MAE116 – Noções de Estatística

Grupos B e D – II Semestre de 2020

Lista de exercícios 1 – Estatística Descritiva I – CASA (gabarito)

Exercício 1.

Os dados a seguir representam medidas de Silício (%) de amostras de 13 sedimentos da cordilheira marinha central do Oceano Índico.

Amostra	1	2	3	4	5	6	7	8	9	10	11	12	13
Silício	20,1	22,3	19,2	24,2	21,5	20,4	46,8	22,2	22,1	19,1	22,6	21,3	17,7

(a) Calcule a média, a mediana, o desvio padrão e os quartis.

Média:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{20,1 + 22,3 + 19,2 + \cdots + 22,6 + 21,3 + 17,7}{13} = 23,03846\ \%,$$

ou seja, em média foi verificado um percentual de 23,04% de silício nos sedimentos observados na cordilheira marinha central do Oceano Índico.

Mediana:

Primeiro ordenamos os dados:

17,7 19,1 19,2 20,1 20,4 21,3 **21,5** 22,1 22,2 22,3 22,6 24,2 46,8

Em seguida, calculamos a posição da mediana

$$\frac{n+1}{2} = \frac{13+1}{2} = 7$$

Portanto,

$$M_d = 21,5\ \%.$$

Desvio padrão:

O desvio padrão corresponde à raiz quadrada da variância,

$$
\begin{aligned}
s =& \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \\
=& \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}} \\
=& \sqrt{\frac{(20,1 - 23,03846)^2 + (22,3 - 23,03846)^2 + \cdots + (17,7 - 23,03846)^2}{13-1}} \\
=& \sqrt{\frac{648,0108}{12}} = \sqrt{54,0009} = 7,34853\ \% \approx 7,35\ \%.
\end{aligned}
$$

De outra maneira,

$$
\begin{aligned}
s =& \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}}{n-1}} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2 - n\overline{x}^2}{n-1}} \\
=& \sqrt{\frac{20,1^2 + 22,3^2 + \cdots + 17,7^2 - (13)(23,03846)^2}{13-1}} \\
=& \sqrt{\frac{7548,03 - (13)(530,7706)}{12}} = \sqrt{\frac{648,0108}{12}} = 7,34853\ \% \approx 7,35\ \%.
\end{aligned}
$$

Quartis:

Quando um conjunto de dados ordenado é dividido em 4 partes iguais, temos 3 quartis, que correspondem aos percentis 25, 50 e 75, calculados como o valor da variável que ocupa a posição $p \times (n+1)$, com $p = 0,25$, $p = 0,50$ e $p = 0,75$.

Dados ordenados: 17,7 19,1 19,2 20,1 20,4 21,3 21,5 22,1 22,2 22,3 22,6 24,2 46,8.

Primeiro Quartil

Q1 = Percentil 25 ⇒ posição 0,25(13 + 1) = 3,5 ⇒ Q1 = (19,2 + 20,1)/2 = 19, 65.

Assim, 25% das amostas de sedimentos da cordilheira marinha central do Oceano Índico possuem quantidade de silício menor ou igual a 19,65%.

Segundo Quartil

Q2 = Percentil 50 = Mediana (já calculada). Assim, 50% das amostas de sedimentos da cordilheira marinha central do Oceano Índico possuem quantidade de silício menor ou igual a 21,5%.

Terceiro Quartil

Q3 = Percentil 75 ⇒ posição 0,75(13 + 1) = 10,5 ⇒ Q3 = (22,3 + 22,6)/2 = 22,45. Assim, 75% das amostas de sedimentos da cordilheira marinha central do Oceano Índico possuem quantidade de silício menor ou igual a 22,45%.

(b) Note que a amostra 7 apresenta um valor atípico. Remova esse valor e refaça o item anterior. Comente as diferenças encontradas.

Média:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{20,1 + 22,3 + 19,2 + \cdots + 22,6 + 21,3 + 17,7}{12} = 21,05833\ \% \approx 21,06\ \%.$$

Mediana:
Os dados ordenados, sem a observação 7, são

17,7 19,1 19,2 20,1 20,4 21,3 21,5 22,1 22,2 22,3 22,6 24,2

E então, calculamos a posição da mediana

$$\frac{n+1}{2} = \frac{12+1}{2} = 6,5$$

Neste caso, como $n = 12$ é par, a mediana será dada pela média dos valores das duas observações centrais, ou seja, média entre os valores que ocupam as posições 6 e 7 dos dados ordenados. Deste modo, temos

$$M_d = \frac{21,3 + 21,5}{2} = 21,4\ \%.$$

Desvio padrão:
O desvio padrão corresponde à raiz quadrada da variância,

$$\begin{aligned}
s =& \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \\
=& \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}} \\
=& \sqrt{\frac{(20,1 - 21,05833)^2 + (22,3 - 21,05833)^2 + \cdots + (17,7 - 21,05833)^2}{12-1}} \\
=& \sqrt{\frac{36,34917}{12}} = \sqrt{3,30447} = 1,81782\ \% \approx 1,82\ \%.
\end{aligned}$$

De outra maneira,

$$
\begin{aligned}
s = & \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}}{n-1}} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2 - n\overline{x}^2}{n-1}} \\
= & \sqrt{\frac{20,1^2 + 22,3^2 + \cdots + 17,7^2 - (12)(21,05833)^2}{12-1}} \\
= & \sqrt{\frac{5357,79 - (12)(443,4534)}{12}} = \sqrt{\frac{36,34917}{12}} = 1,81782\ \% \approx 1,82\ \%.
\end{aligned}
$$

Quartis:

Similar ao item (a), teremos:

Dados ordenados: 17,7 19,1 19,2 20,1 20,4 21,3 21,5 22,1 22,2 22,3 22,6 24,2.

Primeiro Quartil

Q1 = Percentil 25 $\Rightarrow$ posição 0,25(12 + 1) = 3,25 $\Rightarrow$ Q1 = (19,2 + 20,1)/2 = 19, 65.

Assim, 25% das amostras de sedimentos da cordilheira marinha central do Oceano Índico possuem quantidade de silício menor ou igual a 19,65%.

Segundo Quartil

Q2 = Percentil 50 = Mediana (já calculada). Assim, 50% das amostas de sedimentos da cordilheira marinha central do Oceano Índico possuem quantidade de silício menor ou igual a 21,4%.

Terceiro Quartil

Q3 = Percentil 75 $\Rightarrow$ posição 0,75(12 + 1) = 9,75 $\Rightarrow$ Q3 = (22,2 + 22,3)/2 = 22,25. Assim, 75% das amostas de sedimentos da cordilheira marinha central do Oceano Índico possuem quantidade de silício menor ou igual a 22,25%.

Comentários:

Note que, no item (a) a média ficou levemente afastada da mediana e com a remoção da observação 7 (item (b)) a média ficou mais pr'oxima da mediana. Este exemplo ilustra o fato de que a média é influenciada por valores atípicos.

A maior diferença que podemos observar com a remoção da observação 7 está relacionada ao desvio padrão. Com a remoção, o desvio padrão passou de 7,34853 para 1,81782, ou seja, o desvio padrão mais que quadruplica com a permanência da observação 7.

Em relação aos quartis praticamente não houve diferença nos dois casos, o que nos mostra a pouca (ou quase nenhuma) influência de valores atípicos no cálculo destas medidas de posição.

Exercício 2.

Um estudo foi realizado para comparar o peso da carne, em gramas, de mexilhões de dois locais: Sambaqui e Manguezal. Para isso, foram coletados e pesados, 15 mexilhões de cada local. Estatísticas descritivas obtidas para esses dados são apresentadas a seguir (Tabela 1):

Tabela 1: Estatísticas descritivas dos pesos dos mexilhões.

Local	n	Média	Desvio Padrão	Min	Q1	Mediana	Q3	Máx
Sambaqui	15	30,37	7,97	15,25	24,25	28,89	36,22	42,88
Manguezal	15	18,59	6,32	9,49	14,05	17,64	21,6	33,97

(a) Quais são as variáveis do estudo? Classifique-as.

Para cada tipo de local estamos interessados em estudar o peso da carne dos mexilhões. Logo, a variável *peso* corresponde a uma variável quantitativa contínua e o *local* corresponde a uma variável qualitativa nominal.

(b) 50% dos mexilhões do Sambaqui apresentam peso da carne inferior a qual valor? E se considerarmos 75% dos mexilhões?

Note que o valor que divide os dados em 50% é a mediana. Com base na Tabela 1 observa-se que 50% dos mexilhões do Sambaqui apresentam peso inferior ou igual a 28,89 gramas. Já o valor que divide os dados no percentil 75% corresponde ao terceiro quartil. Neste caso, observa-se na Tabela 1 que 75% dos mexilhões do Sambaqui apresentam peso inferior ou igual a 36,22 gramas.

(c) 25% dos mexilhões do Manguezal apresentam peso da carne superior a qual valor? E se considerarmos 75% dos mexilhões?

Podemos observar que 25% dos mexilhões do Manguezal apresentam peso superior ou igual a 21,6 gramas (terceiro quartil), enquanto que 75% dos mexilhões do Manguezal apresentam peso superior ou igual a 14,05 gramas (primeiro quartil).

(d) Escolhendo casualmente um mexilhão do Manguezal, o que seria mais provável: peso da carne maior ou menor que 21,6 gramas? Justifique.

Note que 21,6 gramas corresponde ao terceiro quartil, ou seja, 75% dos dados possuem valores dos pesos menores ou iguais a 21,6 gramas. Logo, se escolhermos ao acaso um mexilhão do Manguezal é mais provável que este tenha um peso inferior a 21,6 gramas.

(e) Utilizando o desvio padrão, compare os dois locais quanto à homogeneidade.

Note que para o Sambaqui temos um desvio padrão de 7,97 gramas, e para o Manguezal o desvio padrão é de 6,32 gramas. Logo, observando esses dois valores, podemos concluir que o Sambaqui é um pouco menos homogêneo quanto aos pesos dos mexilhões, pois apresenta uma maior variabilidade. De fato, o desvio padrão para o Sambaqui é 1,26 vezes maior que o desvio padrão para o Manguezal.

(f) Calcule o coeficiente de variação para cada local. A conclusão é a mesma do item anterior? Qual é a medida de variabilidade mais adequada? Justifique.

O coeficiente de variação (CV) é obtido como a razão entre o desvio padrão e a média. Denote por s_1 e $\overline{x}_1$ o desvio padrão e a média do peso dos mexilhões observados no Sambaqui, e por s_2 e $\overline{x}_2$ o desvio padrão e a média do peso dos mexilhões observados no Manguezal. Com base na Tabela 1, temos que

CV para o local Sambaqui:

$$\text{CV}_1 = \frac{s_1}{\overline{x}_1} = \frac{7,97}{30,37} = 0,26243 \approx 26,24\%.$$

CV para o local Manguezal:

$$\text{CV}_2 = \frac{s_2}{\overline{x}_2} = \frac{6,32}{18,59} = 0,3400 = 34\%.$$

Logo, com base no coeficiente de variação podemos concluir que o local Manguezal é menos homogêneo (possui uma maior variação dos pesos em relação a sua média) quanto aos pesos dos mexilhões.

Note que a conclusão difere de quando usado o desvio padrão. No entanto, a medida de dispersão mais adequada para comparar dois (ou mais) grupos é o CV, pois este nos fornece uma medida de variabilidade relativa à média. Além disso, o CV elimina o efeito de magnitude dos dados e corresponde a uma medida adimensional.

(g) Vocł diria que o crescimento dos mexilhões esá associado ao local?

Com base nas estatísticas descritivas observadas (Tabela 1), nota-se que os mexilhões do Sambaqui possuem medidas de peso superiores às medidas do local Manguezal. Logo, as medidas sugerem que há uma relevante diferença nos pesos dos mexilhões dados os dois locais distintos. Apesar do local Manguezal possuir uma menor variabilidade, este forneceu mexilhões com peso de carne inferior ao local Sambaqui.

Exercício 3

Uma amostra de 150 alunos da USP foi selecionada e para cada estudante foi perguntado o número de vezes que ele prestou o vestibular da FUVEST (x). Observou-se que 75 estudantes prestaram o vestibular da FUVEST, uma única vez, 47 prestaram duas vezes e assim por diante. Os dados estão na tabela abaixo:

Número de vezes que prestou FUVEST (x)	n_i
1	75
2	47
3	21
4	7

(a) Qual é a variável de interesse do estudo? Classifique-a.

A variável de interesse (X) neste estudo corresponde ao número de vezes que o aluno entrevistado prestou o vestibular da FUVEST. Observe que temos uma contagem, logo podemos caracterizar a variável X como sendo uma variável quantitativa discreta.

(b) Calcule a média, a mediana, a moda, a variância e o desvio padrão do número de vezes que o estudante prestou o vestibular da FUVEST.

Média:

Como os valores observados se repetem, a média pode ser calculada da seguinte forma:

$$\overline{x} = \sum_{i=1}^{4} \frac{x_i \times n_i}{n} = \frac{1 \times 75 + 2 \times 47 + 3 \times 21 + 4 \times 7}{150} = \frac{260}{50} = 1,733 \text{ vezes},$$

ou seja, em média os alunos prestam o vestibular da FUVEST 1,73 vezes.

Mediana:

Para calcularmos a mediana neste caso, precisamos primeiro obter a posição da mediana que é dada por $0,5 \times (n+1) = 0,5 \times = 75,5$. Logo, a mediana será a média aritmética entre os valores das posições 75 e 76. Temos que $75 = 1$ e $76 = 2$. Portanto,

$$M_d = \frac{1+2}{2} = 1,5 \text{ vezes}.$$

Note que a mediana não divergiu tanto em relação a média que foi de 1,73 vezes.

Moda:

A moda equivale à observação que é mais frequente. Pela segunda coluna da tabela podemos observar que a observação que mais aparece na base de dados é o valor 1.

Variância:

Pela fórmula da variância temos

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Observe que neste caso em particular, fazemos 75 vezes a diferença $(1 - 1,733)^2$, 47 vezes a diferença $(2 - 1,733)^2$, e assim sucessivamente. Então, podemos reescrever a nossa expressão para o cálculo da variância amostral como sendo

$$s^2 = \frac{\sum_{i=1}^{4} n_i (x_i - \overline{x})^2}{n-1},$$

a qual nos fornece

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{4} n_i (x_i - \overline{x})^2}{n-1} \\ &= \frac{75 \times (1 - 1,733)^2 + 47 \times (2 - 1,733)^2 + 21 \times (3 - 1,733)^2 + 7 \times (4 - 1,733)^2}{150 - 1} \\ &= \frac{113,3333}{149} = 0,76063 \text{ (vez)}^2 \approx 0,76 \text{ (vez)}^2. \end{aligned}$$

Desvio padrão:

O desvio padrão será então

$$s = \sqrt{s^2} = \sqrt{0,76063} = 0,87214 \text{ vez} \approx 0,87 \text{ vezes}.$$

(c) Existe interesse em estudar o gasto dos alunos com as despesas do vestibular. Suponha, para simplificar, que cada aluno tem uma despesa fixa de R$ 5000,00, relativa à preparação e mais R$ 150,00 para cada vestibular prestado. Desta forma tem-se que a despesa é equacionada por d = 5000 + 150x. Calcule a média, a mediana, a moda, a variância e o desvio padrâo da despesa com vestibular da FUVEST.

Calculando as despesas para cada grupo teremos

No. de vezes	n_i	Despesa (d_i)
1	75	5000 + 150×1 = 5150
2	47	5000 + 150×2 = 5300
3	21	5000 + 150×3 = 5450
4	7	5000 + 150×4 = 5600
Σ	150	–

Média:

A média da despesa gasta com o vestibular da FUVEST é dada por

$$\begin{aligned}\overline{d} &= \sum_{i=1}^{4} \frac{n_i \times d_i}{150} \\ &= \frac{75 \times 5150 + 47 \times 5300 + 21 \times 5450 + 7 \times 5600}{150} \\ &= \frac{789000}{150} = 5260 \text{ R\$.}\end{aligned}$$

Mediana:

A posição da mediana é dada por $0,5 \times (n+1) = 0,5 \times (151) = 75,5$. Logo,

$$M_d(d) = \frac{5150 + 5300}{2} = 5225 \text{ R\$.}$$

Moda:

A moda de d é o valor da despesa mais frequente. Logo, a moda é 5150.

Variância:

Análogo ao item (b), a variância da despesa pode ser calculada como

$$
\begin{aligned}
s_d^2 =& \frac{\sum_{i=1}^{n} n_i(d_i - \overline{d})^2}{n-1} \\
=& \frac{75 \times (5150 - 5260)^2 + 47 \times (5300 - 5260)^2 + 21 \times (5450 - 5260)^2 + 7 \times (5600 - 5260)^2}{149} \\
=& \frac{2550000}{149} = 17114,09 \text{ (R\$)}^2.
\end{aligned}
$$

Desvio padrão:

O desvio padrão será então

$$s_d = \sqrt{s_d^2} = \sqrt{17114,09} = 130,82 \text{ R\$}.$$

Formas alternativas para o cálculo da média e do desvio padrão de uma variável y = a + bx:

A seguir apresentamos uma forma alternativa para o cálculo da média e do desvio padrão para o item (c). Pode-se mostrar que se $y = a + bx$, então $\overline{y} = a + b\overline{x}$. Além disso, $s_y^2 = b^2 s_x^2$, bem como $s_y = bs_x$.

Portanto, voltando ao nosso exemplo e sendo $d = 5000 + 150x$, teremos que $\overline{d} = a + b\overline{x}$, em que $a = 5000$ e $b = 150$. Logo,

$$\overline{d} = 5000 + 150 \times \overline{x} = 5000 + 150 \times 1,733 = 5259,95 \approx 5260.$$

E para a variância de d teremos:

$$s_d^2 = b2s_x^2 = 150^2 \times 0,76063 = 17114,17.$$

E o desvio padrão será dado por:

$$s_d = bs_x \Rightarrow s_d = 150 \times 0,87214 = 130,82.$$

Exerccio 4

O método do sequestro do radical livre estável 2,2-difenil-1-picrihidrazil (DPPH) é um método que consiste em avaliar a propriedade antioxidante de alimentos e outras substâncias. Quanto maior o valor do DPPH, maior é a proteção do sistema biológico contra a oxidação. Considere os dados a seguir sobre o valor de DPPH em 600 amostras de vinho tinto.

Classe de DPPH	Frequência	Frequência relativa
5,0 $\vdash$ 8,0	40	0,06
8,0 $\vdash$ 8,5	82	0,14
8,5 $\vdash$ 9,0	120	0,20
9,0 $\vdash$ 9,5	180	0,30
9,5 $\vdash$ 10,0	136	0,23
10,0 $\vdash$ 10,5	42	0,07

Construa os histogramas pelos métodos de frequência e densidade. Qual deve ser utilizado? Justifique.

Calculando a amplitude dos intervalos de classe (base) e a densidade de frequência

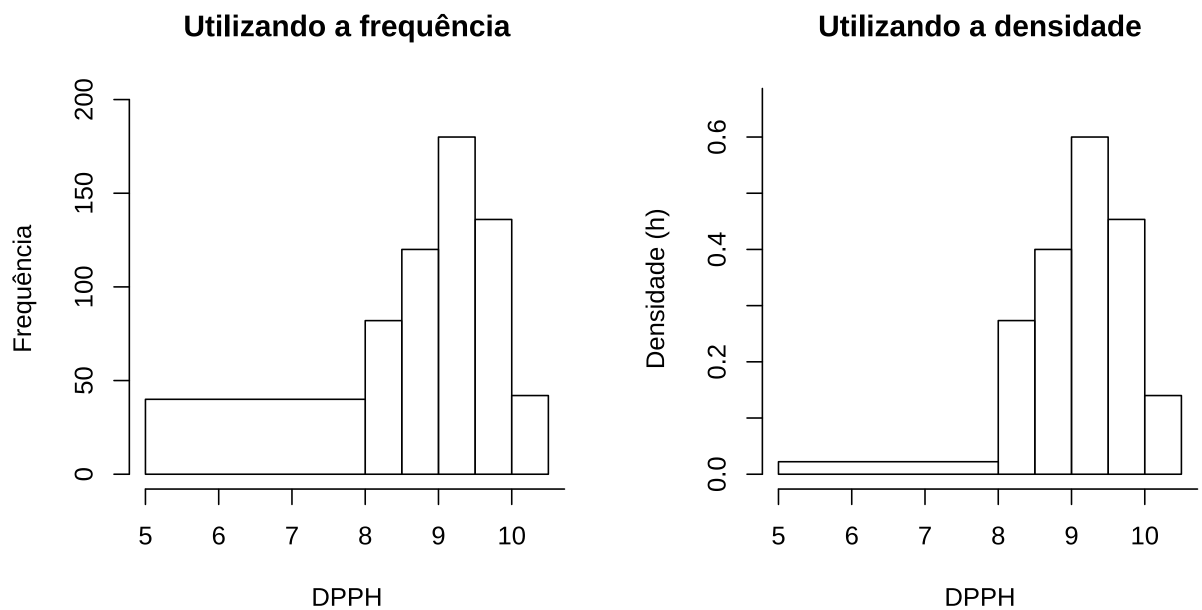
$$h = \frac{\text{frequência relativa}}{\text{base}},$$

temos

Classe de DPPH	Frequência	Frequência relativa	Base	h
5,0 $\vdash$ 8,0	40	0,06	3	0,02
8,0 $\vdash$ 8,5	82	0,14	0,5	0,28
8,5 $\vdash$ 9,0	120	0,20	0,5	0,4
9,0 $\vdash$ 9,5	180	0,30	0,5	0,6
9,5 $\vdash$ 10,0	136	0,23	0,5	0,46
10,0 $\vdash$ 10,5	42	0,07	0,5	0,14

Os histogramas construídos pelo método da frequência e pelo método da densidade são apresentados a seguir.

Como as amplitudes das classes são de tamanhos diferentes, utilizar o histograma construído pelo método de densidade.



Exercício 5.

O arquivo **CopaDoMundo.xls** contém alguns dados das Copas do Mundo de Futebol de 2014 e 2018. As variáveis são: `Gols2014`: número de gols no tempo regulamentar por partida na Copa do Mundo de 2014; `Gols2018`: número de gols no tempo regulamentar por partida na Copa do Mundo de 2018; `Saldo2014`: saldo de gols por partida, incluindo a prorrogação quando houver, na Copa do Mundo de 2014; `Saldo2018`: saldo de gols por partida, incluindo a prorrogação quando houver, na Copa do Mundo de 2018.

Resolva os itens a seguir, com o uso de recurso computacional.

(a) Obtenha média, mediana, variância, desvio padrão e coeficiente de variação para o número de gols no tempo regulamentar em 2014 e 2018

A seguir apresentamos os passos utilizados no Rcmdr.

(1) Ler o conjunto de dados: Clique na guia "Dados" $\rightarrow$ "Importar arquivos de dados" $\rightarrow$ "do Excel". Em seguida defina um nome para o seu conjunto de dados (por exemplo, copa) e clique em "OK". Navegue até o diretório onde o seu conjunto de dados está localizado e clique em "Abrir". Selecione a aba do Excel onde está localizado o conjunto de dados (no nosso caso, "Dados") e clique em "OK".

(2) Cálculo das medidas resumo: Clique na guia "Estatísticas" $\rightarrow$ "Resumos" $\rightarrow$ "Resumos numéricos". Na aba "Dados" selecione as variáveis "Gols2014" e "Gols2018" na lista de variáveis e em seguida na aba “Estatísticas” selecione as estatísticas de interesse, que no nosso caso são: média, desvio padrão, coeficiente de variação e o quartil 0,50 (digite 0.50 em “Quantis”). Clique em “OK”.

O resultado é apresentado a seguir.

	mean	sd	cv	50%	n
Gols2014	2,671875	1,764326	0,660333	3	64
Gols2018	2,59375	1,570752	0,605591	2	64

As variâncias são obtidas pelo quadrado dos desvios-padrão. Na janela “RScript” do Rcmdr digite o seguinte comando e depois submeta (ou Ctrl+R):

```
1.764326^2
```

O resultado estará na janela “Output”.

```
> 1.764326^2
[1] 3.112846
```

Repetindo para o desvio padrão de “Gols2018”, temos:

```
> 1.570752^2
[1] 2.467262
```

A tabela com todas as estatísticas solicitadas é apresentada a seguir.

	mean	sd	var	cv	50%	n
Gols2014	2,671875	1,764326	3,112846	0,660333	3	64
Gols2018	2,59375	1,570752	2,467262	0,605591	2	64

(b) Obtenha média, mediana, variância, desvio padrão e coeficiente de variação para o saldo de gols incluindo a prorrogação em 2014 e 2018;

Repetindo o passo 2 do item anterior para as variáveis “Saldo2014” e “Saldo2018”, temos o seguinte resultado:

	mean	sd	cv	50%	n
Saldo2014	1,421875	1,231913	0,8664	1	64
Saldo2018	1,359375	1,04452	0,768383	1	64

Repetindo o procedimento descrito anteriormente para o cálculo das variâncias temos:

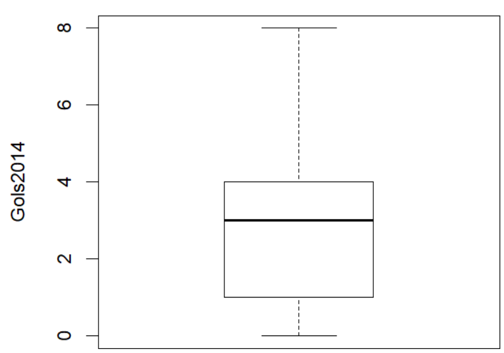
```
> 1.231913^2
[1] 1.51761

> 1.044520^2
[1] 1.091022
```

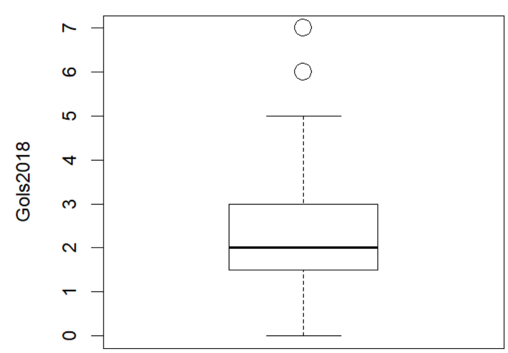
A tabela com todas as estatísticas solicitadas é apresentada a seguir.

	mean	sd	var	cv	50%	n
Saldo2014	1,421875	1,231913	1,51761	0,8664	1	64
Saldo2018	1,359375	1,04452	1,091022	0,768383	1	64

(c) Construa o boxplot do número de gols no tempo regulamentar para as duas copas. Comente sobre a dispersão, pontos extremos, mediana e simetria dos dados, em cada copa

Para construir o Boxplot no Rcmdr vá na guia "Gráficos" → "Boxplot. Na aba "Dados" selecione a variável "Gols2014" na lista de variáveis e clique em "OK". O gráfico será exibido em uma nova janela. Para copiá-lo, clique com o boão direito sobre o gráfico e escolha a opção "Copiar como Bitmap". Em seguida cole no documento da sua lista de exercícios. Repita o mesmo procedimento para a variável "Gols2018". Os gráficos construídos são apresentados a seguir.

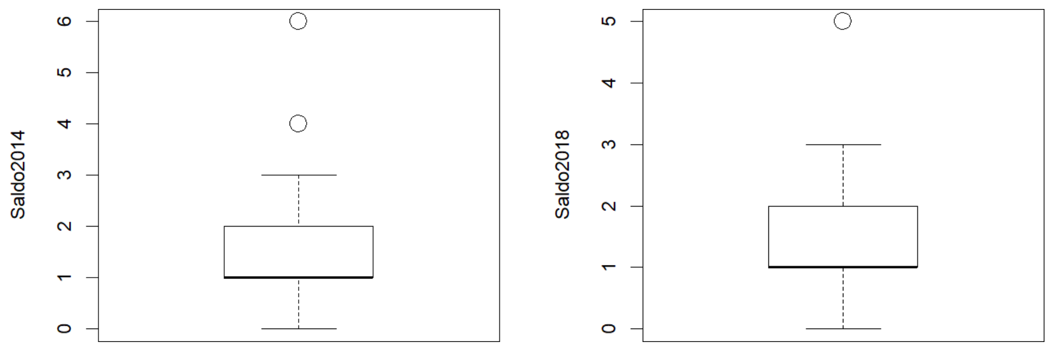




O número mediano de gols em 2014 é maior do que o número mediano de gols em 2018 (3 x 2). A dispersão do número de gols em 2014 é maior do que a do ano de 2018. Isso pode ser visto pelo tamanho das caixas dos boxplots: a caixa do boxplot do ano de 2014 é maior do que a do ano de 2018. O ano de 2018 ainda apresentou 9 partidas com números de gols atípicos (6 ou 7). Observamos também que as distribuições do número de gols em 2014 e em 2018 são assimétricas.

(d) Construa o boxplot do saldo de gols incluindo a prorrogação para as duas copas. Comente sobre a dispersão, pontos extremos, mediana e simetria dos dados, em cada copa

Repetindo o mesmo procedimento descrito no item anterior temos os seguintes gráficos.



As distribuições dos saldos de gols em 2014 e em 2018 são bem parecidas. Nos dois anos o saldo de gols mediano foi igual a 1. Os tamanhos das caixas dos boxplots são próximos indicando que a dispersão de ambas as distribuições são parecidas. Em ambos os anos tivemos saldos de gols atípicos. No ano de 2014, tivemos 4 partidas com saldos de gols atípicos e iguais ou maiores que 4. Em 2018 tivemos duas partidas com saldos de gols atípicos e iguais a 5. As duas distribuições são assimétricas.

Exercício 6.

O artigo “Inquérito Brasileiro de Diálise Crônica 2016”, de autoria de R.C.Sesso, A.A.Lopes, F.S.Thomé, J.R.Lugon e C.T. Martins, publicado no Jornal Brasileiro de Nefrologia, em 2017, mostra a distribuição de 50.807 pacientes com doena renal crônica, em relação ao tipo de diálise a que são submetidos e a fonte pagadora do tratamento (Tabela 3).

Tabela 2: Distribuição de pacientes conforme o Tipo de diálise e Fonte pagadora, Censo de 2016.

	Fonte pagadora		
Tipo de diálise	SUS	Não SUS	Total
HD convencional	38.437	7.279	45.716
HD diária	359	708	1.067
CAPD	933	140	1.073
DPA	2.281	637	2.918
DPI	25	8	33
Total	42.035	8.772	50.807

HD = hemodiálise; CAPD = diálise peritonial ambulatorial contínua; DPA = diálise peritonial automatizada; DPI = diálise peritonial intermitente; SUS = Sistema Único de Saúde.

(a) Identifique as variáveis do estudo e classifique-as

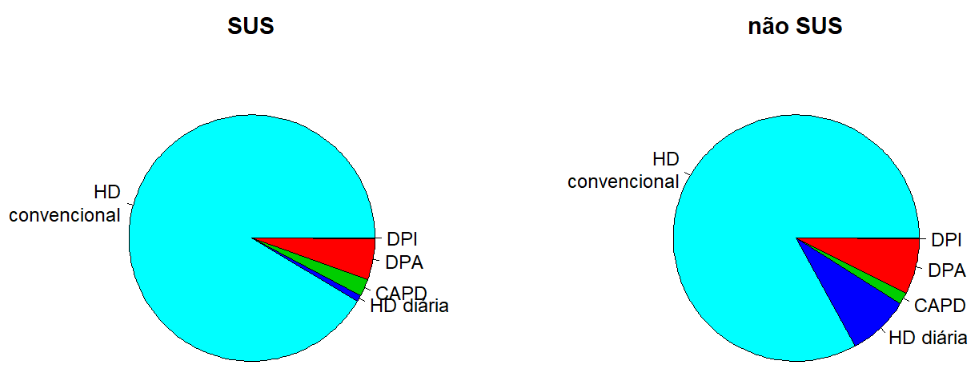
As variáveis do estudo são:

tipo de diálise: qualitativa nominal;

fonte pagadora: qualitativa nominal.

(b) Com o auxílio de recurso computacional, faça um gráfico de setores para os pacientes SUS com a distribuição do tipo de diálise. Faça o mesmo gráfico para os pacientes No SUS. Você diria que as distribuições do tipo de diálise são diferentes? Comente.

Siga o Passo (1) do item (a) do exercício 5 para ler o conjunto de dados **dialise.xlsx** no Rcmdr. Para construir o gráfico de setores para a fonte pagadora SUS vá na guia “Gráficos” → “Gráfico de Pizza”. Na aba “Dados” selecione a variável “SUS” na lista de variáveis e clique em “OK”. O gráfico será exibido em uma nova janela. Para copiá-lo, clique com o botão direito sobre o gráfico e escolha a opção “Copiar como Bitmap”. Em seguida cole no documento da sua lista de exercícios. Repita o mesmo procedimento para a variável “não SUS” da fonte pagadora não SUS. Os gráficos de setores construídos para cada fonte pagadora de acordo com o tipo de diálise são apresentados a seguir.



Para ambas as fontes pagadoras o tipo de diálise predominante é a HD convencional (91,44% x 82,48%), enquanto que o tipo de diálise DPI é o menos frequente (0,06% x 0,09%). O tipo de diálise HD diária é mais frequente na fonte pagadora não SUS do que na fonte pagadora SUS (8,07% x 0,85%). Já o tipo de diálise CAPD é ligeiramente mais frequente na fonte pagadora SUS (2,22% x 1,60%). E o contrário ocorre com o tipo de diálise DPA que é mais frequente na fonte pagadora não SUS (7,26% x 5,43%). Desse modo, podemos dizer que as distribuições dos tipos de diálise são diferentes.