POLITÉCNICO DE LEIRIA | ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Departamento de Matemática

Estatística – 2.º semestre de 2019/2020

Ficha n.º 2 — Apresentação e resumo de dados

Na resolução dos exercícios das fichas deve ser utilizado o *package* **R Commander** do *software*[1] R. Para tal, após abrir o programa R, deverá utilizar uma das duas seguintes opções:

- utilizar o comando: library(Rcmdr).
- *Packages* $\rightarrow$ *Load packages* ... $\rightarrow$ em *Select one* escolher *Rcmdr*.

1. Considere o ficheiro $DadosNBA2013.xls$ disponível na plataforma EAD. Esta base de dados contém informações sobre 505 dos jogadores da NBA, recolhidos durante o ano de 2013. Os dados foram adaptados do ficheiro $nba\_ht\_wt.xls$, disponível em `http://www.stat.ufl.edu/~winner/datasets.html`. As variáveis são as seguintes:

   - *Posição* - posição do jogador em campo (base, poste ou extremo).
   - *Peso* - peso do jogador, em kg.
   - *Altura* - altura do jogador, em cm.
   - $Idade\_grupo$ - classe etária do jogador.

   (a) Indique a unidade estatística e a população em análise.

   (b) Classifique as variáveis em estudo.

   (c) Indique uma variável quantitativa discreta que pudesse ser considerada.

   (d) Indique a dimensão da amostra.

   (e) Importe o ficheiro $DadosNBA2013.xls$ e grave-o com o tipo $DadosNBA2013.Rdata$.

   (f) Introduza uma nova observação com os dados: Extremo, 104, 212, Jovem.

   (g) Verifique a existência de observações omissas na base de dados.

   (h) Construa a variável IMC (Índice de Massa Corporal) definida como $IMC = \dfrac{Peso}{Altura^2}$, considerando *Altura* em metros.

   (i) Agrupe a variável *Altura* num número adequado de classes de igual amplitude.

   (j) Recodifique a variável *Altura* numa nova variável, *Classe_altura*, onde:

   - $Altura \leq 180$ - baixo;
   - $180 < Altura < 195$ - médio;
   - $195 \leq Altura \leq 210$ - alto;
   - $Altura > 210$ - muito alto.

   (k) Grave a base de dados com o nome $DadosNBA2013\_Res.Rdata$.

[1] – O programa R é *freeware*. Para instalar deverá proceder conforme instruções disponibilizadas no Moodle.

2. Considere a base de dados *Carrosusados.Rdata* que contém informações sobre algumas viaturas disponíveis para venda num certo stand, no ano de 2014. As variáveis são as seguintes:

   - *Marca* - marca da viatura;
   - *Ano* - ano de fabrico da viatura (entre 2007 e 2010);
   - *Km* - quilometragem atual da viatura (até 100000 km);
   - *Cavalos* - potência da viatura em cavalos-vapor (CV);
   - *Garantia* - tipo de garantia fornecida pelo vendedor;
   - *Preço* - preço da viatura em euros (até 12000 €).

   (a) Indique qual é a população em estudo e qual é a dimensão da amostra analisada.

   (b) Classifique os atributos em análise.

   (c) Represente os quadros de frequências para as variáveis *Marca* e *Garantia*, ordenadas corretamente.

   (d) Indique a percentagem de automóveis com algum tipo de garantia.

   (e) Indique o número de automóveis da marca *Peugeot*.

   (f) Indique a moda das variáveis *Marca* e *Garantia*.

   (g) Construa uma nova variável que represente o preço do automóvel por cada cavalo de potência.

   (h) Construa uma nova variável que indique se a marca do automóvel é, ou não, de origem francesa (*Citroen*, *Peugeot*, *Renault*).

   (i) Represente graficamente, através de um diagrama de barras e de um diagrama de setores, as variáveis *Marca* e *Garantia*. Comente os resultados obtidos.

   (j) Represente graficamente, através de um histograma, a variável *Preço*. Comente os resultados obtidos.

   (k) Represente graficamente, através de um diagrama de barras, a variável *Marca* em função da variável *Garantia*. Comente os resultados obtidos.

   (l) Construa uma nova base de dados que contenha unicamente os automóveis de origem francesa e resolva novamente a alínea d). Comente os resultados obtidos.

   (m) Para os automóveis de origem francesa, represente graficamente a variável que indica se estes têm ou não mais de 80 CV de potência.

3. Considere a base de dados *RioLena.Rdata* que contém diversas informações sobre o Rio Lena, observadas em 36 ocasiões. As variáveis são as seguintes:

   - *PTempo* - Período de tempo em que as observações são realizadas;
   - *Temperatura* - Temperatura em ^{o}C;
   - *O2* - O2 Dissolvido (mg/L);
   - *Cloretos* - Cloretos (mg/L);
   - *Condutividade* - Condutividade (mS/cm);
   - *SolidosSusp* - Solidos Suspensos Totais (mg/L);

- *Alcalinidade* - Nível de Alcalinidade.

(a) Classifique os atributos *Alcalinidade* e *Cloretos*.

(b) Construa o quadro de frequências para a variável *Alcalinidade*.

(c) Para as situações em que a *Alcalinidade* não é elevada, represente graficamente a variável *Condutividade*. Comente o gráfico obtido.

(d) Calcule as medidas de tendência central (média e mediana) para a variável *Temperatura*. Comente os resultados obtidos.

(e) Calcule as medidas de tendência não central (mínimo, quartis, percentis 12 e 58, máximo) para a variável *Temperatura*. Comente os resultados obtidos.

(f) Calcule as medidas de dispersão (amplitude interquartil, amplitude total, desvio padrão, coeficiente de variação) e assimetria para a variável *Temperatura*. Comente os resultados obtidos.

(g) Represente graficamente o histograma de frequências relativas para a variável *Solidos-Susp*, considerando 5 classes de igual amplitude. Comente o gráfico obtido quanto à assimetria.

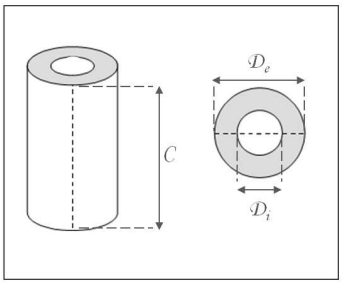
(h) Obtenha o diagrama de extremos e quartis para a variável *Condutividade*.

(i) Obtenha o diagrama de extremos e quartis para a variável *SolidosSusp*. Comente o gráfico obtido quanto à localização, dispersão e assimetria. Identifique o *outlier* detetado.

(j) Em qual dos períodos de tempo considerados é que se observou uma *Temperatura* média mais elevada? E mediana?

(k) Para qual dos níveis de *Alcalinidade* é que a condutividade é mais assimétrica?

4. Num processo de produção foram selecionadas aleatoriamente 45 peças para inspeção, tendo-se registado as seguintes caraterísticas de cada peça:



- *Comprimento*, C (em mm);
- *Diametro_exterior*, D_e (em mm);
- *Diametro_interior*, D_i (em mm);
- *Maquina*, máquina que produziu a peça (M1, M2, M3).

Os resultados encontram-se na base de dados *Produção.Rdata*.

(a) Para a variável *Comprimento* determine e interprete:

i. o valor da média e da mediana;

ii. as medidas de tendência não central: 1.º e 3.º quartis, percentil 5 e percentil 90;

iii. as medidas de dispersão: desvio padrão, amplitude total, amplitude interquartil e coeficiente de variação, comparando a dispersão desta variável com a registada em 3f);

iv. o valor do coeficiente de assimetria.

(b) Qual é o comprimento máximo observado nas 30% de peças mais curtas? E qual é o comprimento mínimo observado nas 20% de peças mais compridas?

(c) Qual das máquinas produz peças com:

i. menor diâmetro exterior médio;
ii. menor diâmetro exterior mediano;
iii. menor variabilidade do diâmetro exterior.

(d) Um dos critérios para classificar cada peça como defeituosa, ou não defeituosa, estabelece que esta será considerada defeituosa se o *Comprimento* se desviar mais do que $3\,mm$ dos $100\,mm$ desejados. Tendo em conta este critério:

i. construa uma nova variável que indique se cada peça inspecionada é defeituosa, ou não;
ii. compare a quantidade de peças defeituosas e não defeituosas com base:
   A. no quadro de frequências;
   B. num gráfico adequado;
iii. determine:
   A. o comprimento médio das peças defeituosas e das peças não defeituosas;
   B. o maior desvio relativamente ao comprimento ideal observado nas peças não defeituosas;
   C. a máquina que produziu mais peças defeituosas.

(e) Sabendo que a densidade do material usado no fabrico destas peças é igual a $\rho = 0.00786\ g/mm^3$, construa uma nova variável que represente o valor da massa das peças:

$$Massa = \rho\frac{\pi}{4}(D_e^2 - D_i^2)C\,,$$

sendo C o comprimento, D_e e D_i os diâmetros exterior e interior, respetivamente. Explique como procedeu e indique o valor médio, o desvio padrão, o mínimo e o máximo da massa das peças inspecionadas.

(f) Estude a existência de observações do diâmetro exterior consideradas *outliers*. No caso de existirem tais observações, identifique o valor do diâmetro observado e a máquina em que a peça foi produzida.

(g) Construa um diagrama de extremos e quartis, para a variável *Comprimento*, por cada máquina existente neste processo de produção. Indique qual é a máquina que produz peças com maior e com menor dispersão dos comprimentos.

(h) Compare a variabilidade do diâmetro exterior com a do diâmetro interior das peças.

(i) Entre as variáveis diâmetro exterior e diâmetro interior, qual é que apresenta uma distribuição mais simétrica?

(j) Estude o tipo de assimetria da variável diâmetro exterior, observada nas peças produzidas pela máquina M3:

i. considerando todas as peças;
ii. excluindo a(s) peça(s) com um diâmetro considerado *outlier*.

Soluções

1. (a) Unidade estatística: cada um dos jogadores da NBA no ano de 2013.
   População: conjunto de todos os jogadores da NBA no ano de 2013.

   (b) *Posição* - variável qualitativa nominal, assumindo que não há uma ordenação única das posições;
   Peso - variável quantitativa contínua;
   Altura - variável quantitativa contínua;
   Idade_grupo - variável qualitativa ordinal.

   (c) Por exemplo, "*Número de jogos efetuados mensalmente no ano de 2013*".

   (d) $n = 505$.

   (e) Data - Import - From Excel.

   (f) Edit data set - Add row.

   (g) Número de observações omissas: *Posição* - 0; *Peso* - 1; *Altura* - 3; *Idade_grupo* - 0.

   (h) Data - Manage - Compute.

   (i) $k = 9$. Data - Manage - Bin.

   (j) Data - Manage - Recode.

   (k) Data - Active - Save.

2. (a) População: conjunto de todas as viaturas disponíveis para venda num certo stand, no ano de 2014. $n = 61$.

   (b) *Marca* - variável qualitativa nominal;
   Ano - variável quantitativa discreta;
   Km - variável quantitativa contínua;
   Cavalos - variável quantitativa contínua;
   Garantia - variável qualitativa ordinal;
   Preço - variável quantitativa discreta (os preços finais são, no máximo, ao cêntimo sendo arredondados se necessário).

   (c)

```
counts:
  Marca
Citroen  Fiat  Ford  Opel  Peugeot  Renault  Smart
      7    14     6     9        7        9      9

percentages:
  Marca
Citroen   Fiat  Ford   Opel  Peugeot  Renault  Smart
  11.48  22.95  9.84  14.75    11.48    14.75  14.75
```

```
counts:
  Garantia
nenhuma  parcial  completa
     21       29        11

percentages:
  Garantia
nenhuma  parcial  completa
  34.43    47.54     18.03
```

   (d) 65.57%.

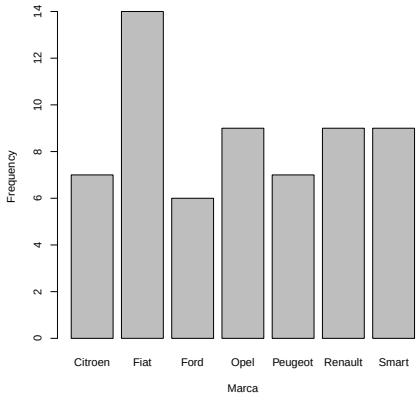
   (e) 7.

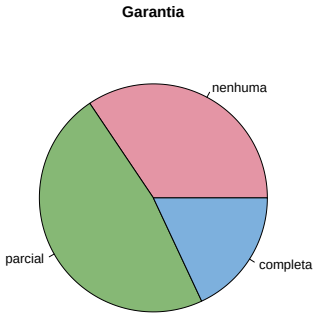
   (f) *Marca* - Fiat; *Garantia* - parcial.

   (g) Data - Manage - Compute.

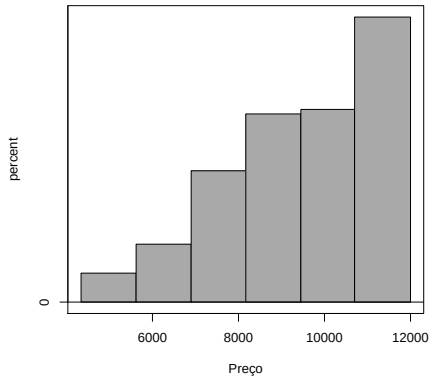
   (h) Data - Manage - Recode.

(i)

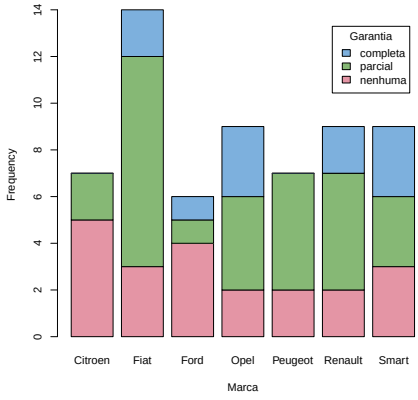




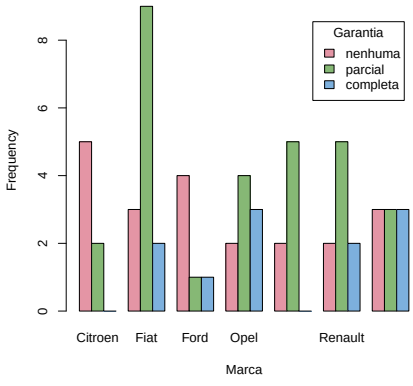
(j)



(k)

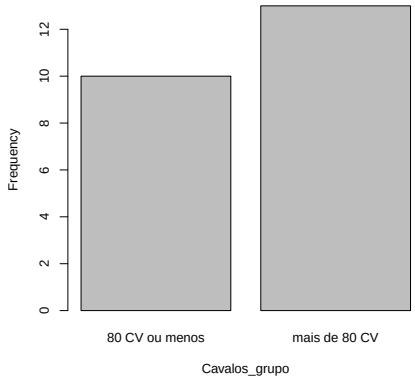


ou



(l) Data - Active - Subset. 60.87%.

(m)



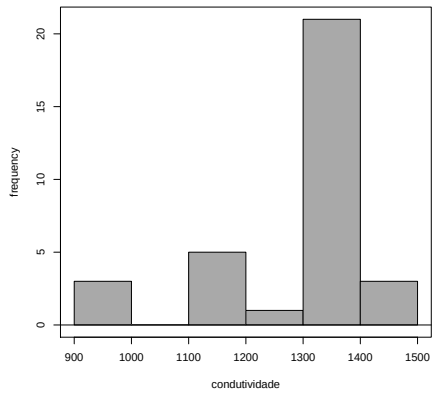
3. (a) *Alcalinidade* - variável qualitativa ordinal;
Cloretos - variável quantitativa contínua.

(b)

```
      counts:
alcalinidade
     pequena  média  elevada
          15     18        3

percentages:
alcalinidade
     pequena  média  elevada
       41.67  50.00     8.33
```

(c)

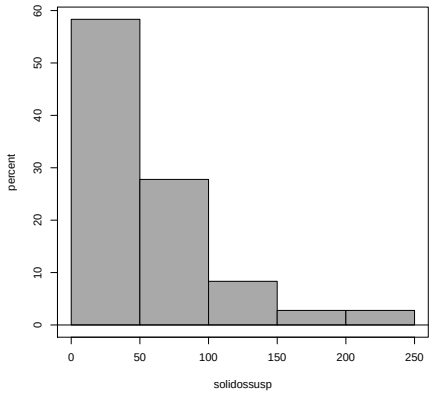


(d) $\overline{x} = 16.9472$ e $Me = 17.2$.

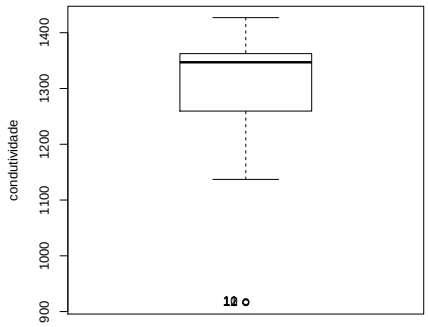
(e) $Min = 15.1$, $q_1 = 16.25$, $q_3 = 17.725$, $P_{12} = 15.22$, $P_{58} = 17.43$ e $Max = 18.6$.

(f) $I_q = 1.475$, $I_T = 3.5$, $s = 1.0880$, $Cv = 0.0642$ e $C_{skew} = -0.5969$.

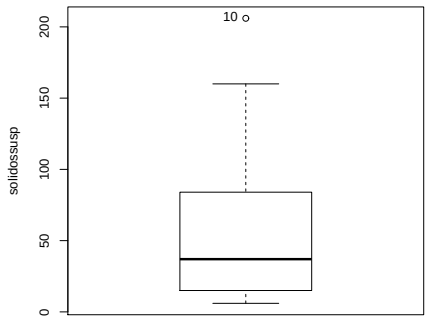
(g)



(h)



(i)



(j) Tarde; Tarde.

(k) Média.

4. (a) i. $\overline{x} = 99.9844$, $Me = 99.9$.
ii. $q_1 = 98.6$, $q_3 = 101.7$, $P_5 = 94.08$, $P_{90} = 102.88$.
iii. $s = 2.9822$, $I_T = 14.2$, $I_q = 3.1$, $Cv = 0.02983$ logo este conjunto de dados menos disperso.
iv. $C_{skew} = 0.0022$.

(b) 98.72 e 102.1.

(c) i. $M1$; ii. $M3$; iii. $M1$.

(d)

i. Data - Manage - Recode.

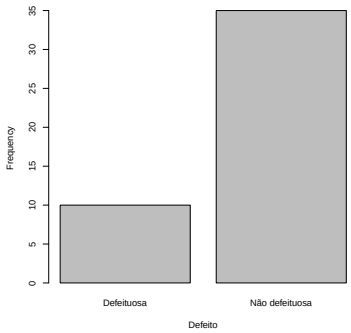
ii.

A.

```
      counts:
      Defeito
   Defeituosa  Não defeituosa
           10              35

 percentages:
      Defeito
   Defeituosa  Não defeituosa
        22.22           77.78
```

B.



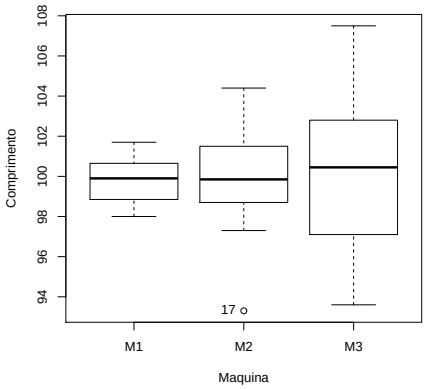
iii.

A. $\overline{x}_D = 99.9200$ e $\overline{x}_{ND} = 100.0029$; B. $2.7\,mm$; C. $M3$ (8 peças).

(e) $\overline{x} = 306.3809$, $s = 45.21298$, $Min = 194.3058$ e $Max = 453.8091$.

(f) Observação 19, $D_e = 31.7$, $M3$; observação 38, $D_e = 28.5$, $M2$.

(g)



(h) $s^2_{D_e} = 0.2812 < 3.3052 = s^2_{D_i}$.

(i) Diâmetro exterior ($c_s = 0.2275$).

(j)

i. Assimetria positiva ($c_s = 1.8576$).

ii. Distribuição aproximadamente simétrica ($c_s = -0.0698$).