Ex.1

No ficheiro de dados econ.xlsx encontra informação relativa a dados económicos dos Estados Unidos providenciados pela empresa FRED. Este conjunto de dados possui as seguintes variáveis: tempo (Data do registo); gcp (gastos de consumo pessoal, em biliões de dólares); pop (população total); tpp (taxa de poupança pessoal); ddesemp (duração mediana do desemprego, em semanas); ndesemp (número de desempregados, em milhares).

Considere as variáveis x_1 = ddesemp e x_2= ndesemp para os anos superiores ou iguais a 1991. Com recurso ao pacote ggplot produza um único gráfico que lhe permita fazer uma análise da evolução dessas duas variáveis para esses anos.

Uma vez que as variáveis podem não ter a mesma escala, antes de construir o gráfico proceda do seguinte modo:

a. Selecione os dados a usar.
b. Faça a seguinte transformação aos dados associados a cada variável:

$$x_k : z_{ik} = \frac{x_{ik} - \overline{x}_k}{s_{xk}}, i = 1,2, \dots, n$$

onde n é a dimensão da amostra, $\overline{x}_k$ e s_{xk} correspondem, respetivamente, à média e desvio-padrão da amostra associada à variável x_k.

Ex.2

O ficheiro TIME_USE_24092022.csv contém uma compilação de dados enviados por diversos países para a OCDE (Organização para a Cooperação e Desenvolvimento Económico) sobre o tempo médio diário (em minutos) despendido pelas pessoas entre os 15 e os 64 anos em diferentes tipos de ocupações.

Leia o ficheiro de dados no R e elimine todos os registos referentes à África do Sul (dados incompletos).

Submeta um ficheiro em formato PDF com uma única página A4, que inclua, num único gráfico, dois diagramas de extremos e quantis que permitam comparar os tempos médios diários registados para Total (Homens+Mulheres) em duas ocupações distintas: Lazer e Trabalho não remunerado.

Ex.3

O ficheiro GENDER_EMP_19032023152556091.txt contém uma compilação de dados sobre emprego enviados por diversos países para a OCDE (Organização para a Cooperação e Desenvolvimento Económico).

Com recurso ao pacote ggplot produza um único gráfico de barras que permita comparar os valores da variável EMP3 (*Unemployment rate, by sex and age group*) entre homens e mulheres nos grupos etários 15–24, 25–54 e 55–64, registados em 2010 no seguinte país: Canada.

Ex.4

Fixando a semente em 1153, gere uma amostra de dimensão k=1087 proveniente de uma distribuição Exponencial de parâmetro λ=9. Os valores gerados correspondem aos tempos entre acontecimentos sucessivos.

Considere agora a soma sucessiva destas observações, i.e., se x_i designar o i-ésimo valor gerado, então $s_j = \sum_{i=1}^{j} x_i$é o instante de ocorrência do j-ésimo acontecimento. Seja $T = [s_{1087}]$ o menor número inteiro maior ou igual ao instante de ocorrência do último acontecimento.

Divida o intervalo]0, T] em intervalos de amplitude unitária e contabilize o número de acontecimentos que ocorreram em cada um desses subintervalos.

Calcule a média do número de acontecimentos por subintervalo e de seguida calcule o desvio absoluto entre este valor e o valor esperado (teórico) do número de acontecimentos num subintervalo. Indique este desvio arredondado a 4 casas decimais.

Ex.5

Ensaios de Bernoulli independentes, cada um dos quais com probabilidade de sucesso 0.35, são sucessivamente realizados. Seja X o número de insucessos até ao primeiro ensaio que resulta em sucesso. A distribuição da variável aleatória X é conhecida por distribuição geométrica de parâmetro p=0.35, cuja função (massa) de probabilidade é dada por:

- $(1-p)^x \cdot p,\ x = 0,1,2,\ldots;$
- $0\ caso\ contrário\ .$

Podemos gerar valores de uma distribuição geométrica a partir de uma distribuição uniforme contínua usando o método de transformação inversa. Nesse sentido, requer-se a execução dos seguintes passos:

i. Simula-se um valor, u, proveniente de uma distribuição uniforme no intervalo]0,1[.
ii. Se $FX(x-1) < u \leq FX(x)$, aceita-se x como um valor simulado de X, onde FX(x) é a função de distribuição de X.

Fixando a semente em 1891, implemente este método de simulação estocástica repetindo os passos anteriores até obter uma amostra de dimensão n=1173.

Indique a proporção de valores simulados que são superiores à soma da média com o desvio padrão amostrais, de entre os que são superiores à respetiva média amostral. Apresente o resultado com 4 casas decimais.

Ex.6

Considere a variável aleatória X que representa o primeiro algarismo de um número inteiro escrito em base decimal. Admita que X possui distribuição de Benford, com função de probabilidade dada por:

$$P(X = x) = \log_{10}\left(1 + \frac{1}{x}\right),\ x \in \{1,2,\ldots,9\}.$$

1. Calcule a probabilidade de X ser igual a 2 ou 8.
2. Obtenha a fração de potências de dois no intervalo $[2^{2,}2^{21}]$ cujo primeiro algarismo é igual a 2 ou 8.
3. Calcule o desvio absoluto entre os valores calculados em 1. e 2.
4. Indique este desvio arredondado a 4 casas decimais.

Ex.7

Fixando a semente em 1526, simule m=2492 amostras de dimensão n=15 de uma população normal de média nula e variância unitária. Para cada uma das amostras, calcule a soma dos quadrados dos valores observados.

Indique a diferença em valor absoluto (arredondado a 4 casas decimais), entre o quantil de probabilidade 0.54 da amostra das somas dos quadrados dos valores observados e o quantil correspondente à distribuição teórica da soma de quadrados de variáveis normais reduzidas independentes.

Nota**:** Use a função quantile com a opção type=2.

Ex.8

Considere uma variável aleatória com distribuição de Cauchy, com parâmetros de localização e escala iguais a −1.4 e 1.8, respetivamente.

Usando o R e fixando a semente em 1961, gere uma amostra de dimensão n=109 desta população.

Represente num único gráfico:

Os valores gerados ordenados por ordem crescente versus os quantis de probabilidade i/(109+1), i=1,…,109desta população;

Os valores gerados ordenados por ordem crescente versus os quantis de probabilidade i/(109+1), i=1,…,109 de uma população normal com valor esperado μ=2.2 e variância σ^2=1.8;

A reta bissetriz dos quadrantes ímpares.

Ex.9

Para a construção de intervalos de confiança para o parâmetro p de uma distribuição de Bernoulli podemos recorrer à variável fulcral:

$$z_1 = \frac{\overline{X} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}}$$

obtida pela aplicação do teorema do limite central a uma amostra aleatória de tamanho n suficientemente grande da referida população. Duas variantes são possíveis:

Método 1

Usando Z1, não é difícil mostrar que os limites do intervalo de confiança são as soluções da seguinte equação do segundo grau em p:

$$\overline{X}^2 - 2p\overline{X} + p^2 - z^2 \cdot \frac{p \cdot (1-p)}{n} = 0$$

em que $\overline{X}$ representa a média amostral e$z = \phi^{-1} \cdot (1 + \gamma^2)$, para um nível de confiança aproximado γ∈]0,1[.

Método 2

Uma segunda aproximação conduz à variável fulcral

$$z_2 = \frac{\overline{X} - p}{\sqrt{\frac{\overline{X} \cdot \left(1 - \overline{X}\right)}{n}}}$$

que permite a construção de intervalos de confiança de uma forma mais simples e habitual.

Com o objetivo de comparar os dois métodos e, em particular, avaliar a adequação da segunda aproximação, implemente os seguintes passos no R:

5. Fixe a semente em 1645 e para cada valor de n ∈ {30,50,100,200,300,500,1000}:
    a. gere k=2000 amostras de tamanho n de uma distribuição de Bernoulli com parâmetro p=0.7;
    b. para cada amostra gerada, calcule a diferença entre os comprimentos dos intervalos de confiança construídos pelo Método 2 e pelo Método 1, com um nível de confiança aproximado γ=0. 97;
    c. calcule a média das k=2000 diferenças anteriores.
6. Construa um gráfico que ilustre a variação das diferenças médias em função do tamanho da amostra.

Ex.10

Considere uma variável aleatória X com distribuição Normal de valor esperado μ desconhecido e variância σ^2=4 Construa um teste de hipóteses H0: μ=50.1, contra H1: μ≠50.1, ao nível de significância de α=0.1.

Com recurso ao R e fixando a semente em 1187, gere m=200 amostras de dimensão n=39 dessa variável, admitindo que μ=51.3. Aplique o teste de hipóteses que construiu para cada amostra gerada, e use o conjunto de resultados para obter uma estimativa da probabilidade de o teste conduzir à não rejeição de H0. Indique o resultado com 3 casas decimais.