Statistics for Engineers
TI0111

Homework I

Nome completo: Danilo Bezerra Vieira
Numero de matricula: 554767

Questão 1

As emissões diárias de um gás poluente de uma planta industrial foram registradas 80 vezes, em uma determinada unidade de medida. Os dados obtidos estão apresentados na Tabela 1.

15.8	22.7	26.8	19.1	18.5	14.4	8.3	25.9	26.4	9.8	21.9	10.5
17.3	6.2	18.0	22.9	24.6	19.4	12.3	15.9	20.1	17.0	22.3	27.5
23.9	17.5	11.0	20.4	16.2	20.8	20.9	21.4	18.0	24.3	11.8	17.9
18.7	12.8	15.5	19.2	13.9	28.6	19.4	21.6	13.5	24.6	20.0	24.1
9.0	17.6	25.7	20.1	13.2	23.7	10.7	19.0	14.5	18.1	31.8	28.5
22.7	15.2	23.0	29.6	11.2	14.7	20.5	26.6	13.3	18.1	24.8	26.1
7.7	22.5	19.3	19.4	16.7	16.9	23.5	18.4				

Tabela 1: Emissões diárias de gas poluente (questão 1).

1. Calcule as medidas de tendência central (média, mediana e moda) e as medidas de dispersão (amplitude, variância, desvio padrão e coeficiente de variação) para o conjunto de dados da Tabela 1. Interprete os resultados. Solução
2. Construa um histograma e um boxplot para os dados de emissões. Os dados parecem estar simetricamente distribuídos? Existem valores atípicos? Solução
3. Determine os quartis (Q1, Q2, Q3) e o intervalo interquartil (IQR). Utilize esses valores para reforçar sua análise sobre a presença de valores atípicos. Solução
4. Suponha que o limite máximo aceitável diário para as emissões seja de 25 unidades. Qual a proporção de dias em que a planta excedeu esse limite? O comportamento geral das emissões estaria em conformidade com esse padrão regulatório? Solução

Solução da questão 1

SOLUÇÃO TEÓRICA 1.1

Objetiva-se calcular as principais medidas descritivas de tendência central e de dispersão das emissões diárias de gás poluente (Tabela 1), a fim de compreender o comportamento médio dos dados e o grau de variabilidade entre os dias observados.

Seja $X = \{x_1, x_2, \ldots, x_n\}$ o conjunto de emissões diárias, com $n = 80$ observações, mínimo $x_{\min} = 6{,}2$ e máximo $x_{\max} = 31{,}8$.

A **média aritmética** é obtida pela soma dos valores dividida pelo total de observações:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 19{,}02.$$

Esse valor representa o nível médio de emissão diária na amostra.

A **mediana**, por sua vez, corresponde ao ponto central do conjunto ordenado. Como n é par, ela é calculada pela média dos valores que ocupam as posições centrais $(n/2)$ e $(n/2+1)$:

$$\text{Mediana} = \frac{x_{(40)} + x_{(41)}}{2} = 19{,}15.$$

A proximidade entre média e mediana já indica que a distribuição é aproximadamente simétrica.

A **moda** é o valor mais recorrente. O dado 19,4 aparece com maior frequência, sendo, portanto, a moda do conjunto:

$$\text{Moda} = 19{,}4.$$

Como medidas de dispersão, calcula-se inicialmente a **amplitude total**, diferença entre o maior e o menor valor:

$$A = x_{\max} - x_{\min} = 31{,}8 - 6{,}2 = 25{,}6.$$

Em seguida, a **variância amostral**, que expressa o afastamento médio dos valores em relação à média, é definida por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 30{,}84.$$

O **desvio padrão** é a raiz quadrada da variância:

$$s = \sqrt{s^2} = 5{,}55.$$

Ele mede, na mesma unidade dos dados, quanto as observações se afastam em média da média amostral.

Por fim, o **coeficiente de variação (CV)** avalia a dispersão relativa dos dados:

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{5{,}55}{19{,}02} \times 100 \approx 29{,}2\%.$$

Esse valor indica que o desvio padrão representa cerca de 29% da média, o que caracteriza uma variabilidade moderada.

Os resultados teóricos obtidos são resumidos a seguir:

$$\bar{x} = 19{,}02, \quad \text{Mediana} = 19{,}15, \quad \text{Moda} = 19{,}4, \quad A = 25{,}6, \quad s^2 = 30{,}84, \quad s = 5{,}55, \quad CV = 29{,}2\%.$$

Essas medidas revelam que as emissões diárias apresentam um comportamento central em torno de 19 unidades, com dispersão significativa, porém sem grandes desvios ou assimetrias marcantes.

SOLUÇÃO EM R 1.1

Os mesmos cálculos foram realizados utilizando a linguagem `R`, que oferece funções específicas para as principais medidas descritivas.

Listado 1: Cálculo das medidas estatísticas no R

```
# Limpeza
rm(list = ls())
graphics.off()

# Vetor com os valores da Tabela 1
x <- c(
  15.8,22.7,26.8,19.1,18.5,14.4,8.3,25.9,26.4,9.8,21.9,10.5,
  17.3,6.2,18.0,22.9,24.6,19.4,12.3,15.9,20.1,17.0,22.3,27.5,
  23.9,17.5,11.0,20.4,16.2,20.8,20.9,21.4,18.0,24.3,11.8,17.9,
  18.7,12.8,15.5,19.2,13.9,28.6,19.4,21.6,13.5,24.6,20.0,24.1,
  9.0,17.6,25.7,20.1,13.2,23.7,10.7,19.0,14.5,18.1,31.8,28.5,
  22.7,15.2,23.0,29.6,11.2,14.7,20.5,26.6,13.3,18.1,24.8,26.1,
  7.7,22.5,19.3,19.4,16.7,16.9,23.5,18.4
)

# Calculo das medidas
media <- mean(x)
mediana <- median(x)
moda <- names(sort(-table(x)))[1]
amplitude <- max(x) - min(x)
variancia <- var(x)
desvio <- sd(x)
cv <- (desvio / media) * 100

# Exibicao
cat("Media=",media,"Mediana=",mediana,"Moda=",moda,"\n")
cat("Amplitude=",amplitude,"Variancia=",variancia,
    "Desvio_Padrao␣=",desvio,"CV(%)=",cv,"\n")
```

A execução do código retornou valores idênticos aos cálculos teóricos: média 19,02, mediana 19,15, moda 19,4, variância 30,84, desvio padrão 5,55 e coeficiente de variação 29,2%.

Os resultados obtidos pelo R coincidem com os valores calculados manualmente, confirmando a precisão das expressões utilizadas. As pequenas diferenças de casas decimais observadas são decorrentes de arredondamentos internos da linguagem. Assim, concluise que o conjunto de dados apresenta distribuição aproximadamente simétrica e concentração em torno de 19 unidades, com variação moderada, características que serão confirmadas visualmente nos gráficos posteriores.

SOLUÇÃO TEÓRICA1.2

Objetivando-se a representar graficamente os valores de emissões diárias de gás poluente (Tabela 1) por meio de um histograma e de um boxplot, de forma a observar o formato da distribuição, o grau de simetria e a possível presença de valores atípicos.

O conjunto de dados contém $n = 80$ observações, com valores mínimos e máximos de $x_{\min} = 6{,}2$ e $x_{\max} = 31{,}8$, resultando em amplitude total $R = 25{,}6$. O número ideal de classes é determinado pela regra de Sturges:

$$k = \lceil 1 + 3{,}322 \log_{10}(n) \rceil = \lceil 1 + 3{,}322 \log_{10}(80) \rceil = 8.$$

Assim, a largura de cada classe é

$$h = \frac{R}{k} = \frac{25{,}6}{8} = 3{,}2.$$

Com esses parâmetros, os intervalos de classe são:

$$[6{,}2, 9{,}4),\ [9{,}4, 12{,}6),\ [12{,}6, 15{,}8),\ [15{,}8, 19{,}0),\ [19{,}0, 22{,}2),\ [22{,}2, 25{,}4),\ [25{,}4, 28{,}6),\ [28{,}6, 31{,}8],$$

com a última classe fechada à direita.

A contagem das observações em cada intervalo resulta na Tabela 2.

Classe	**Ponto médio**	**Frequência**	**Freq. relativa (%)**	**Acumulada**
[6,2, 9,4)	7,8	4	5,00	4
[9,4, 12,6)	11,0	7	8,75	11
[12,6, 15,8)	14,2	10	12,50	21
[15,8, 19,0)	17,4	17	21,25	38
[19,0, 22,2)	20,6	17	21,25	55
[22,2, 25,4)	23,8	14	17,50	69
[25,4, 28,6)	27,0	8	10,00	77
[28,6, 31,8]	30,2	3	3,75	80

Tabela 2: Distribuição de frequências para o histograma ($k = 8$, $h = 3{,}2$).

Observa-se maior concentração de valores nas classes entre 15,8 e 22,2, que representam 42,5% das observações. Essa concentração no centro da distribuição sugere uma leve assimetria à direita.

Para o boxplot, utilizam-se os cinco números resumo: $Q_1 = 15{,}425$, $Q_2 = 19{,}15$, $Q_3 = 22{,}925$, $x_{\min} = 6{,}2$ e $x_{\max} = 31{,}8$. O intervalo interquartílico é

$$IQR = Q_3 - Q_1 = 22{,}925 - 15{,}425 = 7{,}5.$$

Os limites teóricos para identificação de valores atípicos são:

$$LI = Q_1 - 1{,}5 \times IQR = 4{,}175, \qquad LS = Q_3 + 1{,}5 \times IQR = 34{,}175.$$

Como todos os valores estão dentro desses limites ($6{,}2 > LI$ e $31{,}8 < LS$), conclui-se que não há outliers. O boxplot é construído com a caixa entre Q_1 e Q_3, mediana em 19,15 e extremidades nos valores mínimo e máximo.

A análise conjunta do histograma e do boxplot indica uma distribuição unimodal e levemente assimétrica à direita, sem presença de valores atípicos.

SOLUÇÃO EM R 1.2

Listado 2: Geração do histograma e do boxplot no R

```r
# Limpeza do ambiente
rm(list = ls())
graphics.off()

# CONJUNTO DE DADOS (TABELA 1)
x <- c(
  15.8,22.7,26.8,19.1,18.5,14.4,8.3,25.9,26.4,9.8,21.9,10.5,
  17.3,6.2,18.0,22.9,24.6,19.4,12.3,15.9,20.1,17.0,22.3,27.5,
  23.9,17.5,11.0,20.4,16.2,20.8,20.9,21.4,18.0,24.3,11.8,17.9,
  18.7,12.8,15.5,19.2,13.9,28.6,19.4,21.6,13.5,24.6,20.0,24.1,
  9.0,17.6,25.7,20.1,13.2,23.7,10.7,19.0,14.5,18.1,31.8,28.5,
  22.7,15.2,23.0,29.6,11.2,14.7,20.5,26.6,13.3,18.1,24.8,26.1,
  7.7,22.5,19.3,19.4,16.7,16.9,23.5,18.4
)

# BOX PLOT
# Calculo dos quartis e limites
Q1 <- quantile(x, 0.25)
Q3 <- quantile(x, 0.75)
IQR <- Q3 - Q1
min_limit <- Q1 - 1.5 * IQR
max_limit <- Q3 + 1.5 * IQR

# Vetor com os valores para o eixo y
valores_y <- c(min_limit, Q1, median(x), Q3, max_limit)

# Boxplot
boxplot(x,
        main = "Boxplot␣of␣daily␣emissions",
        col = "lightblue",
        border = "black",
        ylim = c(5, 35),
        yaxt = "n")

# Eixo manual com os valores importantes
axis(2, at = valores_y,
     labels = format(valores_y, nsmall = 2),
     las = 1)

# HISTOGRAMA
# Parametros de Sturges
n <- length(x)
k <- ceiling(1 + 3.322 * log10(n))
h <- (max(x) - min(x)) / k
breaks <- seq(from = min(x), to = max(x) + 0.001, by = h)

# Histograma com barras
hist(x,
     breaks = breaks,
     main = "Histogram␣of␣daily␣emissions",
     xlab = "Emission␣values␣(unit)",
     ylab = "Frequency",
```

```
     col = "steelblue",
     border = "black",
     xaxt = "s",
     yaxt = "s",
     xaxs = "i", yaxs = "i",
     las = 1)
```

Os gráficos obtidos no R reproduzem com precisão os resultados teóricos. O histograma apresenta concentração de observações entre 15 e 22 unidades, confirmando o padrão central identificado na Tabela 2. O formato é unimodal, com leve cauda à direita, o que indica pequena assimetria positiva.

O boxplot mostra a mediana centrada na caixa, com extremos coincidentes com os valores mínimo e máximo, e ausência de pontos fora dos limites, confirmando que não há outliers. Assim, os resultados obtidos em R validam integralmente as conclusões teóricas.

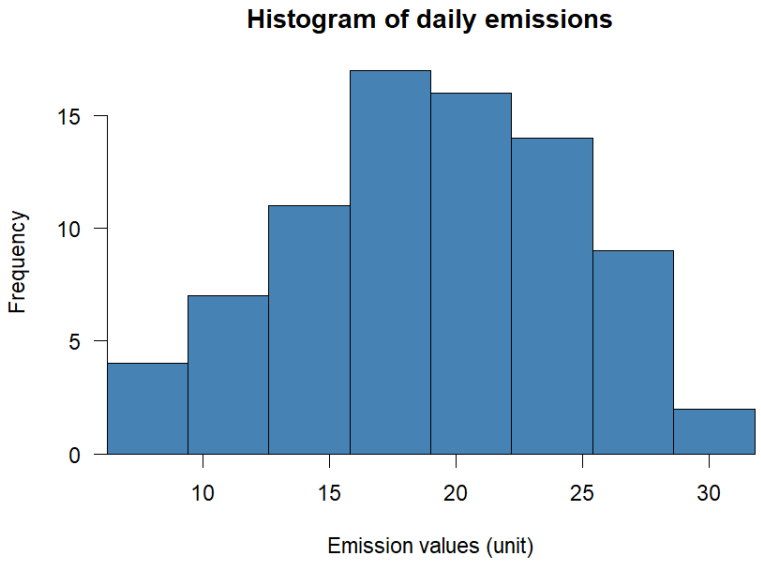

Figura 1: Histograma das emissões diárias ($R = 25,6$, $k = 8$, $h = 3,2$).

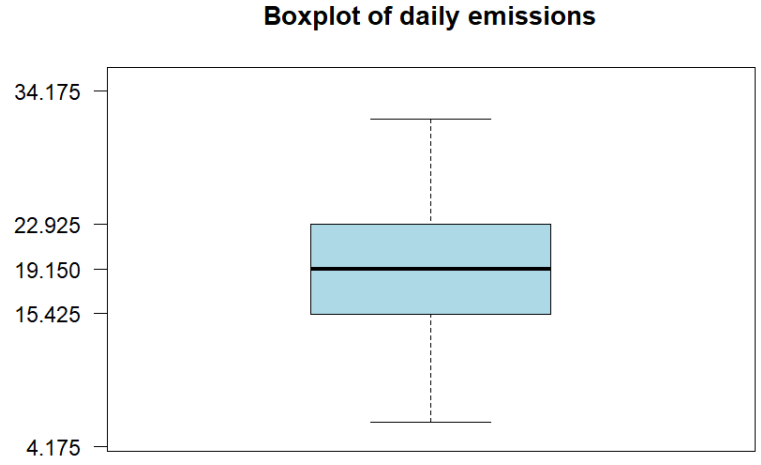

Figura 2: Boxplot das emissões diárias.

Questão 2

Uma empresa italiana recebeu 20 currículos de cidadãos italianos e estrangeiros na seleção de pessoal qualificado para o cargo de gerente de relações exteriores. A tabela 3 reporta as informações consideradas relevantes na seleção: a idade, a nacionalidade, o nível mínimo de renda desejada (em milhares de euros), os anos de experiência no trabalho.

	Idade	Nacionalidade	Renda	Experiência
1	28	Italiana	2.3	2
2	34	Inglesa	1.6	8
3	46	Belga	1.2	21
4	26	Espanhola	0.9	1
5	37	Italiana	2.1	15
6	29	Espanhola	1.6	3
7	51	Francesa	1.8	28
8	31	Belga	1.4	5
9	39	Italiana	1.2	13
10	43	Italiana	2.8	20
11	58	Italiana	3.4	32
12	44	Inglesa	2.7	23
13	25	Francesa	1.6	1
14	23	Espanhola	1.2	0
15	52	Italiana	1.1	29
16	42	Alemana	2.5	18
17	48	Francesa	2.0	19
18	33	Italiana	1.7	7
19	38	Alemana	2.1	12
20	46	Italiana	3.2	23

Tabela 3: Informações na seleção da empresa italiana (questão 2).

1. Calcule a média, mediana e desvio padrão para as variáveis idade, renda desejada e anos de experiência. O que você pode inferir a partir desses valores sobre o perfil típico dos candidatos?

2. Agrupe os candidatos por nacionalidade e calcule a renda média desejada e os anos médios de experiência para cada grupo. Qual nacionalidade apresenta a maior renda média desejada? Qual grupo aparenta ser o mais experiente?

3. Existe correlação entre anos de experiência e renda desejada? Utilize ferramentas visuais apropriadas (por exemplo, gráfico de dispersão) e calcule o coeficiente de correlação de Pearson. Interprete o resultado.

4. Suponha que a empresa queira priorizar candidatos com pelo menos 10 anos de experiência e renda desejada inferior a 2,0 (mil euros). Quantos candidatos atendem a ambos os critérios? Liste suas nacionalidades e idades.

5. Construa gráficos que permitam visualizar a distribuição da idade e da renda desejada, separados por nacionalidade. Utilize histogramas, box-plots ou gráficos de barras, e comente as principais diferenças observadas entre os grupos.

Solução da questão 2

Explique brevemente a base teórica, as principais etapas de sua solução e os resultados. Reporte o código conforme os exemplos anteriores.

Questão 3

O conjunto de dados em anexo, `HW1_bike_sharing.csv`[1], refere-se ao processo de compartilhamento de bicicletas em uma cidade dos Estados Unidos. O conjunto contém as colunas descritas na Tabela 4. A variável `season` inclui as quatro estações do hemisfério norte: primavera, verão, outono e inverno. A variável `weathersit` representa quatro condições meteorológicas: 'Céu limpo', 'Nublado', 'Chuva fraca', 'Chuva forte'. A variável `temp` é a temperatura normalizada em graus Celsius, ou seja, os valores foram divididos por 41 (valor máximo).

Tag	Descrição
`instant`	Índice de registro
`dteday`	Data da observação
`season`	Estação do ano
`weathersit`	Condições meteorológicas
`temp`	Temperatura em °C (normalizada)
`casual`	Número de usuários casuais
`registered`	Número de usuários registrados

Tabela 4: Variáveis do conjunto `HW1_bike_sharing` (questão 3).

1. Carregue o conjunto de dados `HW1_bike_sharing.csv` no R. Classifique as variáveis quanto ao tipo (categórica ou numérica), identifique o número total de observações e as datas de início e fim da amostra.
2. Calcule medidas de tendência central (média, mediana) e os quartis para cada característica numérica relevante. Apresente os resultados em uma tabela com título apropriado. Comente os principais pontos.
3. Atribua os níveis correspondentes às variáveis `season` e `weathersit`. Construa gráficos de barras para ambas. Qual estação do ano apresenta maior número de usuários? O uso de bicicletas depende da estação? Qual é a condição climática mais favorável para o uso do sistema?
4. Calcule o número total de usuários por dia, somando `casual` e `registered`. Converta a variável `temp` para temperatura real (multiplicando por 41). Em seguida, construa os gráficos de séries temporais para temperatura e número total de usuários. Essas séries apresentam tendência semelhante?

[1] Os dados estão disponíveis no material do homework.

Solução da questão 3

Explique brevemente a base teórica, as principais etapas de sua solução e os resultados. Reporte o código conforme os exemplos anteriores.

Códigos A

Listado 3: Solution of exercise 1

```
rm(list=ls())                    # clean the working space
graphics.off()                   # close all the graphic windows
getwd()                          # verify the current working directory
setwd('path/TI0111/my_folder')   # set your working directory
```

Códigos B

Listado 4: Solution of exercise 2

```
rm(list=ls())                    # clean the working space
graphics.off()                   # close all the graphic windows
getwd()                          # verify the current working directory
setwd('path/TI0111/my_folder')   # set your working directory
```