Exercícios – Estatística e Delineamento 2021-22

0 Conceitos introdutórios de Estatística e do programa R

1. Um agricultor instalou um pluviómetro para medir a precipitação num dado terreno. Durante um ano, obteve os seguintes totais mensais (em mm):

Janeiro	101.0	Maio	26.7	Setembro	5.7
Fevereiro	60.7	Junho	10.5	Outubro	51.7
Março	75.1	Julho	2.5	Novembro	50.1
Abril	19.9	Agosto	39.8	Dezembro	170.6

Numa sessão de trabalho no programa R, responda às seguintes alíneas:

(a) Crie um vector com os 12 totais mensais indicados. Designe o objecto criado por `precip`.

(b) Crie o vector `meses` com o nome dos 12 meses do ano.

(c) Associe a cada medição o nome do respectivo mês, utilizando o comando `names` do R.

(d) Calcule, com a ajuda dos comandos estatísticos elementares de que o R dispõe, as seguintes quantidades:

i. A precipitação total anual;
ii. A precipitação mensal média;
iii. A precipitação mensal mediana;
iv. O terceiro quartil das precipitações mensais;
v. A variância das precipitações mensais;
vi. O desvio padrão das precipitações mensais;
vii. A precipitação mensal mínima;
viii. A precipitação mensal máxima;

(e) Aplique o comando summary ao objecto precip e inspeccione o resultado.

(f) Seleccione o subvector

i. da precipitação no mês de Outubro;
ii. das precipitações nos meses de Junho a Setembro (inclusive);

(g) Seleccione o subvector dos meses com precipitação

i. superior a 50 mm;
ii. acima da média.

(h) Identifique, com auxílio de comandos do R:

i. qual o mês onde se verificou a precipitação mínima;
ii. qual o mês onde se verificou a precipitação máxima.

(i) Aplique o comando `plot` ao vector `precip` que criou na alínea 1a. Comente o resultado.

(j) Execute os seguintes comandos e comente o resultado.

```
> plot(precip, type="l")
> plot(precip, type="h")
```

2. O programa R disponibiliza alguns conjuntos de dados. Os seus nomes e breves descrições podem ser consultados através do comando

```
> data()
```

Entre estes dados encontra-se o vector `sunspots`, onde se registam o número médio de manchas solares observadas nos dias de cada mês, entre 1749 e 1983[1]. Os valores podem ser vistos escrevendo apenas o nome do objecto.

(a) Determine o comprimento do vector `sunspots`, utilizando o comando `length`.

(b) Crie um histograma dos valores registados, utilizando o comando `hist`:

i. deixando que o comando defina as classes de valores utilizadas;

ii. pedindo a criação de classes de comprimento 10, começando em zero e acabando em 260.

(c) Calcule, com a ajuda do comando `quantile`:

i. os três quartis (primeiro quartil, mediana e terceiro quartil) dos dados;

ii. o nono decil dos dados.

(d) Aplique o comando `summary` ao objecto `sunspots` e inspeccione o resultado.

(e) Construa um diagrama de extremos e quartis dos dados, utilizando o comando `boxplot`.

3. Uma experiência alimentar com coelhos utilizou quatro diferentes dietas (designadas pelas letras A, B, C e D) e cinco diferentes tratamentos (indicados pelos números de 1 a 5). Ao fim dum certo período de tempo foram medidos os pesos médios dos coelhos submetidos a cada combinação de dieta e tratamento. Eis os resultados obtidos:

peso	dieta	tratamento
1.5	A	1
1.4	A	2
1.4	A	3
1.2	A	4
1.4	A	5
2.7	B	1
2.9	B	2
2.1	B	3
3.0	B	4
3.3	B	5
2.1	C	1
2.2	C	2
2.4	C	3
2.0	C	4
2.5	C	5
1.3	D	1
1.0	D	2
1.1	D	3
1.3	D	4
1.5	D	5

Registe-se que a primeira coluna é numérica, enquanto que as restantes são variáveis categóricas (*factores*), embora numa delas as designações das diferentes categorias (*níveis do factor*) sejam números.

[1] Na realidade, `sunspots` não é um vector, mas um objecto de outro tipo, designado `ts` (as inicias de *time series*, ou seja, série cronológica). No entanto, para aquilo que se pede neste exercicio o objecto comporta-se como um vector.

(a) Crie vectores correspondentes a cada coluna (use os nomes acima indicados para esses vectores). O vector `peso` deverá ser numérico, enquanto que `dieta` e `tratamento` deverão ser definidos como factores (usando o comando `factor`).

(b) A forma mais frequente de armazenar dados no `R` é em objectos da classe *data frame*, criados utilizando o comando `data.frame`. A partir dos três vectores criados na alínea anterior, crie uma *data frame* de nome `coelhos`.

(c) Aplique o comando `summary` à *data frame* `coelhos`. Comente os resultados, e em particular a forma como o comando lida com as colunas de diferente natureza.

(d) Seleccione apenas a coluna correspondente aos pesos e calcule o respectivo valor médio.

(e) Seleccione apenas as linhas correspondentes à dieta `C`.

(f) Usando o comando `apply`, calcule o valor máximo em cada coluna. Comente.

4. A distribuição Normal (ou Gaussiana) é uma distribuição para variáveis aleatórias contínuas que tomam valores em $\mathbb{R}$. Tem dois parâmetros, sendo um o valor médio (ou valor esperado) μ e outro o desvio padrão σ ou, equivalentemente, a variância σ^2. Indica-se que uma variável aleatória X tem distribuição Normal com parâmetros μ e σ escrevendo $X \frown \mathcal{N}(\mu, \sigma)$.
Nota: Nesta disciplina, usar-se-á a convenção de identificar o segundo parâmetro como a variância de X. No entanto, neste exercício utiliza-se o desvio padrão (raíz quadrada da variância), sendo essa a forma de identificar o segundo parâmetro nos comandos do `R` relativos à distribuição Normal.

   Valores duma função densidade Normal podem ser calculados no `R` através do comando `dnorm`; valores da função distribuição cumulativa com o comando `pnorm`; e quantis com o comando `qnorm`. Pode gerar-se uma amostra aleatória de valores duma distribuição Normal através do comando `rnorm`.

   (a) Utilizando o comando `curve` do `R`, trace o gráfico da função densidade duma distribuição Normal com parâmetros $\mu=3$ e $\sigma=2$.

   (b) Seja X uma variável aleatória com distribuição $\mathcal{N}(\mu=3, \sigma=2)$.

      i. Calcule, usando o `R` e usando as tabelas da Normal, $P[X \leq 4]$; $P[X > 4]$ e $P[-1 < X < 4]$.

      ii. Calcule, usando o `R` e usando as tabelas da Normal, o quantil de ordem 0.975 da distribuição de X.

      iii. Usando o `R`, proceda à geração duma amostra aleatória com $n=100$ observações. Construa um histograma e calcule a média amostral $\overline{x}$ e a variância amostral s^2. Comente.

   Responda às seguintes alíneas com base na sua amostra aleatória, admitindo desconhecidos os valores de μ e σ que usou na sua geração. Os quantis da distribuição *t-Student* podem ser obtidos através do comando `qt`. Na página web de Estatística, aqui, encontram-se expressões úteis para o intervalo de confiança e os testes de hipóteses.

      iv. Calcule um intervalo a 95% de confiança para a média populacional μ. Comente.

      v. Teste a hipótese de que a média populacional seja realmente $\mu=3$. Comente.

      vi. Teste a Hipótese Nula que a média populacional seja $\mu \geq 3$. Comente.

      vii. Repita as 3 alíneas anteriores utilizando o comando `t.test` do `R`.

5. A distribuição F (também conhecida por distribuição de Fisher-Snedecor) é uma distribuição para variáveis aleatórias contínuas que tomam valores não negativos. Tem dois parâmetros, que designamos ν_1 e ν_2. Indica-se que uma variável aleatória X tem distribuição F com parâmetros ν_1 e ν_2 escrevendo $X \frown F(\nu_1, \nu_2)$. Valores da respectiva função densidade podem ser calculados no `R` com a função `df`; valores da função distribuição cumulativa com a função `pf`; e quantis com a função `qf`. Pode gerar-se uma amostra aleatória de valores duma distribuição F através do comando `rf`.

(a) Utilizando o comando do `R curve`, trace o gráfico da função densidade duma distribuição F com $\nu_1 = 3$ e $\nu_2 = 10$ graus de liberdade.

(b) Repita a alínea anterior, mas alterando os valores dos parâmetros na distribuição.

(c) Seja X uma variável aleatória com distribuição $F(3, 10)$.

i. Calcule $P[X \leq 4]$, $P[X < 4]$ e $P[X > 4]$.

ii. Calcule a mediana de X.