DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Probabilidade e Estatística

Segundo Semestre 2021-22
Departamento de Matemática
IST-UL

EXERCÍCIOS

1. **Análise preliminar de dados em R**

1.1 Considere os dados disponíveis abaixo, referentes ao peso (em Kg) de 40 bicicletas.

4.3	6.8	9.2	7.2	8.7	8.6	6.6	5.2	8.1	10.9
7.4	4.5	3.8	7.6	6.8	7.8	8.4	7.5	10.5	6.0
7.7	8.1	7.0	8.2	8.4	8.8	6.7	8.2	9.4	7.7
6.3	7.7	9.1	7.9	7.9	9.4	8.2	6.7	8.2	6.5

(a) Obtenha algumas medidas descritivas dos dados e comente.
(b) Determine o intervalo dos 25% menores pesos e o intervalo dos 25% maiores pesos da amostra, bem como a amplitude inter-quantil.
(c) Indique o quantil amostral de 0.68.
(d) Construa o diagrama de caule-e-folhas.
(e) Obtenha o histograma e a caixa de bigodes identificando possíveis *outliers*.

1.2 Num estudo relativo a fatores de risco para as doenças cardiovasculares, os níveis críticos de cotinina (produto metabólico da nicotina) foram registados para um grupo de fumadores e para um grupo de não-fumadores. Os dados recolhidos encontram-se na tabela abaixo.

Nível de cotinina (mg/ml)	Fumadores	Não-Fumadores
0-13	78	3300
14-49	133	72
50-99	142	23
100-149	206	15
150-199	197	7
200-249	220	8
250-259	151	9
300-399	412	11
Total	1539	3445

(a) Construa uma tabela com as frequências absolutas e relativas. Compare as distribuições de frequências das duas populações.
(b) Construa um histograma para comparar as duas populações.
(c) Com base nesta análise, o que pode dizer sobre o efeito do tabaco no nível de cotinina?

1.3 A taxa de mortalidade infantil corresponde ao número médio de mortes, de entre 1000 crianças nascidas vivas, antes de completarem um ano de vida. Os dados referentes à União Europeia (UE), relativos aos anos de 1960 até 2019 estão disponíveis em

https://www.pordata.pt/Europa/Taxa+de+mortalidade+infantil-1589

(a) Identifique a(s) populações e indique a variável em estudo.
(b) Selecione 5 países da zona Euro. Represente graficamente a tabela de distribuições de frequências, para cada um desse país e para a média da zona Euro (considerando os 27 países). Comente os resultados obtidos.

(c) Considere apenas os dados relativos a 1961 e a 2018. Elabore um gráfico que lhe permita visualizar as diferenças nestes anos, para os países selecionados.

(d) Considerando os dados relativos a 2018, indique a taxa média, mediana, variância e coeficiente de variação da taxa de mortalidade infantil relativa aos 31 países disponíveis na base de dados.

(e) Analise a evolução temporal da taxa de mortalidade infantil em Portugal.

(f) Pretendendo estudar uma possível relação entre taxa de mortalidade infantil e o PIB, considere agora os dados relativos ao PIB em Portugal, desde 1961, disponíveis em

https://www.pordata.pt/Portugal/Taxa+de+crescimento+real+do+PIB-2298

Faça uma análise gráfica de forma a explorar uma possível relação entre as duas variáveis.

1.4 Considere os dados *nym.2002*, disponíveis no software R, relativos ao tempo de prova de corredores que terminaram a maratona de Nova Iorque em 2002. Estes dados contêm a seguinte informação: ordenação final na competição, sexo, idade, nacionalidade e tempo de prova. Para aceder a estes dados deverá utilizar os comandos:
> data(nym.2002, package="UsingR")
> nym.2002

(a) Faça uma análise dos dados, em particular para ilustrar, caso exista, uma diferença assinalável nos tempos de prova entre homens e mulheres.

(b) Construa um diagrama de caixa de bigodes para ambas as séries, e comente os resultados obtidos.

1.5 Os dados designados por *Wine Data Set*, disponíveis no software R, contêm informação relativa aos resultados de análises químicas feitas a 178 pés de videira, todos localizados na mesma região (em Itália) mas provenientes de três vinhas distintas (denotadas por 1, 2 e 3 no ficheiro de dados). Os dados reportam-se a 13 variáveis: Teor de alcóol; Ácido málico; Cinzas; Alcalinidade das cinzas; Magnésio; Fenóis totais; Flavonóides; Fenóis não-flavonóides; Proantocianidinas ; Intensidade da cor; Tonalidade; OD280/OD315 e Prolina. Para aceder a estes dados deverá utilizar os comandos:
> install.packages('rattle.data')
> library(rattle.data)
> data(wine)

(a) Analise a variável Magnésio, em termos da suas características amostrais (média, variância, e outras medidas sumárias que achar pertinentes) e apresente um gráfico de frequências. Repita o exercício, mas agora considerando as observações divididas pela vinha (1, 2 ou 3) de onde provêm.

(b) Através da análise de um gráfico adequado, aponte possíveis *outliers* para as variáveis Teor de Alcóol, Fenóis totais e Prolina. Verifique se a análise por vinha altera a identificação dos possíveis *outliers*.

(c) Calcule o coeficiente de correlação amostral entre cada par de variáveis e analise os resultados.

2. Conceitos básicos de probabilidade

2.1 Sejam A e B dois acontecimentos tais que $P(A) + P(B) = x$ e $P(A \cap B) = y$. Determine, em função de x e de y, a probabilidade de:

(a) Não se realizar nenhum dos dois acontecimentos.

(b) Que se realize um e só um dos dois acontecimentos.

(c) Que se realize pelo menos um dos dois acontecimentos.

(d) Que se realize quanto muito um único acontecimento.

2.2 Uma urna contém 5 bolas brancas e 5 bolas pretas. Dois jugadores, A e B, tiram alternadamente e um de cada vez uma bola da urna. O jogador que tirar a primeira bola branca ganha a partida.

(a) Considere a experiência aleatória associada a este jogo e escreva o correspondente espaço de resultados.

(b) Calcule a probabilidade de cada jogador ganhar a partida sabendo que o jogador A é o primeiro a tirar a bola da urna.

(c) Responda às alíneas (a) e (b) mas agora considerando que as bolas são extraídas com reposição.

2.3 Um geólogo crê que existe petróleo numa certa região com probabilidade 0.8 e que, caso haja petróleo, a probabilidade de sair petróleo na primeira perfuração é de 0.5.

(a) Qual a probabilidade de sair petróleo na primeira perfuração?

(b) Tendo-se procedido à primeira perfuração da qual não resultou petróleo, qual é a nova probabilidade atribuída à existência de petróleo na região?

2.4 Para um certo tipo de cancro a taxa de prevalência é 0.005. Um teste diagnóstico para esta doença é tal que i) a probabilidade do teste resultar positivo quando aplicado a um indivíduo com cancro é 0.99; a probabilidade do teste resultar negativo quando o indivíduo não tem cancro é 0.95.

(a) Calcule o valor preditivo de teste, isto é, a probabilidade de um indivíduo ter cancro sabendo que o teste resultou positivo.

(b) Supondo que o teste foi aplicado duas vezes consecutivas ao mesmo doente e que das duas vezes o teste foi positivo, calcule a probabilidade do doente ter cancro (admita que, dado o estado do indivíduo, os resultados do teste em sucessivas aplicações, em qualquer indivíduo, são independentes). O que pode concluir quanto ao valor preditivo da aplicação do teste duas vezes consecutivas?

3. **Variáveis aleatórias discretas e contínuas**

3.1 Numa fábrica existem três máquinas iguais de uma mesma marca, que trabalham independentemente. A probabilidade de cada máquina avariar num dado espaço de tempo é 0.1. Seja X a variável aleatória que representa o número de máquinas que findo esse período de tempo estão a trabalhar. Determine:

(a) A função de probabilidade de X.

(b) A função de distribuição de X.

(c) O valor esperado, moda, mediana e variância de X.

3.2 O número de mensagens electrónicas recebidas por dia ($24h$) numa pequena empresa de entregas rápidas tem distribuição de Poisson com média igual a 10.

(a) Calcule a probabilidade de num dia a empresa não receber mais do que 7 mensagens.

3.3 Considere que o comprimento de uma barra de ferro é uma variável aleatória com distribuição Normal de valor esperado 10cm e desvio padrão 2cm. Só são aceites para comercialização barras com comprimento entre 8cm e 12cm inclusive. Qual a probabilidade de uma barra selecionada ao acaso ser aceite para comercialização?

3.4 Considere que a duração do tempo de vida, em centenares de horas, de uma componente eletrónica é uma variável aleatória com distribuição exponencial de valor esperado 0.5.

(a) Calcule a função de distribuição da variável aleatória X.

(b) Calcule a probabilidade de que a componente electrónica tenha uma duração de vida superior a 150h, sabendo que já funcionou pelo menos durante 100h.

4. **Pares aleatórios**

4.1 Sejam X e Y duas variáveis aleatórias discretas com função de probabilidade conjunta dada por:

$Y \backslash X$	1	2	3
1	1/9	0	1/18
2	0	1/3	1/9
3	1/9	1/6	1/9

(a) Determine:
   (i) A função de probabilidade marginal de X.
   (ii) A função de distribuição marginal de Y.
   (iii) $P(X + Y \leq 4)$.
   (iv) As funções de probabilidade de X condicionais a $Y = 1$ e $Y = 3$.
   (v) $E(X|Y = 1)$.

(b) Defina $E(X|Y)$.

(c) Diga, justificando, se X e Y são variáveis aleatórias independentes.

(d) Calcule a $V(X + Y)$.

4.2 Considere a variável aleatória bidimensional contínua (X, Y) com função densidade de probabilidade conjunta:

$$f_{X,Y}(x, y) = \begin{cases} 2, & 0 < x < y < 1 \\ 0, & \text{caso contrário} \end{cases}.$$

(a) Calcule o coeficiente de correlação entre X e Y.

(b) Calcule a $V(X|Y = y)$.

(c) Verifique que $E(X) = E[E(X|Y)]$.

4.3 O diâmetro interior de um tubo cilíndrico é uma variável aleatória X com distribuição normal de valor esperado 3 cm e desvio padrão 0.02 cm e a espessura Y do mesmo tubo é uma variável com distribuição normal de valor esperado 0.3 cm e desvio padrão 0.005 cm, independente de X.

(a) Calcule o valor esperado e o desvio padrão do diâmetro exterior do tubo.

(b) Calcule a probabilidade de que o diâmetro exterior do tubo exceda 3.62 cm.

5. **Teorema do limite central**

5.1 O tempo (em horas) que João Pestana dorme por noite é uma variável aleatória com distribuição uniforme no intervalo (7,12).

(a) Calcule a probabilidade de João Pestana dormir mais de 11 horas numa noite.

(b) Calcule a probabilidade de, em 20 noites, João Pestana dormir mais de 11 horas em pelo menos 3 dessas noites.

(c) Qual a probabilidade de João Pestana dormir mais de 1100 horas em 100 noites?

6. **Estimação pontual**

6.1 Considere uma urna com bolas brancas e pretas na proporção de 3/1 desconhecendo-se, no entanto, qual a cor dominante. Seja p a probabilidade de sair uma bola preta numa extracção.

Qual a estimativa de máxima verosimilhança de p se, ao extraírmos com reposição 3 bolas da urna, encontrássemos

(a) 1 bola preta?

(b) 2 bolas pretas?

(c) Suponha agora que desconhecíamos qualquer relação entre o número de bolas brancas e pretas. Qual a estimativa de máxima verosimilhança de p, se ao extrairmos 3 bolas com reposição encontrássemos 2 bolas pretas?

6.2 Certo tipo de pilhas tem uma duração (em horas) que se distribui exponencialmente com valor esperado μ. A duração global de 10 pilhas tomadas aleatoriamente foi de 1740 horas. Qual a estimativa de máxima verosimilhança da probabilidade de uma pilha durar mais de 200 horas?

7. **Estimação intervalar**

7.1 Suponha que a intensidade da corrente, em amperes, num certo circuito é uma variável aleatória com distribuição normal. Uma amostra de dimensão 12 desta variável aleatória conduziu aos seguintes resultados:

2.3 1.9 2.1 2.8 2.3 3.6 1.4 1.8 2.1 3.2 2.0 1.9

Construa um intervalo de confiança de 99% para:

(a) O valor esperado da intensidade da corrente.

(b) O desvio padrão da intensidade da corrente.

7.2 Uma amostra de 100 peças de uma linha de produção revelou 17 peças defeituosas.

(a) Determine um intervalo de confiança a 95% para a verdadeira proporção p de peças defeituosas produzidas.

(b) Quantas peças adicionais devemos recolher para estarmos confiantes a 98% que o erro de estimação de p seja menor que 2%?

8. **Testes de hipóteses**

8.1 Da produção diária de determinado fertilizante tiraram-se seis pequenas porções que se analisaram para calcular a percentagem de nitrogénio. Os resultados foram os seguintes:

6.2 5.7 5.8 5.8 6.1 5.9

Sabe-se, por experiência, que o processo de análise fornece valores com distribuição que se pode considerar normal com $\sigma^2 = 0.25$.

(a) Suportam as observações a garantia de que a percentagem esperada de nitrogénio, μ, é igual a 6% ao nível de significância de 10%?

(b) Responda à alínea anterior usando o valor-p.

8.2 Uma máquina de ensacar açúcar está regulada para encher sacos de 16 quilos. Para controlar o funcionamento escolheram-se ao acaso 15 sacos da produção de determinado período, tendo-se obtido os pesos seguintes:

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

Admitindo que o peso de cada saco possui distribuição normal:

(a) Que conclusão pode tirar sobre a regulação da máquina?

(b) Que evidência fornece a concretização de S^2 sobre a hipótese $H_0 : \sigma^2 = 0.25$?

8.3 Uma empresa fabricante de lâmpadas considera que a sua produção é eficaz se a probabilidade de se selecionar ao acaso uma lâmpada não defeituosa for de 90%. Para verificar a qualidade da produção das lâmpadas, foi efetuado um teste a 200 lâmpadas, tendo-se verificado que 24 tinham defeitos. A que conclusão deve chegar o estatístico da empresa? Justifique.

8.4 Suponha que o departamento de defesa acredita que a distribuição de probabilidade do número de avarias, durante uma dada missão, ocorridas numa determinada zona do submarino Polaris segue uma distribuição de Poisson. Os dados relativos a 500 destas missões são os seguintes:

número de falhas por missão	0	1	2	3	4
número de missões	185	180	95	30	10

(a) Teste ao nível de significância de 5% a hipótese da referida variável aleatória possuir uma distribuição de Poisson, com valor esperado igual a 1.

8.5 Numa experiência com tubos de vácuo foram observados os tempos de vida (em horas) de 100 tubos, tendo-se registado as seguintes frequências absolutas:

Intervalo	$]0,30]$	$]30,60]$	$]60,90]$	$]90,+\infty[$
Frequências absolutas	41	31	13	15

Serão os dados consistentes com a hipótese de o tempo de vida de um tubo de vácuo ter distribuição exponencial com valor esperado igual a 50 horas? Calcule um intervalo para o valor-p e comente.

9. **Introdução à regressão linear simples**

9.1 A perda percentual de massa (Y) de uma certa substância metálica (quando exposta a oxigénio seco a 500^oC) depende do período de exposição (x, em hora). Cinco medições conduziram a: $\sum_{i=1}^5 x_i = 12$, $\sum_{i=1}^5 x_i^2 = 32.5$, $\sum_{i=1}^5 y_i = 0.177$, $\sum_{i=1}^5 y_i^2 = 0.006789$, $\sum_{i=1}^5 x_i y_i = 0.4685$, onde

$$[\min_{i=1,\ldots,5} x_i,\ \max_{i=1,\ldots,5} x_i] = [1.0,\ 3.5]\,.$$

Calcule as estimativas de mínimos quadrados dos parâmetros da reta de regressão linear simples de Y em x.

9.2 Por forma a analisar a relação entre a percentagem de humidade relativa no local de armazenagem (x) e o teor de humidade de fibra sintética aí armazenada (Y), foram obtidos os seguintes dados em 10 localizações distintas: $\sum_{i=1}^{10} x_i = 4.50\,,$ $\sum_{i=1}^{10} x_i^2 = 2.1062\,,$ $\sum_{i=1}^{10} y_i = 1.18\,,$ $\sum_{i=1}^{10} y_i^2 = 0.1470\,,$ $\sum_{i=1}^{10} x_i y_i = 0.5551\,,$ onde

$$[\min_{i=1,\ldots,10} x_i,\ \max_{i=1,\ldots,10} x_i] = [0.29,\ 0.61].$$

(a) Após ter enunciado as hipóteses de trabalho que entender convenientes, calcule as estimativas de máxima verosimilhança dos parâmetros da reta de regressão linear simples de Y em x.

(b) Obtenha a estimativa do valor esperado do teor de humidade de fibra sintética quando armazenada num local com humidade relativa de 0.35.

9.3 Para descrever a relação existente entre o volume de uma massa de um gás ideal clássico e a respetiva pressão, registaram-se 10 valores do logaritmo de base 10 do volume, x (com o volume medido em polegadas ao quadrado), e os correspondentes valores experimentais do logaritmo de base 10 da pressão, Y (com a pressão medida em psi). Pretendendo avaliar-se a validade do modelo de regressão linear simples para descrever a relação existente entre o logaritmo da pressão do gás e o logaritmo do seu volume, efetuaram-se os seguintes cálculos: $\sum_{i=1}^{10} x_i = 19.4$, $\sum_{i=1}^{10} x_i^2 = 38.06$, $\sum_{i=1}^{10} y_i = 14.8$, $\sum_{i=1}^{10} y_i^2 = 22.76$, $\sum_{i=1}^{10} x_i y_i = 28.12$

(a) Obtenha as estimativas de mínimos quadrados dos parâmetros da reta de regressão linear simples de Y em x e interprete o significado do sinal da estimativa do parâmetro β_1 do modelo.

(b) Indicando as hipóteses de trabalho convenientes, obtenha um intervalo de confiança a 95% para o parâmetro β_1 do modelo de regressão linear simples de Y em x.

9.4 É geralmente aceite que a frequência cardíaca (Y, em *batimentos por minuto*) é influenciada pela temperatura corporal dos seres humanos (x, em ^oC). Um conjunto de 130 medições independentes conduziu aos seguintes resultados: $\sum_{i=1}^{130} x_i = 4\,784.7$, $\sum_{i=1}^{130} x_i^2 = 176\,121.67$, $\sum_{i=1}^{130} y_i = 9\,589$, $\sum_{i=1}^{130} y_i^2 = 713\,733$, $\sum_{i=1}^{130} x_i y_i = 353\,018.5$. Calcule o valor do coeficiente de determinação e comente a utilidade do modelo ajustado.