Enunciados do projeto de PE 2024/2025

Enunciado do exercício 1

O ficheiro winequality-white-q5.csv contém um conjunto de dados baseado nos dados descritos em:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Neste estudo foram analisadas 4898 amostras de vinho verde branco, com o objetivo de estudar como é que as preferências dos peritos em provas de vinhos se relacionam com as caraterísticas físico-químicas dos vinhos. No ficheiro, as variáveis 1 a 11 referem-se às caraterísticas físico-químicas dos vinhos. A variável 12 representa a avaliação feita por especialistas em vinhos: qualidade do vinho entre 1 (má) e 5 (excelente).

Com recurso ao pacote `ggplot2`, produza um gráfico que represente, através de diagramas de caixa (*box plots*) paralelos, como é que a raiz quadrada da variável **citric.acid** se relaciona com a variável **quality**. No gráfico, deverá realçar as possíveis observações discordantes (*outliers*) e, de alguma forma, tentar contrariar a sua sobreposição.

Por simplicidade, mantenha todo o texto da figura em Inglês.

Submeta um ficheiro em formato PDF com uma única página A4, que inclua:

- O código em R, incluindo os comandos para leitura dos dados do ficheiro.
- O gráfico produzido.

Enunciado do exercício 2

O ficheiro wine_prod_EU.xlsx contém dados coletados pela Comissão Europeia sobre a produção anual de vinho desde 1997 até 2024 nos vários países produtores da União Europeia. Para cada ano, os dados incluem a quantidade de vinho em armazém antes da colheita (**Opening Stock** em 10^3 hL), a produção da colheita (**Production** em 10^3 hL) e a quantidade de vinho disponível após a colheita (**Availability** em 10^3 hL) em diferentes categorias de vinho.

Comece por eliminar todas as observações em que a variável **Category** está em falta e também aquelas em que a variável **Product Group** é igual a `Non-Vinified`.

Recorrendo ao pacote `ggplot2`, produza um único gráfico de barras que permita comparar a distribuição da variável **Production** para as diferentes categorias de vinho no ano de **2022** entre cada um dos seguintes países: **France, Italy, Spain, Germany, Portugal** (em Inglês no ficheiro), e o agrupamento formado pelos restantes países (**Others**).

Tenha em conta que o texto no ficheiro de dados se encontra em Inglês e, por simplicidade, mantenha todo o texto do gráfico nessa língua.

Submeta um ficheiro em formato PDF com uma única página A4, que inclua:

1. O código em `R`, que deve incluir os comandos para leitura e seleção dos dados do ficheiro.
2. O gráfico produzido.

Enunciado do exercício 3

O ficheiro clima.csv contém um exemplo de conjunto de dados de condições climáticas de dados horários de 2010 a 2014 num determinado local. Inclui métricas importantes, como poluição, orvalho, temperatura, velocidade do vento, neve, chuva. (Fonte: kaggle).

Após a leitura desse ficheiro no R, selecione todos os dados referentes ao mês de **março** de **2011**.

Com recurso ao pacote `ggplot2`, produza um único gráfico que ilustre a variação horária da variável **Orvalho** ao longo do referido mês. Adicionalmente, pretende-se representar nesse gráfico a variação da mediana diária dessa variável. Todo o texto do gráfico deverá estar em Português.

Submeta um ficheiro em formato PDF com uma única página A4, que inclua:

1. O código em R, que deve incluir os comandos para leitura e seleção dos dados do ficheiro.
2. O gráfico produzido.

Enunciado do exercício 4

A distribuição de Weibull é aplicada em muitos campos diferentes, como por exemplo, avaliação da fiabilidade de componentes eletrónicas ou de sistemas mecânicos, indústria farmacêutica ou ainda análise de sobrevivência em medicina.

Uma dada variável aleatória X, contínua, não negativa, tem distribuição de Weibull, de parâmetros (λ, k), se a sua função de densidade de probabilidade for dada por

$$f_X(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right], & x \geq 0 \\ 0, & x < 0 \end{cases}$$

onde $\lambda \in]0, +\infty[$ é o chamado parâmetro de escala e $k \in]0, +\infty[$ é o chamado parâmetro de forma. Pode mostrar-se que

$$E(X) = \int_0^{+\infty} x\, f_X(x)\, dx = \lambda \Gamma\left(1 + \frac{1}{k}\right),$$

onde $\Gamma(\cdot)$ é a chamada função *Gama*. Esta função é definida pelo integral impróprio $\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt$.

O valor $\Gamma(x)$ pode ser calculado explicitamente para x inteiro, obtendo-se $\Gamma(x) = (x-1)!$. Para x não inteiro é possível calcular valores aproximados usando integração numérica. A função `gamma` do R faz precisamente isso, o que permite, dados os valores dos parâmetros, λ e k, calcular aproximadamente $E(X)$. Por outro lado, também é possível calcular um valor aproximado de $E(X)$, utilizando o método conhecido como integração de Monte Carlo, que consiste em gerar um número muito elevado de observações de X e calcular a média aritmética respectiva.

Considere uma variável aleatória X que representa o tempo de vida de uma certa componente aeronáutica (em milhares de horas). Admite-se que X tem distribuição de Weibull com parâmetros $\lambda = 29$ e $k = 7$.

1. Calcule o valor esperado de X, usando a função `gamma`.
2. Fixando a semente em 1160, gere uma amostra de dimensão 5000 de X, e use-a para calcular um valor aproximado de $E(X)$.

Indique o valor absoluto da diferença entre os valores obtidos em 1. e 2., arredondado a 4 casas decimais.

Answer: 0.0574 ✔

Enunciado do exercício 5

Conta-se que o Grão-Duque da Toscana (séc. XVI) jogava assiduamente com um adversário o seguinte jogo:

- Em cada jogada, um dos jogadores (sejam A e B) lança três dados cúbicos perfeitos e soma as pontuações obtidas. Se a soma das pontuações for 9 (denotado por "soma 9"), o jogador A ganha a jogada e, se a soma das pontuações for 10 (denotado por "soma 10"), será o jogador B a ganhar; caso contrário, ninguém ganha.

O que intrigava o Grão-Duque era que, apesar de 9 e 10 pontos se poderem decompor cada um em 6 maneiras diferentes, o jogador A ganhava com menor frequência do que o jogador B.

Fixando a semente em $\mathbf{1091}$, simule $n = \mathbf{30000}$ jogadas desse jogo. Reporte o valor da diferença entre as frequências relativas com que a "soma 10" e a "soma 9" são obtidas nessas n jogadas, arredondada a 4 casas decimais.

Nota: caso use a função `sample` na simulação dos lançamentos dos dados, NÃO deverá especificar as probabilidades dos resultados possíveis. Por omissão, estes serão considerados equiprováveis.

Answer: 0.0097 ✔

Enunciado do exercício 6

Sejam $X_1, \ldots, X_n$ variáveis aleatórias contínuas independentes e identicamente distribuídas a $X \sim \textbf{uniforme}(0,1)$. Então $S_n = \sum_{i=1}^{n} X_i$ possui distribuição de Irwin-Hall e

$$P(S_n \leq x) = \frac{1}{n!} \sum_{k=0}^{\lfloor x \rfloor} (-1)^k \binom{n}{k} (x-k)^n,$$

para $0 \leq x \leq n$ e onde $\lfloor x \rfloor$ representa a parte inteira do real x.

1. Obtenha o valor exacto de $p_n = P(S_n \leq x)$, para $n = 11$ e $x = 5.3$.
2. Calcule dois valores aproximados de p_n recorrendo aos métodos seguintes:
   a. **Teorema do limite central** $(p_{n,TLC})$

      Recorra ao TLC, apesar de n ser inferior a 30.

   b. **Simulação** $(p_{n,sim})$

      i. Fixando a semente em 3985, gere $m = 110$ amostras de dimensão $n = 11$ da distribuição de X.

      ii. Calcule um valor simulado de S_n para cada uma das amostras geradas.

      iii. Obtenha a proporção de valores simulados de S_n que não excedem 5.3.
3. Determine o desvio absoluto entre o valor exacto calculado em 1., p_n, e o valor aproximado obtido em 2a., $p_{n,TLC}$.
4. Calcule o desvio absoluto entre p_n e $p_{n,sim}$.
5. Calcule o quociente entre os desvios calculados em 3. e 4. e apresente o resultado arredondado a 4 casas decimais.

Answer: 5.8722 ✔

Enunciado do exercício 7

Considere que a variável aleatória X representa o comprimento (em cm) dos ovos de uma dada espécie de pássaro e que pode ser modelada pela função densidade de probabilidade

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\lambda x},$$

para $x > 0$, onde α e λ são parâmetros com valores positivos desconhecidos e $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\,dt$ denota a função Gama.

Ao deduzir as estimativas de máxima verosimilhança de α e λ, $\hat{\alpha}$ e $\hat{\lambda}$, constatará que $\hat{\lambda}$ se pode escrever como função de $\hat{\alpha}$ mas que não existe uma solução explícita para $\hat{\alpha}$. No entanto, $\hat{\alpha}$ pode ser obtida numericamente por recurso à função `uniroot` do `R`. Use o intervalo $[46.6, 63.4]$ como intervalo inicial de pesquisa e não utilize qualquer outro argumento opcional dessa função.

Assuma que $(X_1, \ldots, X_n)$ é uma amostra aleatória de X e que a observação de $n = 12$ ovos dessa espécie de pássaro resultou em $\sum_{i=1}^n x_i = 61.17$ e $\sum_{i=1}^n \log x_i = 19.45$.

Determine a estimativa de máxima verosimilhança de $(\alpha - 1)/\lambda$, o *comprimento modal* dos ovos dessa espécie de pássaro, indicando-a arredondada a 2 casas decimais.

Answer: 5.02 ✔

Enunciado do exercício 8

Seja X uma variável aleatória que representa o desvio (em milímetros) entre o diâmetro observado e o diâmetro nominal de uma peça mecânica fabricada industrialmente. Considere que X tem distribuição normal com valor esperado μ desconhecido e desvio padrão conhecido e igual a $\sigma = 1$.

1. Usando o `R` e fixando a semente em 1308, gere $m = 1300$ amostras de dimensão $n = 10$ de uma distribuição normal com valor esperado igual a $\mu = 0.5$ e desvio padrão $\sigma = 1$ e determine os respetivos intervalos de confiança para μ ao nível de confiança $\gamma = 0.91$.
2. Obtenha a proporção de intervalos de confiança gerados em 2. que contêm o valor esperado $\mu = 0.5$.

Indique o quociente entre o valor obtido em 2. e o nível de confiança γ, arredondado a 4 casas decimais.

Answer: 0.9958 ✔

Enunciado do exercício 9

Considere que $(X_1, \ldots, X_n)$ é uma amostra aleatória de dimensão n de uma população X com distribuição exponencial. Admita que $E(X) = \mu$ é desconhecido e que

$$T = \frac{2n\,\bar{X}}{\mu} \sim \chi^2_{(2n)},\ \forall \mu \in \mathbb{R}^+.$$

Para testar $H_0 : \mu = \mu_0 = 5$ contra $H_1 : \mu = \mu_1 = 6.2$ pode utilizar-se a estatística de teste T_0 obtida de T admitindo que a hipótese H_0 é verdadeira, rejeitando-se H_0, ao nível de significância α, se $T_0 > F^{-1}_{\chi^2_{(2n)}}(1-\alpha)$.

Fixando a semente em 5707, gere $m = 900$ amostras de dimensão $n = 25$ da distribuição exponencial com valor esperado μ_1. Aplique o teste de hipóteses, ao nível de significância $\alpha = 0.08$, a cada uma das $m = 900$ amostras geradas e calcule uma estimativa, $\hat{\beta}$, da probabilidade teórica do erro de 2ª espécie, β.

Obtenha o quociente de $\hat{\beta}$ e β e indique o resultado arredondado a 4 casas decimais.

Answer: 1.0119 ✔

Enunciado do exercício 10

Uma equipa de engenheiros está a analisar a distribuição da velocidade ($\boldsymbol{X}$, em m/s) do vento dominante em determinado local onde se pretende construir uma aerogare, tendo obtido a seguinte amostra com $\mathbf{200}$ observações:

3.6, 2.6, 3, 1.6, 5.4, 4.9, 3.4, 0.8, 3.7, 2.2, 1.5, 2.6, 1.5, 1.4, 3.4, 1.6, 6.3, 2.9, 3.7, 1.8, 1.1, 4, 0.3, 3.6, 3.6, 1.3, 1.5, 3.9, 1.7, 5.6, 4.5, 6.9, 3.8, 7.9, 3, 4.9, 1.8, 2.9, 2.6, 6.3, 0.7, 5.9, 5.8, 5.9, 4.7, 1.6, 0.7, 3.4, 5.2, 3.7, 2.6, 4.8, 3.2, 1.8, 5.9, 7.7, 2.2, 2.4, 1, 4.6, 2.6, 4.3, 1.7, 3, 5.7, 1.2, 2.5, 5.5, 3.2, 2.9, 4.4, 1.1, 2.9, 1.7, 5.3, 3.2, 1.4, 2.9, 1.8, 4.3, 2.2, 7.2, 1.5, 2.1, 5.9, 4, 1.8, 5.5, 5.2, 2.4, 1.9, 7.4, 1.3, 3.3, 1.4, 2.5, 4.6, 1, 1.8, 0.3, 4.5, 0.2, 0.7, 4.6, 4, 1.7, 1.5, 1.4, 8.6, 5.2, 1.5, 6.2, 3.6, 4.4, 2.8, 1.8, 2, 1.4, 2.2, 1.5, 4.5, 5.4, 6.1, 6.9, 1.6, 0.3, 0.5, 3, 5.2, 2, 2.4, 5, 3, 4.9, 3.8, 3.4, 4, 4.8, 1.9, 0.7, 3.8, 4.1, 6.5, 4.3, 2.9, 2.3, 3.2, 8.4, 3.8, 2.6, 4.1, 0.5, 4.7, 3.8, 2.4, 1.3, 4.6, 2, 3.6, 9.4, 3.8, 4.8, 6.2, 6.5, 6.4, 3.4, 1.2, 1.2, 3.1, 2, 1.6, 5.5, 3.5, 1.7, 1.8, 1.1, 3.6, 3.3, 4.7, 2.4, 4, 4.4, 2.4, 3.5, 2, 2.8, 2.8, 1.4, 5.2, 2.3, 5, 3.4, 4.2, 6, 6.4, 4.3, 2.1, 4.3, 2.3, 3

Os membros da equipa conjecturam que $\boldsymbol{X}$ possui distribuição de Rayleigh com parâmetro de escala σ, i. e., com função de distribuição dada por

$$F_0(x) = 1 - \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad x > 0.$$

Teste $H_0 : X \sim \mathbf{Rayleigh}(\sigma = 2.2)$ contra $H_1 : X \nsim \mathbf{Rayleigh}(\sigma = 2.2)$, procedendo do seguinte modo.

1. Fixe a semente em $\mathbf{4497}$ e selecione ao acaso e sem reposição uma subamostra de dimensão $n = 140$ da amostra original.
2. Divida o suporte da variável aleatória X, $\mathbb{R}^+$, em $k = 6$ classes equiprováveis sob H_0.
3. Agrupe as observações da subamostra selecionada em 1. nas classes definidas em 2. e obtenha o conjunto de frequências absolutas observadas associadas a essas classes.
4. Recorra às frequências absolutas observadas obtidas em 3. e calcule o valor-p do teste de ajustamento do qui-quadrado para as hipóteses referidas.

Com base neste procedimento, indique qual das cinco decisões abaixo deverá tomar a equipa de engenheiros.

Select one:

- a. Rejeitar H_0 aos n.s. de 5% e 10% e não rejeitar H_0 ao n.s. de 1%.
- b. Rejeitar H_0 ao n.s. de 10% e não rejeitar H_0 aos n.s. de 1% e 5%.
- c. Não rejeitar H_0 aos n.s. de 1%, 5% e 10%.
- d. Rejeitar H_0 aos n.s. de 1%, 5% e 10%. ✔
- e. Teste é inconclusivo.