa-vermelha, em cada um dos quais foi implantado um tubo contendo testosterona. Em cada melro avaliou-se o número de anticorpos no sangue antes e após o implante. Os resultados obtidos, em unidades convenientes, foram introduzidos em Python.

Utilize, sempre que possível, os resultados apresentados no **Anexo** para responder às seguintes questões (alguns dos procedimentos apresentados são inadequados).

a) Indique estimativas para o valor médio do número de anticorpos antes e depois do implante.

b) Determine um intervalo de confiança a 95% para a variabilidade do número de anticorpos antes da introdução do implante.

c) A conjetura de que o aumento do nível de testosterona enfraquece o sistema imunitário será compatível com os resultados obtidos? Explicite e valide os pressupostos necessários à resolução do problema.

ANEXO

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> antes=[4.65,4.21,4.91,4.50,4.80,5.01,4.88,4.78,4.98,4.41,4.75,4.70,4.93]
>>> depois=[4.54,3.80,4.98,4.45,5.00,5.00,4.35,4.96,5.02,4.73,4.77,4.60,5.01]
>>> n=len(antes)

>>> print('%.4f' %stat.mean(antes))       >>> print('%.4f' %stat.mean(depois))
  4.7315                                    4.7085

>>> print('%.4f' %stat.variance(antes))   >>> print('%.4f' %stat.variance(depois))
  0.0564                                    0.1300

>>> D=np.array(antes)-np.array(depois)
>>> print('%.4f' % stat.variance(D))
  0.0546

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(antes))
  ShapiroResult(statistic=0.9196, pvalue=0.2477)

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(depois))
  ShapiroResult(statistic=0.8310, pvalue=0.0163)

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(D))
  ShapiroResult(statistic=0.9169, pvalue=0.2275)

>>> F_calc=stat.variance(antes)/stat.variance(depois)
```

```
>>> print(round(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)),4))
  0.1622

>>> T_res=ttest_ind(antes,depois,equal_var=True)
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=0.1927, pvalue=0.8488, df= 24.0 )

>>> ci = T_res.confidence_interval(confidence_level=0.95)
>>> print('ConfidenceInterval(low=%.4f, high=%.4f)'%ci)
  ConfidenceInterval(low=-0.1181, high=0.1642)

>>> T_res=ttest_ind(antes,depois,equal_var=True,alternative='less')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=0.1927, pvalue=0.5756, df= 24.0 )

>>> T_res=ttest_ind(antes,depois,equal_var=True,alternative='greater')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=0.1927, pvalue=0.4244, df= 24.0 )

>>> T_res=ttest_rel(antes,depois)
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=0.3562, pvalue=0.7279, df= 12 )

>>> ci = T_res.confidence_interval(confidence_level=0.95)
>>> print('ConfidenceInterval(low=%.4f, high=%.4f)'%ci)
  ConfidenceInterval(low=-0.1181, high=0.1642)

>>> T_res=ttest_rel(antes,depois,alternative='less')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=0.3562, pvalue=0.6361, df= 12 )

>>> T_res=ttest_rel(antes,depois,alternative='greater')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=0.3562, pvalue=0.3639, df= 12 )
```

R3.11. Um agricultor pretende experimentar dois tipos (I e II) de semeador de milho para comparar a produtividade das duas máquinas.

Para isso, num campo agrícola marcaram-se aleatoriamente 10 parcelas de igual área sendo cada uma delas dividida em duas secções iguais. Em cada parcela sorteou-se a atribuição de um dos tipos de semeador a uma das secções.

A produtividade (em unidades adequadas) de cada um dos semeadores foi registada, e os dados foram introduzidos num interpretador de Python. Utilizando os resultados apresentados abaixo (alguns desnecessários), poderá o agricultor admitir que a produtividade esperada das duas máquinas é igual? Justifique convenientemente a sua resposta.

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro

>>> x=[5.6, 8.4, 8.0, 8.6, 6.4, 7.8, 6.2, 7.7, 8.0, 7.7] #semeador tipo I
>>> y=[6.0, 7.4, 7.3, 7.5, 6.4, 6.0, 5.5, 6.6, 5.6, 6.1] #semeador tipo II

>>> print('%.4f' %stat.mean(x))
  7.4400

>>> print('%.4f' %stat.mean(y))
  6.4400
```

```
>>> print('%.4f' %stat.variance(x))          >>> print('%.4f' %stat.variance(y))
  1.0182                                        0.5449

>>> D=np.array(x)-np.array(y)
>>> print('%.4f' % stat.variance(D))
  0.6800

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(x))
ShapiroResult(statistic=0.8731, pvalue=0.1087)

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(y))
ShapiroResult(statistic=0.9026, pvalue=0.2341)

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(D))
ShapiroResult(statistic=0.9765, pvalue=0.9435)
```

R3.12. (*Exame 31.01.2012*) Um estudante de agronomia pretende estudar a produção de uma certa variedade de milho numa dada região. Para esse efeito, feita a colheita, pesou 50 espigas e organizou os resultados na tabela abaixo:

Peso(g)	x_i'	n_i	N_i	f_i	F_i
] 180, 190]	185	3			
] 190, 200]			12		
] 200, 210]		9		0.18	
] 210, 220]			30		
] 220, 230]		12			0.84
] 230, 240]		8	50		

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

b) Diga o que representa o símbolo no cabeçalho de cada coluna e complete a tabela. Esboce o histograma associado.

c) Calcule um valor aproximado da média e da mediana do peso das espigas.

d) Indique uma estimativa da percentagem de espigas com peso superior a 220 g.

e) O estudante tirou a seguinte conclusão: "Esta variedade de milho tem uma boa produção porque mais de 30% das espigas têm peso superior a 220 g". Está de acordo com o seu colega? Justifique convenientemente.

f) Suponha que outro estudante quer estudar a produção dessa variedade de milho noutra região. Para isso decidiu colher e pesar 50 espigas verificando que 22 tinham peso superior a 220 g. Poder-se-á dizer que a percentagem de espigas com peso superior a 220 g é diferente nas duas regiões? Justifique.

R3.13. Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória retirada de uma população com distribuição normal com valor médio μ e variância σ^2 conhecida.

a) Qual é a probabilidade de o intervalo aleatório

$$\left]\overline{X}-1.96\frac{\sigma}{\sqrt{n}}\,;\overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right[$$

conter o valor μ?

b) Qual é o erro máximo cometido ao usar $\bar{x}$ para estimar μ?

c) Qual é a redução do erro máximo que se obtém quando a dimensão da amostra duplica?

d) Mostre que $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ é um estimador centrado de σ^2.

R3.14. (*Exame 14.01.2013*) Seja X uma variável aleatória com valor esperado μ e variância σ^2 desconhecidos. Recolheram-se aleatoriamente 15 dados da população X, que se introduziram em Python na lista `x`. Responda às seguintes questões utilizando, sempre que possível, o *output* apresentado em Anexo.

a) Pretende-se estimar o valor médio, μ, da população X.

   i) Mostre que $\overline{X}$, média de uma amostra aleatória de dimensão n retirada de X, é um estimador centrado de μ. Exprima a variância de $\overline{X}$ como função dos parâmetros da população.

   ii) Com base na amostra recolhida indique estimativas do valor médio de X e da variância de $\overline{X}$.

b) Pretende-se averiguar se a variância da população é superior a 1, pelo que se decide realizar um teste de hipóteses.

   i) Formule as hipóteses do teste que permite responder a esta questão.

   ii) Indique a estatística de teste e a sua distribuição. Valide os pressupostos necessários à existência dessa distribuição.

   iii) Conclua o teste de hipóteses.

c) Recolheu-se uma amostra de dimensão 15 de outra população Y. Os dados foram introduzidos na lista `y`.

   i) Indique um intervalo de confiança a 95% para a diferença das médias das populações X e Y. Explicite e valide os pressupostos necessários.

   ii) Os dados recolhidos evidenciam que as médias das duas populações são iguais? Justifique convenientemente.

ANEXO

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel

>>> x=[ ... ]
```

```
>>> y=[ ... ]
>>> n=15

>>> print('%.4f' %stat.mean(x))        >>> print('%.4f' %stat.mean(y))
  5.6867                                  4.6467

>>> print('%.4f' %stat.variance(x))    >>> print('%.4f' %stat.variance(y))
  1.0841                                  1.8555

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(x))
  ShapiroResult(statistic=0.9747, pvalue=0.6733)

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(y))
  ShapiroResult(statistic=0.9577, pvalue=0.2704)

>>> D=np.array(x)-np.array(y)
>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(D))
  ShapiroResult(statistic=0.9240, pvalue=0.2214)

>>> F_calc=stat.variance(x)/stat.variance(y)
>>> print(round(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)),4))
  0.3261

>>> T_res=ttest_ind(x, y, equal_var=True)
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=2.3493, pvalue=0.0261, df= 28 )

>>> ci = T_res.confidence_interval(confidence_level=0.95)
>>> print('ConfidenceInterval(low=%.4f, high=%.4f)'%ci)
  ConfidenceInterval(low=0.1332, high=1.9468)

>>> T_res=ttest_rel(x, y)
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=2.1864, pvalue=0.04627, df= 14 )

>>> ci = T_res.confidence_interval(confidence_level=0.95)
>>> print('ConfidenceInterval(low=%.4f, high=%.4f)'%ci)
  ConfidenceInterval(low=0.01978, high=2.0602)
```

R3.15. (*Exame 23.01.2020*) Num inventário florestal de montado de sobro pretende-se comparar o diâmetro das árvores em duas parcelas, A e B. Para isso em cada uma das parcelas registou-se o diâmetro de 25 árvores, em cm. Os dados obtidos foram introduzidos nas listas `parcelaA` e `parcelaB` do Python.

Responda às seguintes questões utilizando, sempre que possível, o *output* apresentado em Anexo.

a) Considere a v.a. X, de valor médio μ e variância σ^2 desconhecidos, que designa o diâmetro dos sobreiros na parcela A. Seja $(X_1, \ldots, X_{25})$ a amostra aleatória retirada de X. Sugerem-se

dois estimadores para μ,

$$T_1 = \frac{\sum_{i=1}^{25} X_i}{25} \qquad \text{e} \qquad T_2 = \frac{2X_1 + 8X_{25}}{10},$$

i) Mostre que T_1 e T_2 são estimadores centrados de μ e obtenha as respetivas variâncias.

ii) Com base nos dados recolhidos indique estimativas para o diâmetro médio dos sobreiros na parcela A, associadas aos estimadores T_1 e T_2.

iii) Determine um intervalo de confiança a 95% para a variância do diâmetro dos sobreiros na parcela A. Responda de forma completa.

b) Face aos resultados obtidos poder-se-á concluir que as árvores da parcela B, apresentam em média um diâmetro superior ao dos sobreiros da parcela A? Explicite e valide os pressupostos necessários à resolução do problema.

ANEXO

```
>>> import numpy as np
>>> import statistics as stat
>>> from scipy.stats import shapiro
>>> from scipy.stats import f
>>> from scipy.stats import ttest_ind
>>> from scipy.stats import ttest_rel
>>> n=25
>>> parcelaA=[45.00, 36.61, 41.70, ...  32.47, 38.52, 38.20,
             36.92, 21.96, 19.42, ...  19.10, 52.20, 28.65, 39.79]
>>> parcelaB=[33.74, 58.89, 32.15, ...  45.84, 49.02, 29.60,
             58.57, 47.75, 62.39, ...  36.29, 37.56, 41.06, 29.92]

>>> print('%.4f' %stat.mean(parcelaA))       >>> print('%.4f' %stat.mean(parcelaB))
  35.1520                                     43.4812

>>> print('%.4f' %stat.variance(parcelaA))   >>> print('%.4f' %stat.variance(parcel
  82.5712                                     93.6749

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(parcelaA))
  ShapiroResult(statistic=0.9352, pvalue=0.1147)

>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(parcelaB))
  ShapiroResult(statistic=0.9757, pvalue=0.7895)

>>> D=np.array(parcelaA)-np.array(parcelaB)
>>> print('ShapiroResult(statistic=%.4f, pvalue=%.4f)' %shapiro(D))
  ShapiroResult(statistic=0.9528, pvalue=0.2890)

>>> F_calc=stat.variance(parcelaA)/stat.variance(parcelaB)
>>> print(round(2*min(f.cdf(F_calc,n-1,n-1), 1-f.cdf(F_calc,n-1,n-1)),4))
  0.7598
```

```
>>> T_res=ttest_ind(parcelaA, parcelaB, equal_var=True, alternative='greater')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=-3.1370, pvalue=0.9985, df= 48 )

>>> T_res=ttest_ind(parcelaA, parcelaB, equal_var=True, alternative='less')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=-3.1370, pvalue=0.001457, df= 48 )

>>> T_res=ttest_rel(parcelaA, parcelaB, alternative='less')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=-2.6753, pvalue=0.006617, df= 24 )

>>> T_res=ttest_rel(parcelaA, parcelaB, alternative='greater')
>>> print('TtestResult(statistic=%.4f, pvalue=%.4f,' % T_res, 'df=',T_res.df,')')
  TtestResult(statistic=-2.6753, pvalue=0.9934, df= 24 )
```

Soluções de alguns Exercícios – Parte C

3.1. a) Como $n = 100$ é "grande", pelo Teorema Limite Central, verifica-se para a média amostral, $\overline{X} = \sum_{i=1}^{100} X_i/n$, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Longrightarrow \frac{\overline{X}-31}{0.24} \sim N(0,1)$. Logo, $P[\overline{X} < 30] \approx \Phi(\frac{30-31}{0.24}) = \Phi(-4.17) = 1-\Phi(4.17) = 1-1 = 0$

b) Seja $S_{100} = \sum_{i=1}^{100} X_i$ o peso total da amostra aleatória. Pelo Teorema Limite Central tem-se

$$\frac{S_{100}-100\mu}{\sigma\sqrt{100}} \sim N(0,1) \Longrightarrow \frac{S_{100}-3100}{24} \sim N(0,1).$$

Então, $P[S_{100} > 3150] = 1-P[S_{100} \leq 3150] \approx 1-\Phi(\frac{3150-3100}{24}) = 1-\Phi(2.08) = 1-0.9812 = 0.0188$.

3.2. a) $\bar{x} = 56.2$ cm e $s = 4$ cm

b) 0.42074

c) 4/5=0.8 cm.

3.3. 1082

3.4. $\approx 1-\Phi(0.4\sqrt{12n})$, para n elevado

3.5. ≈ 0.01267

3.7. Resolução detalhada na coletânea de exames

3.8. a) $]77.27\,;79.33[$

b) 1.03

c) 2654

3.9. 18.3 e $n = 12$

3.10. a) $\bar{x} = 0.959$ kg e $s^2 = 0.0116$ kg^2

c) $]0.00886,\ 0.0155[$

3.11. a) 99%

b) 0.185

3.12. a) i) 27 ii) 0.95

b) $]14.76\,;64.77[$

3.13. a) 25.85

b) 95%

c) O IC não inclui o valor 28 m. Logo, com 95% de confiança, pode-se afirmar que o silvicultor não tem razão na afirmação.

d) i) $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ ii) IC para variância a 95% de confiança $=]5.9; 11.7[$

3.18. Resolução detalhada na coletânea de exames

3.19. a) Unidade estatística: planta de tabaco; variável: nº de folhas por planta

b) $\bar{x} = 2967/150 = 19.78$ e $\tilde{x} = 20$

3.21. Resolução detalhada na coletânea de exames

3.25. a) Verdadeira

b) Falsa

3.28. Resolução detalhada na coletânea de exames

3.32. Resolução detalhada na coletânea de exames

3.33. Resolução detalhada na coletânea de exames

3.35. Soluções com alguns detalhes na coletânea de exames

3.36. Há um total de $N = 141$ pintos na segunda geração, sendo dado no enunciado o número de pintos observados com cada uma das três cores de penas. São igualmente dadas no enunciado as probabilidades associadas a cada cor de penas, a serem verdade os pressupostos genéticos referidos. Assim, o número esperado de pintos de penas brancas seria $E_b = N \times \pi_b = 141 \times \frac{1}{4} = 35.25$, o número esperado de pintos de penas pretas igualmente $E_p = N \times \pi_p = 141 \times \frac{1}{4} = 35.25$ e o número esperado de pintos de penas azuis $E_a = N \times \pi_a = 141 \times \frac{2}{4} = 70.50$. Sintetizando num quadro:

	Brancas	Pretas	Azuis
Observados (O_i)	36	32	73
Esperados (E_i)	35.25	35.25	70.50

Os valores esperados satisfazem claramente o critério de Cochran, pelo que pode admitir-se que, a serem corretas as probabilidades acima referidas, a estatística de Pearson tem distribuição χ^2 com $k - 1 = 3 - 1 = 2$ graus de liberdade. O valor calculado da estatística de Pearson é:

$$X^2_{calc} = \frac{(36-35.25)^2}{35.25} + \frac{(32-35.25)^2}{35.25} + \frac{(73-70.50)^2}{70.50} = 0.4042553 \ .$$

Sabemos que a região crítica do teste χ^2 associado à estatística de Pearson é unilateral direita. Neste caso concreto (e ao nível de significância $\alpha = 0.05$), será a região à direita do valor $\chi^2_{0.05(2)} = 5.991465$. Assim, o valor calculado da estatística não pertence à região crítica, pelo que não se rejeita H_0, ou seja, não se rejeita a validade da teoria genética de recessividade e dominância indicada no enunciado.

Cálculos em Python:

```
import numpy as np
from scipy.stats import chisquare
Oi=[36,32,73]
pi=[0.25,0.25,0.5]
Ei=np.array(pi)*sum(Oi)
res=chisquare(f_obs=Oi, f_exp=Ei)
print('ChiSquare Result(statistic=%.4f, pvalue=%.4f)' %res)
```

```
ChiSquare Result(statistic=0.4043, pvalue=0.8170)
```

3.38. Se o número de cachos por pé segue uma distribuição Poisson com parâmetro $\lambda = 4$, então a probabilidade de existirem x cachos por pé é dada por:

$$P[X = x] \quad = \quad \mathrm{e}^{-\lambda}\,\frac{\lambda^x}{x!} \quad = \quad \mathrm{e}^{-4}\,\frac{4^x}{x!} \quad (x \in \mathbb{N}_0).$$

Nesse caso, a probabilidade de cada um dos valores de 0 a 8, é dada (tendo em conta que, por convenção, $0! = 1$ e $4^0 = 1$) por:

$$\pi_0 = \mathrm{e}^{-4} \approx 0.01832 \qquad \pi_1 = 4\,\mathrm{e}^{-4} \approx 0.07326 \qquad \pi_2 = 8\,\mathrm{e}^{-4} \approx 0.14653$$

$$\pi_3 = \frac{32}{3}\,\mathrm{e}^{-4} \approx 0.19537 \qquad \pi_4 = \frac{32}{3}\,\mathrm{e}^{-4} \approx 0.19537 \qquad \pi_5 = \frac{128}{15}\,\mathrm{e}^{-4} \approx 0.15629$$

$$\pi_6 = \frac{256}{45}\,\mathrm{e}^{-4} \approx 0.10420 \qquad \pi_7 = \frac{1024}{315}\,\mathrm{e}^{-4} \approx 0.05954 \qquad \pi_8 = \frac{512}{315}\,\mathrm{e}^{-4} \approx 0.02977$$

Em Python, estes valores podem ser obtidos da seguinte forma:

```
import numpy as np
from scipy import stats
prob=stats.poisson.pmf(range(0,9),4)
print("Probabilidades:", *(f"{p:.5f}" for p in prob))
```

```
Probabilidades: 0.01832 0.07326 0.14653 0.19537 0.19537 0.15629
                0.10420 0.05954 0.02977
```

A probabilidade da última classe (número de cachos maior que 8), ou seja, $P[X > 8]$, obtém-se somando as probabilidades anteriores e subtraindo a 1, ou seja, $P[X > 8] = 1 - P[X \leq 8]$. Ou, em Python:

```
>>> print(round(1-stats.poisson.cdf(8,4), 5))
  0.02136
```

Assim, a probabilidade de ter oito ou mais observações é:

$$\pi_{>8} \;=\; 1-\sum_{i=0}^{8}\pi_i \;=\; 1-0.9786366 \approx 0.02136\ .$$

O cálculo dos valores esperados em cada categoria obtem-se multiplicando o número total de pés de videira observados ($N=200$) pelas respetivas probabilidades de categoria, obtendo-se:

$$E_0 = 3.663 \quad E_1 = 14.653 \quad E_2 = 29.305 \quad E_3 = 39.073 \quad E_4 = 39.073$$

$$E_5 = 31.259 \quad E_6 = 20.839 \quad E_7 = 11.908 \quad E_8 = 5.954 \quad E_{>8} = 4.273\ .$$

Como apenas dois dos dez valores esperados são inferiores a 5 e nenhum é inferior a 1, considera-se que a situação é adequada para aproximar a distribuição da estatística de Pearson por uma χ^2. Os graus de liberdade desta χ^2 são 9 (porque há $k=10$ classes e uma restrição, resultante de fixar o número total de observações em $N=200$). Vamos proceder ao cálculo da estatística de Pearson, calculando as suas dez parcelas:

Cachos	0	1	2	3	4	5	6	7	8	>8
O_i	2	20	29	47	54	29	14	4	1	0
E_i	3.663	14.653	29.305	39.073	39.073	31.259	20.839	11.908	5.954	4.273
$\frac{(O_i-E_i)^2}{E_i}$	0.755	1.952	0.003	1.608	5.702	0.163	2.245	5.252	4.122	4.273

A estatística de Pearson calculada vem: $X^2_{calc} = 26.07420$. Como numa distribuição χ^2_9 o valor que deixa à direita uma região de probabilidade $\alpha = 0.05$ é $\chi^2_{0.05(9)} = 16.91898$, *rejeita-se a hipótese nula* e considera-se que o número de cachos por pé *não* segue uma distribuição $P(4)$.

Nota: Esta rejeição não invalida que a distribuição possa ser uma Poisson *de parâmetro diferente*.

Em Python:

```
import numpy as np
from scipy import stats
Oi=[2, 20, 29, 47, 54, 29, 14, 4, 1, 0]
pi=list(stats.poisson.pmf(range(0,9),4))+[1-stats.poisson.cdf(8,4)]
Ei=np.array(pi)*sum(Oi)
res=stats.chisquare(f_obs=Oi, f_exp=Ei)
print('ChiSquare Result(statistic=%.4f, pvalue=%.4f)' %res)
```

```
  ChiSquare Result(statistic=26.0742, pvalue=0.0020)
```

O baixo valor do *p-value* (0.002) significa que, mesmo num teste com nível de significância $\alpha = 0.01$ ou $\alpha = 0.005$ optar-se-ia por rejeitar H_0, em favor da hipótese de outra distribuição.

3.39. Uma distribuição Binomial surge normalmente quando se conta o número de êxitos em *n provas de Bernoulli* (provas independentes com resultado dicotómico: êxito/fracasso), havendo probabilidade p de êxito em cada uma das n provas. Se $X \frown \mathscr{B}(n,p)$, X toma valores inteiros entre 0 e n, sendo a probabilidade do valor x dada por:

$$P[X=x] \quad = \quad \binom{n}{x} p^x (1-p)^{n-x}$$

Neste caso, a hipótese nula do teste é $X \frown \mathscr{B}(4, 0.455)$, sendo a hipótese alternativa $X \not\frown \mathscr{B}(4, 0.455)$. As probabilidades ao abrigo de H_0 podem-se obter com a função `binom.pmf` da biblioteca `scipy`, módulo `stats`:

```
import numpy as np
from scipy import stats
prob=stats.binom.pmf(range(0,5),4,0.455)
print("Probabilidades:", *(f"{p:.5f}" for p in prob))
```

```
  Probabilidades: 0.08822 0.29462 0.36895 0.20535 0.04286
```

O *número esperado de observações em cada classe* será o número total de observações ($N = 100$) vezes as probabilidades de cada classe:

$E_0 = 8.822$ $E_1 = 29.462$ $E_2 = 36.895$ $E_3 = 20.535$ $E_4 = 4.286$

Verificam-se as condições de Cochran para admitir a distribuição assintótica da estatística de Pearson. Tendo em conta que há $k = 5$ classes, teremos uma distribuição assintótica $\chi^2_{(4)}$. O cálculo da estatística de Pearson dá:

$$X^2_{calc} \quad = \quad \sum_{i=0}^{4} \frac{(O_i - E_i)^2}{E_i} \quad = \quad \frac{(15-8.822)^2}{8.822} + \frac{(20-29.462)^2}{29.462} + \frac{(40-36.895)^2}{36.895} + \frac{(18-20.535)^2}{20.535} + \frac{(7-4.286)^2}{4.286} \quad = \quad 9.657347\ .$$

O valor que define uma região crítica unilateral direita, ao nível $\alpha = 0.05$ é $\chi^2_{0.05(4)} = 9.487729$, que se obtém em Python com `stats.chi2.ppf(0.95,4)`. Logo, estamos perante um valor que é significativo, isto é, *rejeita-se a hipótese nula de que a distribuição seja Binomial com probabilidade de sucesso de* 0.455.

É possível obter o valor calculado da estatística de teste e o *p-value* do teste com:

```
import numpy as np
from scipy import stats
Oi=[15, 20, 40, 18, 7]
pi=list(stats.binom.pmf(range(0,5),4,0.455))
Ei=np.array(pi)*sum(Oi)
res=stats.chisquare(f_obs=Oi, f_exp=Ei)
print('ChiSquare Result(statistic=%.4f, pvalue=%.4f)' %res)
```

```
  ChiSquare Result(statistic=9.6573, pvalue=0.0466)
```

Note-se que o *p-value* está muito próximo de 0.05, pelo que a rejeição da hipótese nula com este nível de significância não é clara.

3.40. Pede-se para testar se (hipótese H_0) as contagens seguem uma distribuição de Poisson com parâmetro $\lambda = 2$. Caso esta hipótese seja verdadeira:

$$\begin{aligned}
P[X=0] &= \mathrm{e}^{-2}\frac{2^0}{0!} = \mathrm{e}^{-2} \approx 0.1353 \qquad (\text{convenciona-se que } 0! = 1)\\
P[X=1] &= \mathrm{e}^{-2}\frac{2^1}{1!} = 2\mathrm{e}^{-2} \approx 0.2707\\
P[X=2] &= \mathrm{e}^{-2}\frac{2^2}{2!} = 2\mathrm{e}^{-2} \approx 0.2707\\
P[X=3] &= \mathrm{e}^{-2}\frac{2^3}{3!} = \frac{4}{3}\mathrm{e}^{-2} \approx 0.1804\\
P[X=4] &= \mathrm{e}^{-2}\frac{2^4}{4!} = \frac{2}{3}\mathrm{e}^{-2} \approx 0.0902\\
P[X\geq 5] &= 1-P[X\leq 4] = 1-\mathrm{e}^{-2}\left[1+2+2+\frac{4}{3}+\frac{2}{3}\right] \approx 0.0527
\end{aligned}$$

Em Python estas probabilidades obtêm-se através de

```
import numpy as np
from scipy import stats
pi=list(stats.poisson.pmf(range(0,5),2))+[1-stats.poisson.cdf(4,2)]
print("Probabilidades:", *(f"{p:.5f}" for p in pi))
```

```
  Probabilidades: 0.13534 0.27067 0.27067 0.18045 0.09022 0.05265
```

Note-se que a última classe foi considerada como uma classe de 5 *ou mais* valores. De facto, *a soma das probabilidades de todas as classes deve dar* 1. Não se observaram mais do que 5 poros por casca, mas tais valores não são impossíveis. Assim, considera-se que a última classe corresponde ao acontecimento "cinco ou mais poros".

Logo, os valores esperados são obtidos multiplicando estas probabilidades pelo número total de observações ($N = 320$). Tem-se:

Poros	0	1	2	3	4	≥ 5
E_i	43.307	86.615	86.615	57.743	28.872	16.849
O_i	18	74	139	70	17	2
$\frac{(O_i-E_i)^2}{E_i}$	14.789	1.837	31.683	2.602	4.881	13.086

A estatística calculada é $X^2_{calc} = 68.8787$. Este valor observado deve ser comparado com uma distribuição χ^2 com 5 graus de liberdade. (Todos os valores esperados são superiores a 5, pelo que não é necessário agrupar classes).

O valor fronteira duma região crítica ao nível $\alpha = 0.05$ é $\chi^2_{0.05((5))} = 11.07$. Assim, *rejeita-se a hipótese nula de que haja uma distribuição de Poisson com* $\lambda = 2$. Esta rejeição é muito enfática, como se pode confirmar calculando o valor de prova (*p-value*) associado ao valor calculado da estatística, `1-stats.chi2.cdf(68.8787,5)` que é da ordem de 1.75×10^{-13}. O teste em Python pode ser realizado com:

```
import numpy as np
from scipy import stats
Oi=[18, 74, 139, 70, 17, 2]
pi=list(stats.poisson.pmf(range(0,5),2))+[1-stats.poisson.cdf(4,2)]
Ei=np.array(pi)*sum(Oi)
res=stats.chisquare(f_obs=Oi, f_exp=Ei)
print('ChiSquare Result(statistic=%.4f, pvalue=%10.3E)' %res)

  ChiSquare Result(statistic=68.8787, pvalue= 1.754E-13)
```

Da análise das parcelas individuais da estatística de Pearson observa-se que as que mais contribuem para a rejeição de H_0, isto é, para o elevado valor da estatística, são as parcelas associadas às contagens 0, 2 e 5. No entanto, as parcelas dos valores extremos resultam de haver bastante menos observações do que seria de esperar à luz da hipótese de distribuição Poisson ($O_i \ll E_i$), enquanto que o elevado valor da parcela para a contagem 2 resulta de um muito maior número de observações do que seria de esperar ($O_i \gg E_i$). Assim, a distribuição observada é muito mais *concentrada* (de menor dispersão) do que a Poisson.

3.41. No enunciado são indicadas as quatro probabilidades associadas a cada combinação de cor e tipo de superfície, resultantes dos pressupostos genéticos que foram admitidos. Indicando por π_{ij} a probabilidade de se ter a cor i (onde $i = 1$ corresponde a amarelo e $i = 2$ a verde) e uma superfície de tipo j (onde $j = 1$ indica lisa e $j = 2$ rugosa), temos $\pi_{11} = 9/16$, $\pi_{12} = 3/16$, $\pi_{21} = 3/16$ e $\pi_{22} = 1/16$.

a) Uma vez que existem ao todo $N = 994$ observações, os valores esperados são, respetivamente, $E_{11} = N \times \pi_{11} = 994 \times \frac{9}{16} = 559.125$, $E_{12} = N \times \pi_{12} = 994 \times \frac{3}{16} = 186.375$, $E_{21} = N \times \pi_{21} = 994 \times \frac{3}{16} = 186.375$ e $E_{22} = N \times \pi_{22} = 994 \times \frac{1}{16} = 62.125$. Resumindo numa única tabela os valores observados e (entre parênteses) esperados ao abrigo dos pressupostos genéticos referidos no enunciado, temos:

	Superfície	
Côr	Lisa	Rugosa
Amarelas	556 (559.125)	184 (186.375)
Verdes	193 (186.375)	61 (62.125)

Todos os valores esperados são grandes, pelo que não há problemas em admitir que a estatística de Pearson tem distribuição χ^2, neste caso com $ab - 1 = 3$ graus de liberdade. Assim, tem-se:

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(556-559.125)^2}{559.125} + \frac{(184-186.375)^2}{186.375} + \frac{(193-186.375)^2}{186.375} + \frac{(61-62.125)^2}{62.125} = 0.3036\,.$$

A região crítica (ao nível $\alpha = 0.05$ pedido no enunciado) tem fronteira $\chi^2_{0.05(3)} = 7.8147$. Logo, o valor calculado da estatística não pertence à região crítica, pelo que não se rejeita a hipótese nula, isto é, consideram-se admissíveis os pressupostos genéticos de dominância/recessividade e segregação independente das caraterísticas referidas.

Para efetuar estes cálculos em Python, pode-se proceder como no caso unidimensional, e criar um vetor de valores observados e outro de probabilidades sob H_0, tendo apenas o cuidado de especificar a mesma ordem (por linhas ou por colunas da tabela), quer para os valores observados, quer para as probabilidades. Assim, por exemplo:

```
import numpy as np
from scipy.stats import chisquare
Oij=[556,184,193,61]
pij=np.array([9,3,3,1])/16
Eij=pij*sum(Oij)
res=chisquare(f_obs=Oij, f_exp=Eij)
print('ChiSquare Result(statistic=%.4f, pvalue=%.4f)' %res)

  ChiSquare Result(statistic=0.3036, pvalue=0.9594)
```

b) Agora existem ao todo $N^* = 30 \times N = 30 \times 994 = 29820$ observações (onde o asterisco indica a nova situação desta alínea). Todos os valores esperados são assim 30 vezes maiores do que eram antes, ou seja, $E^*_{ij} = N^* \times \pi_{ij} = 30 \times N \times \pi_{ij}$, para qualquer i e j. Mas se as proporções observadas em cada célula se mantiveram iguais, é porque os valores observados em cada célula também são 30 vezes maiores do que os observados antes. Assim, $O^*_{ij} = 30 \times O_{ij}$, para todo o i e j.
O novo valor calculado da estatística de teste é também 30 vezes maior, já que:

$$X^*_{calc} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O^*_{ij} - E^*_{ij})^2}{E^*_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(30O_{ij} - 30E_{ij})^2}{30E_{ij}} = 30\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 30 \times 0.3036 = 9.108.$$

A região crítica não se alterou, uma vez que os graus de liberdade associados à estatística do teste se mantêm iguais. Mas o valor calculado da estatística alterou-se (é 30 vezes maior) e pertence agora à região crítica para $\alpha = 0.05$, pelo que se rejeita a hipótese nula, isto é, não se consideram admissíveis os pressupostos genéticos de dominância/recessividade e segregação independente das caraterísticas referidas.

Nota: Todos os valores esperados são maiores do que na alínea anterior, pelo que não há problemas em admitir que a estatística de Pearson tem distribuição χ^2_3.

3.42. Tal como no Exercício anterior, as probabilidades resultantes da teoria genética, associadas a cada combinação de comprimento e cor do pêlo são completamente especificados no enunciado. Os valores esperados para cada uma dessas células resultam assim do produto $E_{ij} = N \times \pi_{ij}$ onde $N = 482$ é o número total de cobaias observadas na segunda geração, $i = 1,2$ indica o comprimento do pêlo (pela ordem de linha da tabela do enunciado) e $j = 1,2,3$ indica a cor do pêlo (pela ordem de coluna da tabela do enunciado), sendo π_{ij} a probabilidade da combinação de comprimento e cor do pêlo referidas. Por exemplo, o número esperado de cobaias de pêlo longo e branco será $E_{23} = 482 \times \frac{1}{16} = 30.125$. Eis a tabela com os valores observados e (entre parênteses) os

correspondentes valores esperados ao abrigo da teoria genética (que verificam as condições de Cochran):

	Côr		
Pelo	Creme	Amarelo	Branco
Curto	178 (180.750)	93 (90.375)	89 (90.375)
Longo	62 (60.250)	29 (30.125)	31 (30.125)

A estatística de Pearson tem assim o seguinte valor calculado:

$$X^2_{calc} = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(178-180.750)^2}{180.750} + ... + \frac{(31-30.125)^2}{30.125} = 0.2573 \ .$$

O enunciado não especifica qualquer nível de significância, mas optando por $\alpha = 0.05$, rejeita-se H_0 se $X^2_{calc} > \chi^2_{0.05(5)} = 11.0705$. Uma vez que esta desigualdade não se verifica, não se rejeita H_0, sendo admissível que haja segregação independente da cor e comprimento do pêlo das cobaias. Em Python:

```
import numpy as np
from scipy.stats import chisquare
Oij=[178,93,89,62,29,31]
pij=np.array([6,3,3,2,1,1])/16
Eij=pij*sum(Oij)
res=chisquare(f_obs=Oij, f_exp=Eij)
print('ChiSquare Result(statistic=%.4f, pvalue=%.4f)' %res)
```

```
  ChiSquare Result(statistic=0.2573, pvalue=0.9984)
```

3.43. a) O objetivo é o de saber se se pode admitir que a distribuição pelas três categorias de resultados (morte, calo e enraizamento bem sucedido) são idênticas para os quatro tratamentos utilizados na experiência. Dito de outra maneira, queremos saber se a probabilidade de morte é igual, qualquer que seja o nível do fator tratamento (em cujo caso, pode falar-se apenas em π_{Morte}) e, de forma análoga, se há uma única probabilidade de criar calo (π_{Calo}) qualquer que seja o tratamento, e uma única probabilidade de enraizamento (π_{Enraiz}) qualquer que seja o nível do fator Tratamento. Uma forma de explicitar melhor esta hipótese será considerar que $\pi_{j|i}$ indica a probabilidade de, no tratamento i ($i = 1,2,3,4$) o resultado ser j ($j = 1,2,3$, associados respetivamente a $Morte, Calo, Enraizamento$), e escrever:

$$H_0 \ : \ \begin{cases} \pi_{Morte|1} = \pi_{Morte|2} = \pi_{Morte|3} = \pi_{Morte|4} & [= \pi_{Morte} = \pi_{.1}] \\ \pi_{Calo|1} = \pi_{Calo|2} = \pi_{Calo|3} = \pi_{Calo|4} & [= \pi_{Calo} = \pi_{.2}] \\ \pi_{Enraiz|1} = \pi_{Enraiz|2} = \pi_{Enraiz|3} = \pi_{Enraiz|4} & [= \pi_{Enraiz} = \pi_{.3}] \end{cases}$$

A hipótese alternativa H_1 será que pelo menos uma das igualdades acima referidas não é verdadeira. A tabela de contingências tem os totais de linha (número de observações para cada um dos quatro tratamentos) fixado à partida pelo experimentador.

Estamos assim perante um *teste de homogeneidade*. A haver uma distribuição comum pelos três tipos de resultados, as probabilidades associadas a cada possível resultado podem ser estimadas a partir das frequências relativas marginais:

$$\hat{\pi}_{Morte} = \hat{\pi}_{.1} = \frac{N_{.1}}{N} = \frac{121}{240} = 0.50417$$
$$\hat{\pi}_{Calo} = \hat{\pi}_{.2} = \frac{N_{.2}}{N} = \frac{83}{240} = 0.34583$$
$$\hat{\pi}_{Enraiz} = \hat{\pi}_{.3} = \frac{N_{.3}}{N} = \frac{36}{240} = 0.15000$$

A ser verdade a hipótese de distribuição homogénea nos quatro tratamentos, o valor esperado para cada categoria é dado por: $\hat{E}_{ij} = N_{i.} \times \hat{\pi}_{.j}$. Como os totais de cada linha são todos iguais (60), os valores esperados estimados de cada resultado também vêm iguais nos quatro tratamentos ($i = 1,2,3,4$):

$$\hat{E}_{i1} = 60 \times 0.5041667 = 30.25 \qquad \hat{E}_{i2} = 60 \times 0.3458333 = 20.75 \qquad \hat{E}_{i3} = 60 \times 0.15 = 9\ .$$

Eis a tabela de valores observados e esperados estimados (estes últimos entre parênteses):

	Resultado			
Tratamento	Morte	Com calo	Enraizamento	Total
Sem incisão/sem boro	26 (30.25)	18 (20.75)	16 (9)	60
Com incisão/sem boro	32 (30.25)	22 (20.75)	6 (9)	60
Sem incisão/com boro	24 (30.25)	24 (20.75)	12 (9)	60
Com incisão/com boro	39 (30.25)	19 (20.75)	2 (9)	60
Total	121	83	36	240

O cálculo do valor da estatística de Pearson produz:

$$\begin{aligned}\sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} &= \frac{(26-30.25)^2}{30.25} + \frac{(18-20.75)^2}{20.75} + \frac{(16-9)^2}{9} + ... + \frac{(2-9)^2}{9} \\ &= 0.5971074 + 0.3644578 + 5.444444 + 0.1012397 + ... + 5.444444 \\ &= 18.50593\end{aligned}$$

Não havendo violação do critério de Cochran, o valor calculado da estatística (18.50593) pode ser comparado com a fronteira duma região crítica unilateral direita numa distribuição χ^2_6. Esse valor fronteira, para um nível de significância $\alpha = 0.05$, é $\chi^2_{0.05(6)} = 12.59159$. Em Python este valor obtém-se através do comando que pede o quantil 0.95 da distribuição $\chi^2_{(6)}$:

```
from scipy import stats
print(stats.chi2.ppf(0.95,6))
```

```
12.591587243743977
```

Uma vez que $\chi^2_{calc} > \chi^2_{0.05(6)}$, rejeita-se H_0, ou seja, rejeita-se (ao nível de significância 0.05) a hipótese de haver homogeneidade na distribuição dos resultados do enraizamento, para os quatro tratamentos.

Estes cálculos podem igualmente ser feitos em Python.

```
>>> import numpy as np
>>> import pandas as pd
>>> from scipy.stats import chi2_contingency
>>> # Criar a matriz de frequências observadas
>>> Oij = np.array([[26, 18, 16],
>>>                 [32, 22,  6],
>>>                 [24, 24, 12],
>>>                 [39, 19,  2]])
>>> nomes_colunas=['Morte', 'Calo', 'Enraizamento']
>>> nomes_linhas=['sI/sB', 'cI/sB', 'sI/cB','cI/cB']
>>> # Executar o teste do qui-quadrado
>>> X2calc, pval, gl, Eij = chi2_contingency(Oij)
>>> # Mostrar os resultados
>>> print('Frequências Observadas:\n', pd.DataFrame(Oij,
          columns=nomes_colunas, index=nomes_linhas))
  Frequências Observadas:
         Morte  Calo  Enraizamento
  sI/sB     26    18            16
  cI/sB     32    22             6
  sI/cB     24    24            12
  cI/cB     39    19             2

>>> print('Frequências esperadas:\n', pd.DataFrame(Eij,
           columns=nomes_colunas, index=nomes_linhas))
  Frequências esperadas:
         Morte   Calo  Enraizamento
  sI/sB  30.25  20.75           9.0
  cI/sB  30.25  20.75           9.0
  sI/cB  30.25  20.75           9.0
  cI/cB  30.25  20.75           9.0

>>> ET=pd.DataFrame((Oij-Eij)**2/Eij, columns=nomes_colunas, index=nomes_linhas)
>>> print('Parcelas da Estatistica de Teste:\n', ET)
  Parcelas da Estatistica de Teste:
              Morte      Calo  Enraizamento
  sI/sB  0.597107  0.364458      5.444444
  cI/sB  0.101240  0.075301      1.000000
  sI/cB  1.291322  0.509036      1.000000
  cI/cB  2.530992  0.147590      5.444444

>>> print('X2 calc: ', X2calc)
  X2 calc:  18.50593558808236

>>> print('Graus de liberdade:', gl)
  Graus de liberdade: 6

>>> print('p-value: ', pval)
  p-value:  0.005084732324308631
```

O comando `chi2_contingency` parte do pressuposto que se pretende efetuar ou um teste de homogeneidade ou um teste de independência, para os quais os procedimentos de cálculo são idênticos. Repare-se que os graus de liberdade indicados são iguais a $(a-1)(b-1)$, onde a indica o número de linhas da matriz e b o seu número de colunas. Este valor corresponde aos graus de liberdade nos dois testes referidos.

O valor de prova 0.005085 indica que para qualquer nível de significância maior do que esse valor, a conclusão do teste seria a rejeição da hipótese de homogeneidade.

b) A fim de perceber as causas duma tal rejeição, podemos analisar as parcelas da soma que gera o valor calculado da estatística. As três parcelas de maior valor são a parcela da linha 1, coluna 3 (associada ao enraizamento, no tratamento sem incisão e sem boro), de valor 5.44444; a parcela da linha 4, coluna 3 (enraizamento no tratamento com incisão e com boro), igualmente de valor 5.44444; e a parcela da linha 4, coluna 1 (morte no tratamento com incisão e com boro), de valor 2.530992. Só por si, a soma destas três parcelas já excede a fronteira da região crítica, sendo assim estas combinações de resultados e tratamentos as mais responsáveis pela conclusão de rejeição de H_0. Nos três casos há discrepâncias importantes entre valores esperados e valores observados. No entanto, essas discrepâncias são de sinal diferente. Os enraizamentos observados no tratamento sem incisão, nem boro, são em número muito maior (16) do que o esperado (9). Pelo contrário, os enraizamentos observados no tratamento com incisão e com boro são muito menos (2) do que o esperado (9). Para este último tratamento, as mortes observadas são bastante mais numerosas (39) do que o esperado (30.25). Em suma, pode afirmar-se que a falta de homogeneidade está sobretudo associada aos dois tratamentos extremos (sem intervenção e com os dois tipos de intervenção), sendo que o enraizamento é mais bem sucedido quando não há qualquer tipo de intervenção nas estacas.

3.44. Um teste de independência corresponde, neste caso, à pior das situações possíveis: a classificação no terreno e a classificação com base nas imagens de satélite não terem qualquer correspondência. É de desejar que haja uma claríssima rejeição desta hipótese, dado o contexto do problema. A hipótese de independência gera, como valores esperados estimados em cada célula, os valores $\hat{E}_{ij} = N\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} = \frac{N_{i.}\times N_{.j}}{N}$. Tem-se $N_{1.} = 35$, $N_{2.} = 40$, $N_{3.} = 25$, $N_{.1} = 31$, $N_{.2} = 42$, $N_{.3} = 27$ e $N = 100$. Logo,

$$\begin{array}{lll} \hat{E}_{11} = \frac{35\times 31}{100} = 10.85 & \hat{E}_{12} = \frac{35\times 42}{100} = 14.7 & \hat{E}_{13} = \frac{35\times 27}{100} = 9.45 \\ \hat{E}_{21} = \frac{40\times 31}{100} = 12.4 & \hat{E}_{22} = \frac{40\times 42}{100} = 16.8 & \hat{E}_{23} = \frac{40\times 27}{100} = 10.8 \\ \hat{E}_{31} = \frac{25\times 31}{100} = 7.75 & \hat{E}_{32} = \frac{25\times 42}{100} = 10.5 & \hat{E}_{33} = \frac{25\times 27}{100} = 6.75 \end{array}$$

A estatística de Pearson toma assim o valor:

$$\begin{aligned} X^2_{calc} &= \frac{(16-10.85)^2}{10.85} + \frac{(15-14.7)^2}{14.7} + \frac{(4-9.45)^2}{9.45} + \frac{(15-12.4)^2}{12.4} + \frac{(22-16.8)^2}{16.8} + \frac{(3-10.8)^2}{10.8} \\ &\qquad + \frac{(0-7.75)^2}{7.75} + \frac{(5-10.5)^2}{10.5} + \frac{(20-6.75)^2}{6.75} \\ &= 2.4445 + 0.0061 + 3.1431 + 0.5452 + 1.6095 + 5.6333 + 7.7500 + 2.8810 + 26.0093 \\ &= 50.02194 \end{aligned}$$

Nos testes de independência, e como resultado da estimação das distribuições marginais, há $(a-1)(b-1)$ graus de liberdade, onde a designa o número de linhas e b o número de colunas. No nosso caso, são 4 graus de liberdade. Usando $\alpha = 0.05$, tem-se $\chi^2_{0.05(4)} = 9.4877$, pelo que a rejeição da hipótese de independência é clara.

Da análise das parcelas da estatística X^2_{calc} é possível constatar que é sobretudo a última parcela (correspondente à classe $(3,3)$) que é responsável pelo grande valor da estatística, e em menor medida, também as parcelas correspondentes às classes $(3,1)$ e $(2,3)$. Inspecionando de novo a tabela dos valores observados constata-se que, como seria de esperar, há uma grande concentração de parcelas no canto inferior direito da tabela, sendo que a maioria das parcelas classificadas como sendo de regadio ao abrigo duma técnica, também o são ao abrigo da outra técnica. No entanto, o mesmo não se pode dizer das parcelas de sequeiro/não cultivadas. Para estas duas classes, a situação em pouco se distingue da que seria de esperar ao abrigo da hipótese de independência. Admitindo que a classificação feita por inspeção direta no terreno é a verdadeira, pode afirmar-se que a classificação por satélite consegue separar parcelas de regadio, mas não consegue distinguir entre culturas de sequeiro e parcelas não cultivadas.

O comando `chi2_contingency` também permite efetuar um teste de independência a partir duma matriz de contingências.

```
>>> import numpy as np
>>> import pandas as pd
>>> from scipy.stats import chi2_contingency
>>> Oij = np.array([[16, 15,  4],
                    [15, 22,  3],
                    [ 0,  5, 20]])

>>> nomes=['NCult', 'Seq', 'Reg']

>>> X2calc, pval, gl, Eij = chi2_contingency(Oij)

>>> print('Frequências Observadas:\n', pd.DataFrame(Oij,
           columns=nomes, index=nomes))
  Frequências Observadas:
          NCult  Seq  Reg
  NCult     16   15    4
  Seq       15   22    3
  Reg        0    5   20

>>> print('Frequências esperadas:\n',  pd.DataFrame(Eij,
           columns=nomes, index=nomes))
  Frequências esperadas:
          NCult   Seq    Reg
  NCult  10.85  14.7   9.45
  Seq    12.40  16.8  10.80
  Reg     7.75  10.5   6.75

>>> ET=pd.DataFrame((Oij-Eij)**2/Eij, columns=nomes, index=nomes)
>>> print('Parcelas da Estatistica de Teste:\n', ET)
  Parcelas da Estatistica de Teste:
             NCult       Seq        Reg
  NCult  2.444470  0.006122   3.143122
  Seq    0.545161  1.609524   5.633333
  Reg    7.750000  2.880952  26.009259

>>> print('X2 calc: ', X2calc)
  X2 calc: 50.0219442615756
```

```
>>> print('Graus de liberdade:', gl)
  Graus de liberdade: 4

>>> print('p-value: ', pval)
  p-value:  3.572970251840667e-10
```

R3.2. 0.7127

R3.3. a) Estimativa do valor médio μ é $\bar{x} = 23.88\ ppm$ e de σ^2 é $s^2 = 1.9441\ ppm^2$

b) O grau de confiança é 99%.

R3.4. a) $\bar{x} = 8.179$

c) $n \geq 16$

R3.7. a)

x_i	Freq. Abs. n_i	Freq. Rel. f_i	Freq. Rel Acum. F_i
0	9	0.09	0.09
1	26	0.26	0.35
2	34	0.34	0.69
3	22	0.22	0.91
4	8	0.08	0.99
5	1	0.01	1.00

b) $\bar{x} = 197/100$ e $\tilde{x} = 2$

c) i) $\hat{p} = 197/500 = 0.394$ ii) 0.0428

R3.8. IC para p $]0.7417; 0.8423[$

Como o IC não contém 0.9, a afirmação do vendedor não é correta.

R3.9. a) 29.01%

b) $]0.2765; 0.3037[$

R3.10. Resolução detalhada na coletânea de exames