Statistics for Engineers
TI0111

Homework I

Nome completo: Savlio Carvalho Pontes
Numero de matricula: 567715
Nome completo: Thaís Sousa Barros Leal
Numero de matricula: 565548

Base Teórica

A base teórica apresentada a seguir é fundamental para a compreensão e execução de todas as análises estatísticas deste trabalho. Ela abrange as principais medidas de tendência central, dispersão e posição, além de conceitos gráficos e procedimentos de conformidade usados ao longo das questões.

Medidas de Tendência Central

- **Média ($\bar{x}$):** Soma de todos os valores dividida pelo número total de observações (n).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Mediana (Md):** Valor central de um conjunto de dados ordenado.
  - Se n é ímpar: posição $(n+1)/2$.
  - Se n é par: média dos dois valores centrais.
- **Moda (Mo):** Valor que aparece com maior frequência no conjunto de dados.

Medidas de Dispersão

- **Amplitude (A):** Diferença entre o maior e o menor valor do conjunto.

$$A = x_{\max} - x_{\min}$$

- **Variância (s^2):** Mede a dispersão dos dados em relação à média.

  Fórmula teórica (populacional):

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

  Fórmula amostral:

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

- **Desvio Padrão (s):** Raiz quadrada da variância.

$$s = \sqrt{s^2}$$

- **Coeficiente de Variação (CV):** Mede a variabilidade relativa dos dados.

$$CV = \frac{s}{\bar{x}} \times 100\%$$

- **Coeficiente de Correlação de Pearson (r):** Mede o grau de relação linear entre duas variáveis quantitativas. O valor de r varia de -1 a 1, onde:
  - $r = 1$: correlação linear positiva perfeita;
  - $r = -1$: correlação linear negativa perfeita;
  - $r = 0$: ausência de correlação linear.

  A fórmula é dada por:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Quartis e Outliers

- **Quartis:** Dividem o conjunto ordenado em quatro partes iguais.
  - **Fórmula geral:**

$$Q_k = \left(\frac{k(n+1)}{4}\right)\text{-ésimo valor da amostra}, \quad k = 1, 2, 3$$

  - $Q1$: 25% dos dados estão abaixo (primeiro quartil).
  - $Q2$: 50% dos dados estão abaixo (mediana).

  - $Q3$: 75% dos dados estão abaixo (terceiro quartil).

- **Intervalo Interquartil (IQR):** Diferença entre $Q3$ e $Q1$.

$$IQR = Q3 - Q1$$

- **Outliers:** Valores que ultrapassam os limites:

$$\text{Limite Inferior} = Q1 - 1.5 \times IQR \quad \text{e} \quad \text{Limite Superior} = Q3 + 1.5 \times IQR$$

Representações Gráficas

- **Histograma:** Mostra a distribuição dos dados em intervalos (classes), permitindo observar a forma da distribuição (simétrica, assimétrica etc.).

- **Boxplot:** Representa a variação dos dados com base nos quartis, destacando a mediana, a dispersão e os outliers.

Classificação dos tipos de variáveis

- **Categórica:** representa categorias determinadas por características que não podem ser medidas numericamente. Exemplos: cor dos olhos, doenças, etc.

- **Numéricas:** são mensuráveis em números que podem ser manipulados por operações matemáticas.

1 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.

Tabela 1: Emissões diárias de gás poluente.

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				

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.

2. Construa um histograma e um boxplot para os dados de emissões. Os dados parecem estar simetricamente distribuídos? Existem valores atípicos?

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.

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 da Questão 1

1.1) Para o cálculo das medidas de tendência central e dispersão, utilizamos as fórmulas apresentadas na Base Teórica, bem como o software R para visualização e manipulação dos dados.

```
emissions <- scan(); # Permite ler todos os dados agrupados
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
```

Listado 1: Leitura dos dados da Tabela

```
# --- ITEM 1 --- #

# Medidas de tendência central:
mean_emissions <- mean(emissions); # Calcula a média das emissões
median_emissions <- median(emissions); # Calcula a mediana
# Função para moda:
get_mode <- function(x){
  ux <- unique(x); # Vetor contendo cada elemento único do vetor original
  count <- tabulate(match(x,ux)); # Conta quantas vezes cada elemento ocorre
  if(all(count == 1)){
    return(NA); # Se nenhum elemento se repete, não há moda
  }else{
    return(ux[which.max(count)]); # Se há repetições, obtém o elemento mais
    frequente
  }
}
mode_emissions <- get_mode(emissions); # Retorna a moda

# Medidas de dispersão:
amplitude_emissions <- diff(range(emissions));
# Calcula a amplitude (diferença entre maior e menor valor)

variance_emissions <- var(emissions); # Calcula a variância
stDev_emissions <- sqrt(variance_emissions); # Calcula o desvio padrão
coefVar_emissions <- (stDev_emissions/mean_emissions) * 100;
# Calcula o coeficiente de variação

# Tabela para medidas de tendência central
```

```r
tabela_tendencia <- data.frame(
  Medida = c("Média", "Mediana", "Moda"),
  Valor = c(
    mean_emissions,
    median_emissions,
  )
)

# Tabela para medidas de dispersão
tabela_dispersao <- data.frame(
  Medida = c("Amplitude", "Variância", "Desvio Padrão", "Coeficiente de Variação")
    ,
  Valor = c(
    amplitude_emissions,
    variance_emissions,
    stDev_emissions,
    paste(coefVar_emissions, "%")
  )
)

# Exibir as tabelas
cat("Medidas de Tendência Central:")
tabela_tendencia

cat("Medidas de Dispersão:")
tabela_dispersao
```

Listado 2: Cálculo das medidas de tendência central e de dispersão

```
Medidas de Tendência Central:
   Medida    Valor
1   Média 19.02125
2 Mediana 19.15000
3    Moda 19.40000
>

Medidas de Dispersão:
                   Medida              Valor
1               Amplitude               25.6
2               Variância   30.8414414556962
3           Desvio Padrão   5.55350713114661
4 Coeficiente de Variação 29.1963311093993 %
```

Listado 3: Saída do código no console com os dados calculados

Cálculos teóricos e valores obtidos no R Tabela de Dados Ordenados

Os dados de emissões foram organizados em ordem crescente, conforme mostrado abaixo.

Tabela 2: Tabela de Dados Ordenados

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

- **Número de observações (n):** Indica a quantidade total de medições realizadas, neste caso, 80 dias de registro das emissões.

$$n = 80 \quad \text{(No R: 80)}$$

- **Somatório das emissões ($\sum x_i$):** Representa a soma de todos os valores medidos ao longo do período de observação, totalizando 1521,7 unidades.

$$\sum x_i = 1521{,}7 \quad \text{(No R: 1521,7)}$$

- **Média ($\bar{x}$):** Indica o valor médio diário das emissões. O valor de 19,02 mostra que, em média, as emissões se mantêm abaixo do limite de 25 ppm.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1521{,}7}{80} = 19{,}021 \quad \text{(No R: 19,02125)}$$

- **Mediana (Md):** Representa o ponto central da distribuição — metade dos dias apresentou emissões abaixo de 19,15 e metade acima. Isso reforça que os dados estão distribuídos de forma relativamente simétrica em torno da média.

$$Md = 19{,}15 \quad \text{(No R: 19,15)}$$

- **Moda (Mo):** Corresponde ao valor que mais se repete. O valor modal de 19,4 indica que essa concentração é a mais comum entre as medições.

$$Mo = 19{,}4 \quad \text{(No R: 19,4)}$$

- **Amplitude (A):** Mede a diferença entre o maior e o menor valor observado, refletindo a variação total dos dados. A amplitude de 25,6 indica uma considerável dispersão entre os valores extremos.

$$A = x_{\max} - x_{\min} = 31{,}8 - 6{,}2 = 25{,}6 \quad \text{(No R: 25,6)}$$

- **Variância amostral (s^2):** Indica o grau de dispersão dos dados em relação à média. O valor de 30,84 mostra que há uma variação moderada entre as medições. O R utiliza a **correção de Bessel** (divisão por $n-1$) para obter uma estimativa não tendenciosa.

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = 30{,}84 \quad \text{(No R: 30,84144)}$$

- **Desvio padrão (s):** É a raiz quadrada da variância e indica, em média, quanto os valores se afastam da média. O desvio de 5,55 sugere que a maior parte das medições se concentra entre 13,5 e 24,5 ppm.

$$s = \sqrt{s^2} = \sqrt{30{,}84} = 5{,}55 \quad \text{(No R: 5,553507)}$$

- **Coeficiente de variação (CV):** Mede a dispersão relativa dos dados em relação à média, em termos percentuais. O valor de 29,2% indica uma **variabilidade moderada**, mostrando que as emissões têm certa flutuação, mas permanecem estáveis em torno da média.

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{5{,}55}{19{,}02} \times 100\% \approx 29{,}20\% \quad \text{(No R: 29,19633\%)}$$

1.2) Com os dados obtidos e as funções `hist()` e `boxplot()`, foi possível construir os gráficos para análise das emissões:

```r
# --- ITEM 2 --- #
par(mfrow = c(1,2)); # Divide a área de plotagem em duas colunas

# Cria nosso histograma
hist(emissions,
     freq = FALSE,
     xlab = "Valores de Emissões",
     ylab = "Densidade",
     main = "Histograma de Emissões", # Título ajustado
     col = "skyblue",
     border = "black")
```

```r

# Cria nosso box plot
boxplot(emissions,
        main = "Boxplot de Emissões", # Título ajustado
        col = "skyblue",
        border = "black")

par(mfrow = c(1, 1)); # Retorna a configuração da área de plotagem ao normal
```

Listado 4: Funções utilizadas para a construção dos gráficos

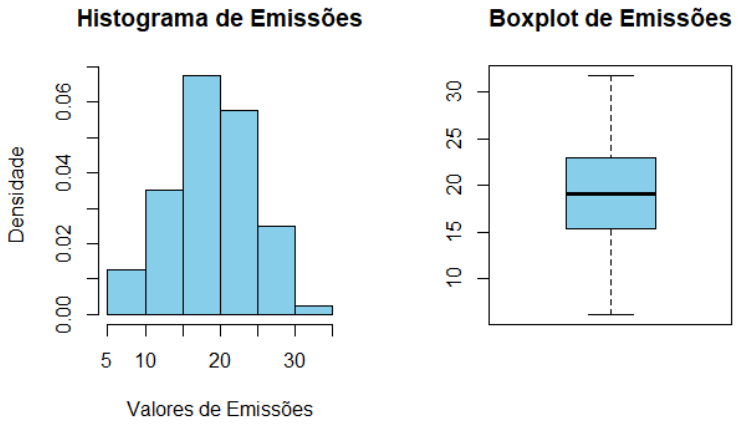


Figura 1: Histograma e boxplot das emissões gerados no R.

Análise: Observando o histograma, percebe-se que os dados não são perfeitamente simétricos, pois há uma concentração maior de valores em faixas menores (próximas de 10) e uma menor frequência nas faixas superiores (em torno de 30). Já o boxplot indica uma distribuição relativamente equilibrada entre os quartis, o que sugere uma leve assimetria, mas sem grandes desvios. Não foram identificados valores atípicos (*outliers*) nos dados.

1.3) Com o uso da função `quantile()` em R, foi possível calcular os quartis e o intervalo interquartil (IQR) dos dados de emissões:

```r
# --- ITEM 3 --- #

q1_emissions = quantile(emissions, 0.25); # Calcula o Quartil 1
q2_emissions = quantile(emissions, 0.5); # Calcula o Quartil 2
q3_emissions = quantile(emissions, 0.75); # Calcula o Quartil 3
IQR_emissions = IQR(emissions); # Calcula o IQR das emissões
```

```r

tabela_quartis <- data.frame(
  Q1 = q1_emissions,
  Q2 = q2_emissions,
  Q3 = q3_emissions,
  IQR = IQR_emissions
)
# Remove o nome da linha
rownames(tabela_quartis) <- NULL

cat("Quartis das Emissões:")
tabela_quartis
```

Listado 5: Cálculo dos valores de quartis e intervalo interquartil

```
Quartis das Emissões:
      Q1    Q2     Q3 IQR
1 15.425 19.15 22.925 7.5
```

Listado 6: Saída do código no console com os dados calculados

Antes de apresentarmos os cálculos teóricos, é importante destacar que o R, por padrão, utiliza o método de interpolação **tipo 7** para calcular quartis. Nesse método, o quantil de ordem p é dado por:

$$Q_p^R = (1-\gamma)\, x_j + \gamma\, x_{j+1}, \quad \text{onde } j = \lfloor (n-1)p \rfloor + 1,\ \gamma = (n-1)p + 1 - j$$

Ou seja, o R pondera linearmente os valores das posições adjacentes de acordo com a fração γ, gerando pequenas diferenças em relação ao cálculo teórico que faz a média simples entre os dois elementos mais próximos.

Cálculos teóricos e valores obtidos no R

Utilizando os dados ordenados da Tabela 2:

- **Número de observações:**

$$n = 80$$

- **Cálculo do Quartil 1 (Q1):**

$$P = 0.25 \times (n+1) = 0.25 \times 81 = 20.25$$

Como P não é inteiro, fazemos a média entre o 20^{o} e o 21^{o} dado:

$$Q1 = \frac{15{,}5 + 15{,}8}{2} = 15{,}65 \quad \text{(No R: 15,4)}$$

- **Cálculo do Quartil 2 (Q2) — Mediana:**

$$P = 0.5 \times (n + 1) = 0.5 \times 81 = 40.5$$

Como P não é inteiro, calculamos a média do 40^{o} e 41^{o} dados:

$$Q2 = \frac{19{,}0 + 19{,}1}{2} = 19{,}05 \quad \text{(No R: 19,1)}$$

- **Cálculo do Quartil 3 (Q3):**

$$P = 0.75 \times (n + 1) = 0.75 \times 81 = 60.75$$

P não é inteiro, então:

$$Q3 = \frac{23{,}9 + 24{,}1}{2} = 24{,}0 \quad \text{(No R: 22,9)}$$

- **Cálculo do Intervalo Interquartil (IQR):**

$$IQR = Q3 - Q1 = 24{,}0 - 15{,}65 = 8{,}35 \quad \text{(No R: 7,5)}$$

- **Limites para verificação de outliers:**

$$\text{Limite Inferior} = Q1 - 1.5 \times IQR = 15{,}65 - 12{,}525 = 3{,}125 \quad \text{(No R: 4,9)}$$

$$\text{Limite Superior} = Q3 + 1.5 \times IQR = 24{,}0 + 12{,}525 = 36{,}525 \quad \text{(No R: 34,15)}$$

Todos os dados estão entre 6,2 e 31,8, portanto não há valores fora desses limites.

1.4) Supondo que o limite máximo aceitável diário para as emissões seja de 25 unidades, podemos calcular a proporção de dias em que a planta excedeu esse limite a partir do código em R a seguir.

```r
# --- ITEM 4 --- #
days_exceeded <- sum(emissions > 25); # Contagem de dias que excederam o limite
total_days <- length(emissions); # Número total de registros de emissões
exceed_proportion <- days_exceeded/total_days; # Proporção de dias excedidos

```

```r
cat("Proporção de dias em que a planta excedeu o limite de 25 unidades:")
exceed_proportion
```

Listado 7: Medição dos dias que ultrapassaram o limite de 25 unidades

```
Proporção de dias em que a planta excedeu o limite de 25 unidades:
[1] 0.1375
```

Listado 8: Saída do código no console

Cálculo teórico utilizando os dados ordenados da Tabela 2:

O número de emissões que ultrapassam 25 unidades é igual a 11. Portanto, a proporção de dias com excesso é:

$$\text{Proporção de excedimento} = \frac{\text{dias excedentes}}{n} = \frac{11}{80} = 0,1375 \approx 13,75\%$$

Interpretação: Como a grande maioria dos valores, 86,25%, está dentro do padrão desejado, o comportamento geral das emissões da planta está em conformidade com o limite regulatório.

2 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.

Tabela 3: Informações dos candidatos para a vaga de gerente.

ID	Idade	Nacionalidade	Renda (mil euros)	Experiência (anos)
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	Alemã	2.5	18
17	48	Francesa	2.0	19
18	33	Italiana	1.7	7
19	38	Alemã	2.1	12
20	46	Italiana	3.2	23

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?

2.2) 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?

2.3) 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.

2.4) 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.

2.5) 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

2.1) A seguir, o código em R para o cálculo da média, mediana e desvio padrão da idade, renda desejada e anos de experiência dos candidatos.

```r
#Questao 2
#Dados dos candidatos na ordem da tabela
idade <- c(28, 34, 46, 26, 37, 29, 51, 31, 39, 43, 58, 44, 25, 23, 52, 42, 48,
           33, 38, 46)
nacionalidade <- c("Italiana", "Inglesa", "Belga", "Espanhola", "Italiana", "
    Espanhola", "Francesa", "Belga", "Italiana", "Italiana","Italiana", "Inglesa",
     "Francesa", "Espanhola", "Italiana","Alemana", "Francesa", "Italiana", "
    Alemana", "Italiana"
)
renda <- c(2.3, 1.6, 1.2, 0.9, 2.1, 1.6, 1.8, 1.4, 1.2, 2.8, 3.4, 2.7, 1.6, 1.2,
    1.1, 2.5, 2.0, 1.7, 2.1, 3.2)
experiencia <- c(2, 8, 21, 1, 15, 3, 28, 5, 13, 20, 32, 23, 1, 0, 29, 18, 19, 7,
    12, 23)

curriculos <- data.frame(idade, nacionalidade, renda, experiencia) #

# Estatisticas descritivas
#item 1
#Informacoes
Informacoes <- c("Idade", "Renda", "Experiencia")
#Medias
media_idade <- mean(idade)
media_Renda <- mean(renda)
media_Experiencia <- mean(experiencia)
Media <- c(media_idade, media_Renda, media_Experiencia)
#Medianas
mediana_idade <- median(idade)
mediana_Renda <- median(renda)
mediana_Experiencia <- median(experiencia)
Mediana <- c(mediana_idade, mediana_Renda, mediana_Experiencia)
```

```
#Desvio padrao
desvio_idade <- sd(idade)
desvio_Renda <- sd(renda)
desvio_Experiencia <- sd(experiencia)
Desvio_Padrao <- c(desvio_idade, desvio_Renda, desvio_Experiencia)

#Display dos dados obtidos
estatisticas = data.frame(Informacoes,Media, Mediana, Desvio_Padrao)
print(estatisticas)
```

Listado 9: Estatísticas descritivas dos currículos

```
     Informacoes Media Mediana Desvio_Padrao
       Idade 38.65   38.50     9.9275003
       Renda  1.92    1.75     0.7134792
 Experiencia 14.00   14.00    10.2700382
```

Listado 10: Saída do código no console com os dados calculados

Interpretação do perfil dos candidatos a partir dos resultados obtidos:

- **Idade:** Com a proximidade entre a média (38,65) e a mediana (38,50) das idades, percebe-se que se tratam de dados simétricos, ou seja, não existem muitas pessoas nos extremos, isto é, muitos novos ou muito velhos, o que provocaria uma distorção da média. Além disso, o desvio padrão (9,93) demonstra que a maioria da idade dos candidatos está na faixa de 29-49 anos.
- **Renda:** Seguindo o mesmo padrão da idade, os dados da renda desejada pelos candidatos também podem ser considerados simétricos, estando a maioria dos dados no intervalo de variação de 0,7 mil euros da média considerada (1,92) de acordo com o desvio padrão.
- **Experiência:** Apesar da média e da mediana serem iguais (14.00), o que indica uma distribuição equilibrada da experiência dos candidatos, o desvio padrão de 10 anos (10,27) evidencia uma grande variação nesse parâmetro, o que pode torná-lo um fator decisivo na contratação do gerente.

Cálculos teóricos e valores obtidos no R Os dados das idades, renda desejada e anos de experiência foram organizados em ordem crescente, conforme mostrado abaixo.

Tabela de Dados Ordenados

Tabela 4: Idades dos candidatos em ordem crescente

23	25	26	28	29	31	33	34	37	38	39	42
43	44	46	46	48	51	52	58				

Tabela 5: Renda desejada (mil euros) em ordem crescente

0.9	1.1	1.2	1.2	1.2	1.4	1.6	1.6	1.6	1.7	1.8	2.0
2.1	2.1	2.3	2.5	2.7	2.8	3.2	3.4				

Tabela 6: Anos de experiência em ordem crescente

0	1	1	2	3	5	7	8	12	13	15	18
19	20	21	23	23	28	29	32				

A partir desses dados e utilizando a base teórica apresentada, foram feitos os cálculos teóricos abaixo e comparados com os resultantes do código em R:

- **Número de observações:**

$$n = 20 \quad \text{(No R: 20)}$$

- **Idade (anos)**

$$\bar{x}_{\text{idade}} = \frac{\sum x_i}{n} = \frac{773}{20} = 39{,}65 \quad \text{(No R: 39,65)}$$

$$Md_{\text{idade}} = \frac{38 + 39}{2} = 38,50 \quad \text{(No R: 38,50)}$$

$$s_{\text{idade}} = \sqrt{\frac{9.56}{20}} \approx 9{,}93 \quad \text{(No R: 9,9275003)}$$

- **Renda desejada (mil euros)**

$$\bar{x}_{\text{renda}} = \frac{38,4}{20} = 1{,}92 \quad \text{(No R: 1,92)}$$

$$Md_{\text{renda}} = \frac{1,7 + 1,8}{2} = 1,75 \quad \text{(No R: 1,75)}$$

$$s_{\text{renda}} = \sqrt{\frac{9.67}{20}} \approx 0.714 \quad \text{(No R: 0,7134792)}$$

- **Experiência (anos)**

$$\bar{x}_{\text{exp}} = \frac{280}{20} = 14{,}00 \quad \text{(No R: 14,00)}$$

$$Md_{\text{exp}} = \frac{13 + 15}{20} = 14,00 \quad \text{(No R: 14, 00)}$$

$$s_{\text{exp}} = \sqrt{\frac{2004}{19}} \approx 10{,}27 \quad \text{(No R: 10,2700382)}$$

2.2) A seguir, o código em R para o agrupamento dos candidatos de acordo com a sua nacionalidade e o cálculo da renda média desejada e dos anos médios de experiência de cada grupo.

```r
#item 2
#bibliotecas usadas
library(dplyr)
library(tidyverse)
grupos_nacionalidade <- curriculos %>%
  group_by(nacionalidade) %>% # agrupa por nacionalidade
  summarise(
    # media da renda em cada nacionalidade
    Renda_media = mean(renda),
    # media de experiencia em cada nacionalidade
    Experiencia_media = mean(experiencia)
  )
cat("Tabela completa:\n")
print(grupos_nacionalidade)
```

Listado 11: Nacionalidades e sua renda e experiência média

```
Tabela completa:
  nacionalidade Renda_media Experiencia_media
  <chr>               <dbl>             <dbl>
 Alemana              2.3               15
 Belga                1.3               13
 Espanhola            1.23               1.33
 Francesa             1.8               16
 Inglesa              2.15              15.5
 Italiana             2.22              17.6
```

Listado 12: Saída do código no console

Nacionalidade de maior renda média desejada: pela observação da tabela, percebe-se que é a "Alemana"(2,3), informação que pode ser também extraída pelo código abaixo.

```
maior_renda <- grupos_nacionalidade %>%
  slice_max(Renda_media)%>% # seleciona a maior renda media
  pull(nacionalidade) # retorna a nacionalidade da media obtida

cat("\nNacionalidade com maior renda media desejada:\n")
print(maior_renda)
```

Listado 13: Maior renda média desejada

```
 Nacionalidade com maior renda media desejada:
[1] "Alemana"
```

Listado 14: Saída do código no console

Nacionalidade que aparenta ser mais experiente: pela observação da tabela, percebe-se que é a "Italiana", já que possui mais ano de experiência(17,6), informação que também pode ser extraída pelo código abaixo.

```
mais_experiente <- grupos_nacionalidade %>%
  slice_max(Experiencia_media) %>% # seleciona maior experiencia media
  pull(nacionalidade) # retorna a nacionalidade da media obtida

cat("Nacionalidade que aparenta ser mais experiente:")
print(mais_experiente)
```

Listado 15: Maior renda média desejada

```
Nacionalidade que aparenta ser mais experiente:[1] "Italiana"
```

Listado 16: Saída do código no console

Cálculos teóricos e valores obtidos no R:A partir dos dados da tabela e utilizando a base teórica apresentada, foram feitos os cálculos teóricos abaixo e comparados com os resultantes do código em R:

Renda média e experiência média por nacionalidade

- **Alemã**

$$\bar{x}_{\text{renda, Alemã}} = \frac{2{,}5 + 2{,}1}{2} = 2{,}3 \quad \text{(No R: 2,3)}$$

$$\bar{x}_{\text{exp, Alemã}} = \frac{18 + 12}{2} = 15 \quad \text{(No R: 15)}$$

- **Belga**

$$\bar{x}_{\text{renda, Belga}} = \frac{1{,}2 + 1{,}4}{2} = 1{,}3 \quad \text{(No R: 1,3)}$$

$$\bar{x}_{\text{exp, Belga}} = \frac{21 + 5}{2} = 13 \quad \text{(No R: 13)}$$

- **Espanhola**

$$\bar{x}_{\text{renda, Espanhola}} = \frac{0{,}9 + 1{,}6 + 1{,}2}{3} \approx 1{,}23 \quad \text{(No R: 1,23)}$$

$$\bar{x}_{\text{exp, Espanhola}} = \frac{1 + 3 + 0}{3} \approx 1{,}33 \quad \text{(No R: 1,33)}$$

- **Francesa**

$$\bar{x}_{\text{renda, Francesa}} = \frac{1{,}8 + 2{,}0 + 1{,}6}{3} = 1{,}8 \quad \text{(No R: 1,8)}$$

$$\bar{x}_{\text{exp, Francesa}} = \frac{28 + 19 + 1}{3} = 16 \quad \text{(No R: 16)}$$

- **Inglesa**

$$\bar{x}_{\text{renda, Inglesa}} = \frac{1{,}6 + 2{,}7}{2} = 2{,}15 \quad \text{(No R: 2,15)}$$

$$\bar{x}_{\text{exp, Inglesa}} = \frac{8 + 23}{2} = 15{,}5 \quad \text{(No R: 15,5)}$$

- **Italiana**

$$\bar{x}_{\text{renda, Italiana}} = \frac{2{,}3 + 2{,}1 + 1{,}2 + 2{,}8 + 3{,}4 + 1{,}1 + 1{,}7 + 3{,}2}{8} = 2{,}22 \quad \text{(No R: 2,22)}$$

$$\bar{x}_{\text{exp, Italiana}} = \frac{2 + 15 + 13 + 20 + 32 + 29 + 7 + 23}{8} = 17{,}6 \quad \text{(No R: 17,6)}$$

2.3) A seguir, o código em R utilizado para a contrução do gráfico de dispersão e para o cálculo do Coeficiente de Pearson, se modo a determinar a correlação entre anos de experiência e renda desejada dos canidatos.

```
1 #item3
2 coeficienteP <- cor(experiencia, renda) # Coeficiente de Pearson
3
```

```r
ggplot(curriculos, aes(x = experiencia, y = renda)) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", color = "red") + # linha de regressao
  labs(title = "Grafico de Dispersao",
       x = "Experiencia",
       y = "Renda desejada") +
  theme_minimal()
```

Listado 17: Funções para a construção do gráfico e cálculo do coeficiente de Pearson

```
   Coeficiente de Pearson: [1] 0.4977672
```

Listado 18: Saída do código no console

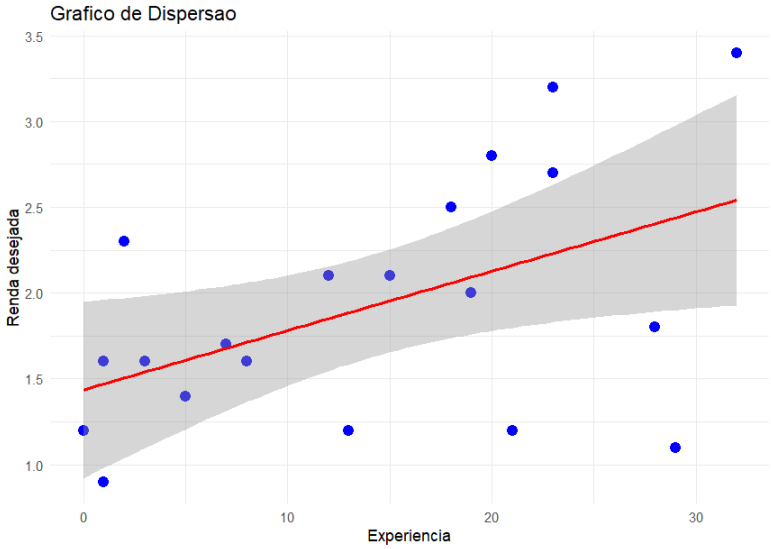


Figura 2: Gráfico de dispersão gerado pelo código

Interpretação dos resultados obtidos: O coeficiente de Pearson calculado mostra uma relação direta entre os anos de experiência e a renda média desejada, uma vez que o valor é positivo. No entanto, o seu módulo moderado evidencia que não se trata de uma correlação muito forte entre esses parâmetros, análise corroborada pelo gráfico de dispersão construído, que apesar de possuir um número considerável de pontos próximos à linha que representaria uma relação linear entre os eixos, ainda apresenta muitos casos que não correspondem ao comportamento esperado.

Cálculos teóricos e valores obtidos no R:

A partir dos dados da tabela e utilizando a base teórica apresentada, foram feitos os cálculos do Coeficiente de correlação de Pearson entre experiência e renda.

$$r = \frac{\sum_{i=1}^{20}(x_i - \bar{x}_{\text{exp}})(y_i - \bar{y}_{\text{renda}})}{\sqrt{\sum_{i=1}^{20}(x_i - \bar{x}_{\text{exp}})^2 \sum_{i=1}^{20}(y_i - \bar{y}_{\text{renda}})^2}} = \frac{176{,}9}{\sqrt{2004 \cdot 63{,}99}} \approx 0{,}498 \quad \text{(No R: 0,4977672)}$$

2.4) A seguir, o código em R utilizado para filtrar os candidatos com pelo menos 10 anos de experiência e renda desejada inferior a 2,0 (mil euros).

```
#item 4
prioridade <- curriculos %>%
  filter(experiencia > 10, renda < 2) %>% # filtra as caracteristicas selecionadas
  select(idade, nacionalidade)

print("Idade e nacionalidade dos candidatos que atendem aos criterios de
    prioridade: \n")
print(prioridade)

print("Numero de candidatos que atendem aos criterios de prioridade:")
print(nrow(prioridade))
```

Listado 19: Seleção dos candidatos piorizados

```
 Idade e nacionalidade dos candidatos que atendem aos criterios de prioridade:
  idade nacionalidade
    46         Belga
    51      Francesa
    39      Italiana
    52      Italiana

Numero de candidatos que atendem aos criterios de prioridade:
[1] 4
```

Listado 20: Saída do código no console

Nesse item, a conclusão sem o R é facilmente feita pela análise da tabela, selecionando os candidatos que atendem aos critérios escolhidos. Dessa forma, chega-se na mesma conclusão fornecida pelo código.

2.5) A seguir, o código em R utilizado para analisar a distribuição das idades dos candidatos de acordo com a nacionalidade usando box-plots.

```r
# item 5
#IDADE
par(mfrow = c(2, 3))#divide o quadro em 6 colunas
nacionalidade_unicas <- unique(curriculos$nacionalidade)
#cria um vetor com as nacionalidades
cores <- c("lightgreen", "skyblue", "gold", "red", "violet", "gray")
#vetor de cores para diferenciar as nacionalidades

for (i in 1:length(nacionalidade_unicas)) { #for para cada nacionalidade
  nac <- nacionalidade_unicas[i]#cada loop trata os dados de uma nacionalidade

  dados_filtrados <- curriculos %>%
    filter(nacionalidade == nac) %>%
    pull(idade)#cria um vetor de idades da nacionalidade atual do loop

  if (i == 1) {#Na primeira iteracao cria o label do grafico
    ylab_text <- "Idade"
    yaxt_setting <- 's'
  }
  #cria o boxplot
  boxplot(dados_filtrados,
          main = nac,
          ylab = ylab_text,
          col = cores[i],
          border = "black",
          ylim = c(20, 55), #define o limite do eixo y
          yaxt = yaxt_setting)
}
```

Listado 21: Distribuição das idades

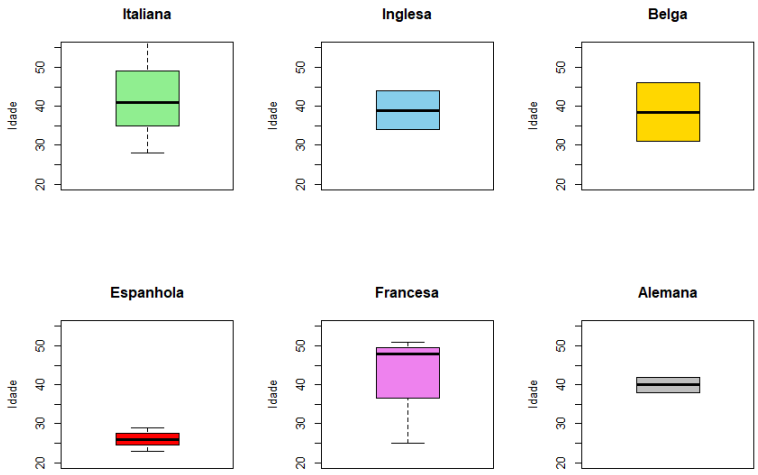


Figura 3: Box-plots da distribuição das idades de acordo com a nacionalidade.

Análise das diferenças entre os grupos: Ao observar os box-plots da distribuição das idades podemos observaar uma grande variedade de padrões. A França apresenta a maior idade típica (mediana próxima a 50 anos) e uma das maiores dispersões de idade, indicando grande variabilidade no grupo. Em contraste, a Espanha é o grupo mais jovem, com a mediana mais baixa (abaixo de 30 anos) e a menor dispersão, sugerindo uma concentração de indivíduos mais jovens. Os grupos Italiana, Inglesa, Belga e Alemã possuem medianas agrupadas em torno da faixa dos 40 anos, mas diferem na variabilidade: a Italiana é mais dispersa (maior interquartil), enquanto a Alemã é a mais homogênea desse subgrupo, com idades muito concentradas em torno da média.

A seguir, o código em R utilizado para analisar a distribuição das idades dos candidatos de acordo com a nacionalidade usando box-plots.

```r
#RENDAS
par(mfrow = c(2,3))
for (i in 1:length(nacionalidade_unicas)) { #for para cada nacionalidade
  nac <- nacionalidade_unicas[i] #cada loop trata os dados de uma nacionalidade

  dados_filtrados <- curriculos %>%
    filter(nacionalidade == nac) %>%
    pull(renda) #cria um vetor de renda desejada da nacionalidade atual do loop

  if (i == 1) { #Na primeira iteracao cria o label do grafico
```

```
    ylab_text <- "Renda"
    yaxt_setting <- 's'
  }
  #cria o boxplot
  boxplot(dados_filtrados,
          main = nac,
          ylab = ylab_text,
          col = cores[i],
          border = "black",
          ylim = c(0.5,3.5),
          yaxt = yaxt_setting)
}
```

Listado 22: Distribuição da renda desejada

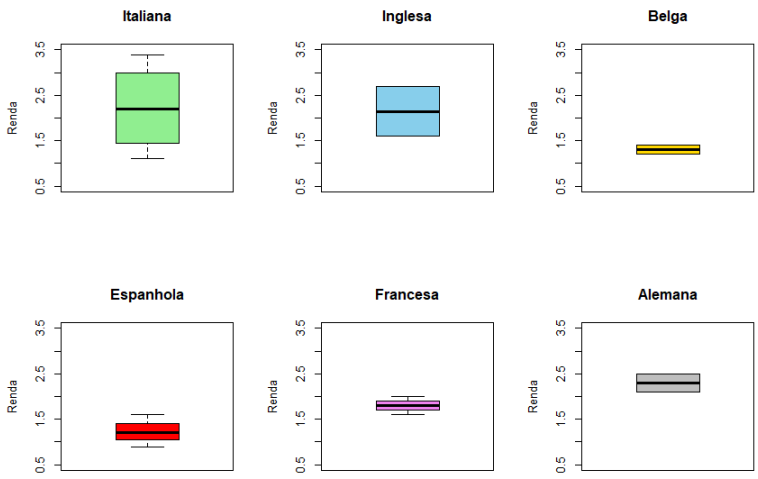


Figura 4: Box-plots da distribuição da renda desejada de acordo com a nacionalidade.

Análise das diferenças entre os grupos: A análise da distribuição de Renda Desejada mostra uma clara divisão entre as nacionalidades: Italiana, Inglesa e Alemã apresentam as maiores rendas típicas desejadas (medianas superiores a 2.0), com a Italiana destacando-se pela maior dispersão e variabilidade de valores solicitados (maior interquartil), indicando uma maior heterogeneidade de expectativas salariais dentro desse grupo. Por outro lado, Belga, Espanhola e Francesa possuem rendas típicas desejadas mais baixas (medianas entre 1.25 e 1.75) e são notavelmente mais homogêneas, o que sugere que a maioria dos candidatos desses países apresenta

expectativas salariais mais alinhadas e com pouca variação entre si.

Cálculos teóricos e valores obtidos no R – Item 5

Utilizando a base teórica para a construção de box-plots, foram feitos os cálculos necessários para encontrar os valores dos quartis, limites superiores e inferiores e o IQR dos box-plots de cada nacionalidade, como mostrado abaixo.

Tabela 7: Estatísticas da Renda por Nacionalidade - Cálculos Detalhados

Nacionalidade	Q1	Q2 (Mediana)	Q3	IQR	Limite inferior	Limite superior
Alemã	2.10	2.30	2.50	0.40	2.1	2.5
Belga	1.30	1.30	1.40	0.10	1.2	1.4
Espanhola	1.05	1.20	1.40	0.35	0.9	1.6
Francesa	1.70	1.90	2.00	0.30	1.6	2.0
Inglesa	2.15	2.15	2.70	0.55	1.6	2.7
Italiana	1.65	2.10	2.80	1.15	1.1	3.4

Cálculos Básicos por Nacionalidade:

- **Alemã (n=2):** (Dados ordenados: 2.1, 2.5)
  - **Q1** = 2.10
  - **Q2** = $\frac{2.1+2.5}{2}$ = 2.30 (mediana)
  - **Q3** = 2.50
  - **IQR** = 2.5 - 2.1 = 0.4
  - **Lim. Inf.** = 2.1 - 1.5 × 0.4 = 1.5 → 2.1 (valor mínimo)
  - **Lim. Sup.** = 2.5 + 1.5 × 0.4 = 3.1 → 2.5 (valor máximo)
- **Belga (n=2):** (Dados ordenados: 1.2, 1.4)

  - **Q1** = $\frac{1.2+1.2}{2} = 1.20$
  - **Q2** = $\frac{1.2+1.4}{2} = 1.30$
  - **Q3** = $\frac{1.4+1.4}{2} = 1.40$
  - **IQR** = 1.4 - 1.2 = 0.2
  - **Lim. Inf.** = 1.2 - 1.5 × 0.2 = 0.9 → 1.2 (valor mínimo)
  - **Lim. Sup.** = 1.4 + 1.5 × 0.2 = 1.7 → 1.4 (valor máximo)
- **Espanhola (n=3):** (Dados ordenados: 0.9, 1.2, 1.6)
  - **Q1** = $\frac{0.9+1.2}{2} = 1.05$ (média do 1º e 2º valores)
  - **Q2** = 1.20 (mediana)
  - **Q3** = $\frac{1.2+1.6}{2} = 1.40$ (média do 2º e 3º valores)
  - **IQR** = 1.4 - 1.05 = 0.35
  - **Lim. Inf.** = 1.05 - 1.5 × 0.35 = 0.525 → 0.9 (valor mínimo)
  - **Lim. Sup.** = 1.4 + 1.5 × 0.35 = 1.925 → 1.6 (valor máximo)
- **Francesa (n=3):** (Dados ordenados: 1.6, 1.8, 2.0)
  - **Q1** = $\frac{1.6+1.8}{2} = 1.70$ (média do 1º e 2º valores)
  - **Q2** = 1.80 (valor central - 2º posição)
  - **Q3** = $\frac{1.8+2.0}{2} = 1.90$ (média do 2º e 3º valores)
  - **IQR** = 1.9 - 1.7 = 0.2
  - **Lim. Inf.** = 1.7 - 1.5 × 0.2 = 1.4 → 1.6 (valor mínimo)
  - **Lim. Sup.** = 1.9 + 1.5 × 0.2 = 2.2 → 2.0 (valor máximo)
- **Inglesa (n=2):** (Dados ordenados: 1.6, 2.7)
  - **Q1** = $\frac{1.6+1.6}{2} = 1.60$ (média dos 2 primeiros valores)
  - **Q2** = $\frac{1.6+2.7}{2} = 2.15$ (média dos 2 valores centrais)
  - **Q3** = $\frac{2.7+2.7}{2} = 2.70$ (média dos 2 últimos valores)
  - **IQR** = 2.7 - 1.6 = 1.1
  - **Lim. Inf.** = 1.6 - 1.5 × 1.1 = -0.05 → 1.6 (valor mínimo)
  - **Lim. Sup.** = 2.7 + 1.5 × 1.1 = 4.35 → 2.7 (valor máximo)
- **Italiana (n=8):** (Dados ordenados: 1.1, 1.2, 1.7, 2.1, 2.3, 2.8, 3.2, 3.4)
  - **Q1** = $\frac{1.2+1.7}{2} = 1.45$ (média do 2º e 3º valores - 25%)
  - **Q2** = $\frac{2.1+2.3}{2} = 2.20$ (média do 4º e 5º valores - 50%)
  - **Q3** = $\frac{2.8+3.2}{2} = 3.00$ (média do 6º e 7º valores - 75%)
  - **IQR** = 3.0 - 1.45 = 1.55

- **Lim. Inf.** = 1.45 - 1.5 × 1.55 = -0.875 → 1.1 (valor mínimo)
- **Lim. Sup.** = 3.0 + 1.5 × 1.55 = 5.325 → 3.4 (valor máximo)

Tabela 8: Estatísticas da Idade por Nacionalidade - Cálculos Detalhados

Nacionalidade	Q1	Q2 (Mediana)	Q3	IQR	Limite inferior	Limite superior
Alemã	40.0	42.0	44.0	4.0	38	42
Belga	36.0	40.0	44.0	8.0	34	46
Espanhola	24.5	26.0	27.5	3.0	23	29
Francesa	34.0	39.0	49.5	15.5	25	51
Inglesa	31.5	34.0	39.0	7.5	34	44
Italiana	31.5	38.5	45.0	13.5	28	52

Cálculos Básicos por Nacionalidade:

- **Alemã (n=2):** (Dados ordenados: 38, 42)
  - **Q1** = 38.0
  - **Q2** = $\frac{38+42}{2}$ = 40.0 (mediana)
  - **Q3** = 42.0
  - **IQR** = 42.0 - 38.0 = 4.0
  - **Lim. Inf.** = 38.0 - 1.5 × 4.0 = 32.0 → 38 (valor mínimo)
  - **Lim. Sup.** = 42.0 + 1.5 × 4.0 = 48.0 → 42 (valor máximo)
- **Belga (n=2):** (Dados ordenados: 34, 46)
  - **Q1** = $\frac{34+34}{2}$ = 34.0
  - **Q2** = $\frac{34+46}{2}$ = 40.0
  - **Q3** = $\frac{46+46}{2}$ = 46.0
  - **IQR** = 46.0 - 34.0 = 12.0

  - **Lim. Inf.** = 34.0 - 1.5 × 12.0 = 16.0 → 34 (valor mínimo)
  - **Lim. Sup.** = 46.0 + 1.5 × 12.0 = 64.0 → 46 (valor máximo)
- **Espanhola (n=3):** (Dados ordenados: 23, 26, 29)
  - **Q1** = $\frac{23+26}{2}$ = 24.5 (média do 1º e 2º valores)
  - **Q2** = 26.0 (mediana)
  - **Q3** = $\frac{26+29}{2}$ = 27.5 (média do 2º e 3º valores)
  - **IQR** = 27.5 - 24.5 = 3.0
  - **Lim. Inf.** = 24.5 - 1.5 × 3.0 = 20.0 → 23 (valor mínimo)
  - **Lim. Sup.** = 27.5 + 1.5 × 3.0 = 32.0 → 29 (valor máximo)
- **Francesa (n=3):** (Dados ordenados: 25, 48, 51)
  - **Q1** = $\frac{25+48}{2}$ = 36.5 (média do 1º e 2º valores)
  - **Q2** = 48.0 (valor central - 2º posição)
  - **Q3** = $\frac{48+51}{2}$ = 49.5 (média do 2º e 3º valores)
  - **IQR** = 49.5 - 36.5 = 13.0
  - **Lim. Inf.** = 36.5 - 1.5 × 13.0 = 17.0 → 25 (valor mínimo)
  - **Lim. Sup.** = 49.5 + 1.5 × 13.0 = 69.0 → 51 (valor máximo)
- **Inglesa (n=2):** (Dados ordenados: 34, 44)
  - **Q1** = $\frac{34+34}{2}$ = 34.0 (média dos 2 primeiros valores)
  - **Q2** = $\frac{34+44}{2}$ = 39.0 (média dos 2 valores centrais)
  - **Q3** = $\frac{44+44}{2}$ = 44.0 (média dos 2 últimos valores)
  - **IQR** = 44.0 - 34.0 = 10.0
  - **Lim. Inf.** = 34.0 - 1.5 × 10.0 = 19.0 → 34 (valor mínimo)
  - **Lim. Sup.** = 44.0 + 1.5 × 10.0 = 59.0 → 44 (valor máximo)
- **Italiana (n=8):** (Dados ordenados: 28, 29, 33, 37, 39, 43, 46, 52)
  - **Q1** = $\frac{29+33}{2}$ = 31.0 (média do 2º e 3º valores - 25%)
  - **Q2** = $\frac{37+39}{2}$ = 38.0 (média do 4º e 5º valores - 50%)
  - **Q3** = $\frac{43+46}{2}$ = 44.5 (média do 6º e 7º valores - 75%)
  - **IQR** = 44.5 - 31.0 = 13.5
  - **Lim. Inf.** = 31.0 - 1.5 × 13.5 = 10.75 → 28 (valor mínimo)
  - **Lim. Sup.** = 44.5 + 1.5 × 13.5 = 64.75 → 52 (valor máximo)

3 Questão 3

O conjunto de dados em anexo, `HW1_bike_sharing.csv`, refere-se ao processo de compartilhamento de bicicletas em uma cidade dos Estados Unidos. O conjunto contém as colunas descritas na Tabela 9.

Tabela 9: Variáveis do conjunto de dados de compartilhamento de bicicletas.

TAG	DESCRIÇÃO	UNIDADES
instant	Índice de registro	
dteday	Data da observação	
season	Estação do ano (1:inverno, 2:primavera, 3:verão, 4:outono)	
weathersit	Condições meteorológicas (1:Céu limpo, 2:Nublado, 3:Chuva fraca, 4:Chuva forte)	
temp	Temperatura em °C (normalizada)	
casual	Número de usuários casuais	
registered	Número de usuários registrados	

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.

3.2) 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.

3.3) 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?

3.4) 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?

Solução da questão 3

3.1) Considerando as definições apresentadas na base teórica, pode-se classificar as variáveis da descrição em númericas ou categóricas como mostrado abaixo

- **Índice de registro:** categórica
- **Data de observação:** categórica
- **Estações do ano:** categórica
- **Condições meteorológicas:** categórica
- **Temperatura em ∘C (normalizada):** numérica
- **Número de usuários casuais:** numérica
- **Número de usuários registrados:** numérica

A seguir, o código em R utilizado para indentificar o número total de observações e as datas de início e fim da amostra como pedido na questão.

```
# ---- Item 1 ----#
#função para obter o número de linhas (=número de observações)
observationNumber <- nrow(dataBikes);
#função para obter o prmeiro elemento do vetor com os dias das observações
startDate <- dataBikes$dteday[1];
#função para obter o último elemento do vetor com os dias das observações
endDate <- dataBikes$dteday[observationNumber];

cat("Numero de observações:")
print(observationNumber)

cat("Data de início da amostra:")
print(startDate)

cat("Data de fim da amostra:")
print(endDate)

```

Listado 23: Número de observações e datas de início e fim.

```
Número de observações:
[1] 731
Data de início da amostra:
[1] "2011-01-01"
Data de fim da amostra:
[1] "2012-12-31"

```

Listado 24: Saída do código no console

3.2) A seguir, o código em R utilizado para calcular a média, a mediana e os quartis das categorias numéricas (temperatura em ºC, número de usuários casuais e número de usuários registrados).

```r
# ---- Item 2 ----#
#função para obter a média -> mean
#função para obter a mediana -> median
#função para obter os quartis -> quantile

#TEMPERATURA
tempMean <- mean(dataBikes$temp);
tempMedian <- median(dataBikes$temp);
tempQuantile <- quantile(dataBikes$temp,
                         probs = c(0.25, 0.5, 0.75))
#USUÁRIOS CASUAIS
casualMean <- mean(dataBikes$casual);
casualMedian <- median(dataBikes$casual);
casualQuantile <- quantile(dataBikes$casual,
                           probs = c(0.25, 0.5, 0.75));
#USUÁRIOS REGISTRADOS
registeredMean <- mean(dataBikes$registered);
registeredMedian <- median(dataBikes$registered);
registeredQuantile <- quantile(dataBikes$registered,
                               probs = c(0.25, 0.5, 0.75));
# Tabela com os dados
centralTendTable <- data.frame(
  Variáveis = c("Temperatura", "Usuários casuais", "Usuários registrados"),
  Média = c(tempMean, casualMean, registeredMean),
  Mediana_Q2 = c(tempMedian, casualMedian, registeredMedian),
  Q1 = c(tempQuantile[1], casualQuantile[1], registeredQuantile[1]),
  Q3 = c(tempQuantile[3], casualQuantile[3], registeredQuantile[3])
)
cat("Medidas de tendência central e quartios calculadas por funções do R:\n")
print(centralTendTable)

```

Listado 25: Número de observações e datas de início e fim.

```
Medidas de tendência central e quartios calculadas pelo R:
            Variáveis      Média Mediana_Q2     Q1     Q3
1          Temperatura   20.31122       20.4   13.8   26.9
2     Usuários casuais  848.17647      713.0  315.5 1096.0
3 Usuários registrados 3656.17237     3662.0 2497.0 4776.5
```

Listado 26: Saída do código no console

Análise dos dados obtidos: As medidas apresentadas permitem compreender o comportamento geral das variáveis analisadas:

- **Temperatura:** A média (20,31) e a mediana (20,4) são praticamente iguais, indicando uma distribuição simétrica das temperaturas. A diferença entre $Q1 = 13{,}8$ e $Q3 = 26{,}9$ mostra uma variação moderada dos valores ao longo do período observado.
- **Usuários casuais:** A média (848,18) é maior que a mediana (713), sugerindo a presença de valores altos ocasionais (assimetria à direita). Isso indica que, em alguns dias, houve um número significativamente maior de usuários ocasionais.
- **Usuários registrados:** A média (3656,17) e a mediana (3662) são muito próximas, apontando para uma distribuição equilibrada no número de usuários frequentes. O intervalo interquartil ($4776{,}5 - 2497{,}0 = 2279{,}5$) mostra uma variação considerável, indicando que a atividade dos usuários registrados oscila de forma mais ampla.

Em síntese, observa-se que a variável **usuários casuais** apresenta a maior variabilidade e leve assimetria, enquanto **temperatura** e **usuários registrados** mantêm comportamento mais uniforme e centrado.

Cálculos teóricos:

- **Média:** nesse caso, o cálculo realizado pela função *mean* do R é o mesmo do convencional mostrado na base teórica, logo os resultados são iguais.

$$\text{Média} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{731} x_i}{731}$$

- **Mediana (Q2):** o cálculo do elemento que representa a mediana também é o mesmo nos dois casos. Assim, tanto a função *median* quanto o método convencional retornarão o valor de posição de número 366 do vetor ordenado considerado, como mostrado abaixo.

$$\text{Mediana (posição)} = 0,5 * (n + 1) = 0,5 * (731 + 1) = 366$$

- **Q1 e Q3:** nesse caso, já podem-se observar diferenças entre os valores calculados, pois a função *quantile* utilizada (de tipo 7 por padrão) aplica uma fórmula diferente do método convencional, como mostrado abaixo.

$$\text{Função } \textit{quantile}\text{: } Q_p = (1 - \gamma) \cdot x_j + \gamma \cdot x_{j+1}$$

onde:

$$
\begin{aligned}
p &= \text{porcentagem de dados no quartil } (p_{q1} = 0,25,\ p_{q3} = 0,75) \\
j &= \lfloor p \cdot (n-1) + 1 \rfloor \\
\gamma &= p \cdot (n-1) + 1 - j \\
n &= \text{número de observações} \\
x &= \text{dados ordenados } (x_1 \leq x_2 \leq \ldots \leq x_n)
\end{aligned}
$$

$$
\text{Fórmula geral: } Q_k = \left(\frac{k(n+1)}{4}\right)\text{-ésimo valor da amostra}, \quad k = 1, 2, 3
$$

Para estabelecer a comparação entre esses resultados, a fórmula convencional foi aplicada por meio da criação de uma função em R, como mostrado abaixo.

```r
# Função para calcular quartis usando o método convencional
calcular_quartis_convencional <- function(x) {
  # Ordenar os dados
  x_ordenado <- sort(x)
  n <- length(x_ordenado)

  # Função para calcular posição do quartil
  calcular_posicao <- function(percentil, n) {
    pos <- percentil * (n + 1)
    return(pos)
  }

  # Calcular posições
  pos_q1 <- calcular_posicao(0.25, n)
  pos_q2 <- calcular_posicao(0.50, n)
  pos_q3 <- calcular_posicao(0.75, n)

  # Função para obter valor baseado na posição
  obter_valor <- function(pos, dados) {
    if (pos == floor(pos)) {
      # Posição inteira
      return(dados[pos])
    } else {
      # Posição não inteira
      pos_inf <- floor(pos)
      pos_sup <- ceiling(pos)
      frac <- pos - pos_inf
      valor <- (1 - frac) * dados[pos_inf] + frac * dados[pos_sup]
      return(valor)
```

```
    }
  }

  # Calcular quartis
  q1 <- obter_valor(pos_q1, x_ordenado)
  q2 <- obter_valor(pos_q2, x_ordenado)
  q3 <- obter_valor(pos_q3, x_ordenado)

  # Retornar resultados
  resultados <- c(Q1 = q1, Q2 = q2, Q3 = q3)
  return(resultados)
}
```

Listado 27: Cálculo dos quartis no modo convencional

```
quartio1 <- calcular_quartis_convencional(dataBikes$temp)
quartio2 <- calcular_quartis_convencional(dataBikes$casual)
quartio3 <- calcular_quartis_convencional(dataBikes$registered)

#tabela com os valores resultantes
tabela_quartis_convencional <- data.frame(
  Variáveis = c("Temperatura", "Usuários casuais", "Usuários registrados"),
  Média = c(mean(dataBikes$temp), mean(dataBikes$casual), mean(dataBikes$
    registered)),
  Mediana_Q2 = c(quartio1["Q2"], quartio2["Q2"], quartio3["Q2"]),
  Q1 = c(quartio1["Q1"], quartio2["Q1"], quartio3["Q1"]),
  Q3 = c(quartio1["Q3"], quartio2["Q3"], quartio3["Q3"])
)
cat("Medidas de tendência central e quartios calculadas convencionalmente:")
tabela_quartis_convencional
```

Listado 28: Aplicação da função nos dados da questão

```
Medidas de tendência central e quartios calculadas convencionalmente:
             Variáveis      Média Mediana_Q2     Q1     Q3
1          Temperatura   20.31122       20.4   13.8   26.9
2     Usuários casuais  848.17647      713.0  315.0 1097.0
3 Usuários registrados 3656.17237     3662.0 2493.0 4790.0
```

Listado 29: Saída do código no console

3.3) A seguir, o código em R utilizado para atribuir os níveis correspondentes às variáveis *season* e *weathersit* e obter a frequência de usuários de cana nível nas duas categorias.

```r
    # ---- Item 3 ----#
# Substituição dos números pelos seus nomes de representação
dataBikes$season <- factor(dataBikes$season,
                           levels = c(1, 2, 3, 4),
                           labels = c("Primavera", "Verão",
                                      "Outono", "Inverno"));
dataBikes$weathersit <- factor(dataBikes$weathersit,
                               levels = c(1, 2, 3, 4),
                               labels = c("Céu limpo", "Nublado",
                                          "Chuva fraca", "Chuva forte"));
# Mostra a frequência de cada estação
freqSeason <- table(dataBikes$season);
cat("Frequência de usuários em cada estção do ano:")
freqSeason
# Mostra a frequência de cada condição climática
freqWeathersit <- table(dataBikes$weathersit);
cat("Frequência de usuários em cada condição climática:")
freqWeathersit
```

Listado 30: Renomeação e tabelas com frequências

```
Frequência de usuários em cada estção do ano:

Primavera     Verão    Outono   Inverno
      181       184       188       178

Frequência de usuários em cada condição climática:

  Céu limpo     Nublado Chuva fraca Chuva forte
        463         247          21           0
```

Listado 31: Saída do código no console

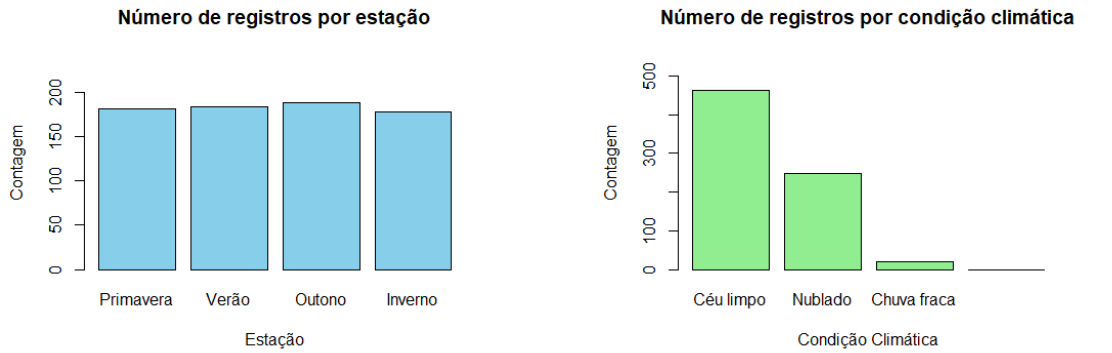
A seguir, o código em R utilizado para a construção dos gráficos de barras para ambas variáveis consideradas.

```r
#gráfico de barras- Estações
barplot(freqSeason,
        col = "skyblue", #cor
        main = "Número de registros por estação", #título
        ylab = "Contagem", #eixo y
        xlab = "Estação", #eixo x
        ylim = c(0, 240) #amplitude eixo y
)
```

```r

#gráfico de barras - Condições Climáticas
barplot(freqWeathersit,
        col = "lightgreen",#cor
        main = "Número de registros por condição climática", #título
        ylab = "Contagem", #eixo y
        xlab = "Condição Climática", #eixo x
        ylim = c(0,550) #amplitude eixo y
)
```

Listado 32: Gráficos de barras para o número de usuários (registros no sistema)



(a) Número de Registros por Estação

(b) Número de Registros por Condição Climática

Figura 5: Gráficos de barras gerados pelo código em R.

Análise da frequência das estações do ano: Para determinar a estação do ano com maior número de usuários e se esse critério inlfuencia o uso do sistema, analisa-se a tabela construída acima para o número de usuários. Assim, percebe-se que, apesar de no Outono esse valor ser maior, a diferença em relação às demais estações é muito baixa, ou seja, ele permaneceu praticamente constante durante o ano todo, independentemente da estação, o que permite afirmar que não se trata de um fator decisivamente influente na frequência de uso do sistema.

Esse comportamento, apesar de facilmente indentificado é corroborado pelo cálculo do desvio padrão desse conjunto de valores, como mostrado abaixo.

Cálculo do Desvio Padrão

Dados: $178, 188, 184, 181$

Média: $\bar{x} = \dfrac{178 + 188 + 184 + 181}{4} = 182.75$

Variância:

$$s^2 = \frac{(178 - 182.75)^2 + (188 - 182.75)^2 + (184 - 182.75)^2 + (181 - 182.75)^2}{4} = 18.25$$

Desvio padrão: $s = \sqrt{18.25} \approx 4.272$

Esse resultado ressalta exatamente a baixa variabilidade desses dados, o que confirma a ideia anterior sobre a influência das estações.

A seguir, o código em R utilizado para retornar a estação com mais usuários.

```r
# Retorna a estação mais frequente
maxSeason <- names(freqSeason)[which.max(freqSeason)];
cat("Estação do ano com maior número de usuários:")
maxSeason
```

Listado 33: Estação com maior número de usuários

```
Estação do ano com maior número de usuários:
[1] "Outono"
```

Listado 34: Saída do código no console

Condição climática mais favorável para o uso do sistema: Nesse caso, pela análise da tabela construída anteriormente, percebe-se que a condição de "Céu Limpo"é a que mais favorece o uso das bicicletas, uma vez que ela possui um número significativamente maior de usuários em relação às outras condições; demonstrando, então, a heterogeneidade desses dados em relação ao critério considerado, fato evidenciado pelo alto desvio padrão desses valores, como mostrado abaixo.

Cálculo do Desvio Padrão

Dados: $463, 247, 21, 0$

$$\text{Média: } \bar{x} = \frac{463 + 247 + 21 + 0}{4} = 182.75$$

$$\text{Variância: } s^2 = \frac{\sum(x_i - \bar{x})^2}{3} = 38806.92$$

$$\text{Desvio padrão: } s = \sqrt{38806.92} \approx 196.99$$

Esse resultado ressalta exatamente a alta variabilidade desses dados, o que confirma a forta influência das condições climáticas no uso do sistema, comportamento que faz sentido logicamente, já que é obviamente mais confortável e prazeroso andar de bicicleta quando o céu está limpo do que quando está chovendo ou se suspeita que irá chover em breve.

A seguir, o código em R utilizado para retornar a condição climática na qual se registrou mais usuários.

```r
# Retorna a condição climática mais frequente
maxWeathersit <- names(freqWeathersit)[which.max(freqWeathersit)];
cat("Condição climática mais favorável para o uso do sistema, ou seja, na qual há
    o maior número de usuários:")
maxWeathersit
```

Listado 35: Condição climática com maior número de usuários

```
Condição climática mais favorável para o uso do sistema, ou seja, na qual há o
    maior número de usuários:
[1] "Céu limpo"
```

Listado 36: Saída do código no console

3.4) A seguir, o código em R utilizado para calcular o número total de usuários por dia, somando *casual* e *registered*, para converter a variável temp para temperatura real (multiplicando por 41) e para a construção dos gráficos de séries temporais para essas duas categorias.

```r
    # ---- Item 4 ----#
# Calcula o número total de usuários por dia
dataBikes$totalUsers <- dataBikes$casual + dataBikes$registered;
# Converte a temperatura normalizada para seu valor real
dataBikes$tempReal <- dataBikes$temp * 41;

par(mfrow = c(2,1))
#Gráfico de usuários
plot(dataBikes$X, dataBikes$totalUsers, type = "l",
     col = "brown", lwd = 2,
     main = "Contagem total de usuários ao longo do tempo",
     xlab = "Dias (Índice)", ylab = "Número de Usuários"
)
#Gráfico de temperatura
plot(dataBikes$X, dataBikes$temp, type = "l",
     col = "red", lwd = 2,
```

```
17      main = "Tendência da temperatura ao longo do tempo",
18      xlab = "Dias (Índice)", ylab = "Temperatura"
19  )
```

Listado 37: Definição das variáveis e construção dos seus gráficos.

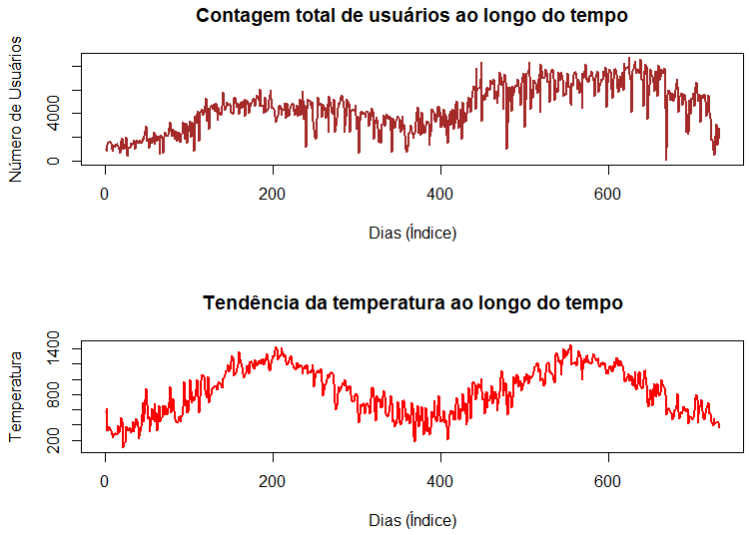


Figura 6: Gráficos gerados pelo código em R

Análise da tendência das séries dos gráficos: Pela análise dos gráficos, observa-se que as séries apresentam uma tendência aproximadamente semelhante, ou seja, percebe-se um aumento no número de usuários em períodos de temperatura mais alta e mais amena e uma certa diminuição quando está mais frio. Esse comportamento pode ser atribuído ao fato de que andar de bicicleta é um esporte ao ar livre, ou seja, a temperatura do ambiente influencia na sua execução.