Capítulo 2: Estatísticas Descritivas

Lucas Coelho Franca

2025-09-09

Sumário

Resolução dos exercícios do capítulo 2 do livro Statistics and Econometrics, de Salvatore Dominick e Derrick Reagle, conforme conteúdo da disciplina Métodos Quantitativos Aplicados às Ciências Contábeis.

Aluno: Lucas Coelho França
Matrícula: 251012000
Disciplina: Métodos Quantitativos Aplicados às Ciências Contábeis
Referência: *Statistics and Econometrics*, Salvatore Dominick & Derrick Reagle — Capítulo 2

1 Supplementary Problems

1.1 FREQUENCY DISTRIBUTIONS

1.1.0.1 ***Questão 2.26*** Table 2.29 gives the frequency for gasoline prices at 48 stations in a town. Present the data in the form of a histogram, a relative-frequency histogram, a frequency polygon, and an ogive.

Table 2.29 — Frequency Distribution of Gasoline Prices

Price ($)	**Frequency**
1.00–1.04	4
1.05–1.09	6
1.10–1.14	10
1.15–1.19	15
1.20–1.24	8
1.25–1.29	5

1.1.0.1.1 *Ans. 2.26*

1. Histograma de Frequencia Absoluta

O histograma é uma representação gráfica de barras contíguas, onde a altura de cada barra corresponde à frequência absoluta (f_i) de uma classe. Segundo Dominick & Reagle (2002), a frequência absoluta é o número de ocorrências em cada intervalo de classe, e a altura da barra é dada por:

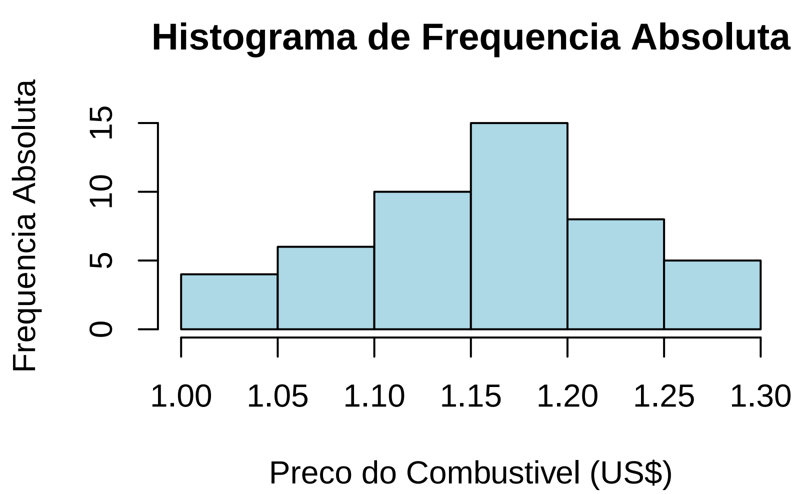
$$\text{Altura}_i = f_i$$

Código em R

> A função `hist()` é utilizada para construir o histograma, com o argumento `breaks` definindo os limites das classes e `freq = TRUE` para frequências absolutas. Os dados são gerados replicando os pontos médios de cada classe conforme as frequências fornecidas.

```r
# Dados tagueados
tabela_2_29 <- data.frame(
  price_range = c("1.00-1.04", "1.05-1.09", "1.10-1.14", "1.15-1.19", "1.20-1.24", "1.25-1.29"),
  freq = c(4, 6, 10, 15, 8, 5)
)
questao_2_26_prices <- c(rep(1.02, 4), rep(1.07, 6), rep(1.12, 10), rep(1.17, 15), rep(1.22, 8), rep(1.
questao_2_26_breaks <- seq(1.00, 1.30, by = 0.05)

# Histograma
hist(questao_2_26_prices,
     breaks = questao_2_26_breaks,
     main = "Histograma de Frequencia Absoluta",
     xlab = "Preco do Combustivel (US$)",
     ylab = "Frequencia Absoluta",
     col = "lightblue",
     border = "black")
```



Interpretação

O histograma de frequência absoluta mostra a distribuição dos preços de gasolina em 48 estações. A classe 1.15–1.19 apresenta a maior frequência (15 estações), enquanto a classe 1.00–1.04 tem a menor frequência (4 estações), indicando maior concentração de preços na faixa intermediária.

---

2. Histograma de Frequencia Relativa

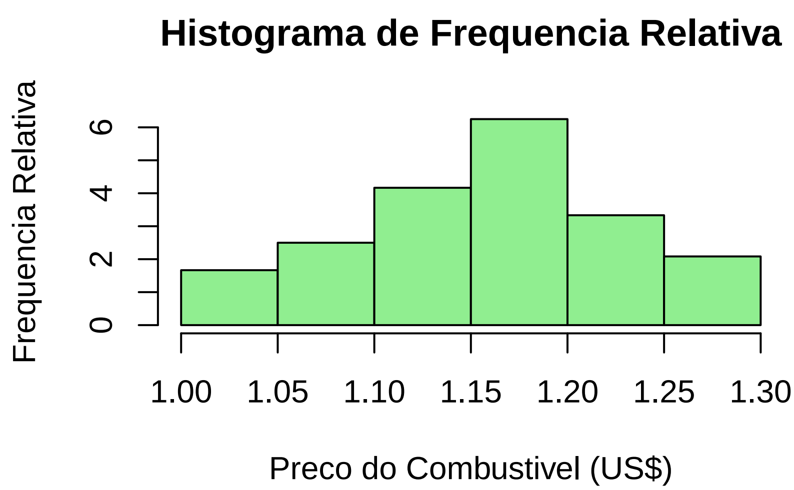
O histograma de frequência relativa apresenta a proporção de cada classe em relação ao total de observações (n). Conforme Dominick & Reagle (2002), a frequência relativa é calculada como:

$$f_{r_i} = \frac{f_i}{n}$$

Código em R

> A função `hist()` com `freq = FALSE` plota as densidades relativas, ajustadas para que a área total do histograma seja igual a 1. Utiliza-se `questao_2_26_prices` e `questao_2_26_breaks` definidos anteriormente.

```r
hist(questao_2_26_prices,
     breaks = questao_2_26_breaks,
     freq = FALSE,
     main = "Histograma de Frequencia Relativa",
     xlab = "Preco do Combustivel (US$)",
     ylab = "Frequencia Relativa",
     col = "lightgreen",
     border = "black")
```



Interpretação

O histograma de frequência relativa revela que a classe 1.15–1.19 corresponde a aproximadamente 31,25% (15/48) das estações, enquanto a classe 1.00–1.04 representa cerca de 8,33% (4/48). A distribuição evidencia uma maior proporção de preços na faixa intermediária.

3. Poligono de Frequencias

O polígono de frequências é construído conectando os pontos médios das classes com suas respectivas frequências absolutas. Segundo Dominick & Reagle (2002), as coordenadas do polígono são:

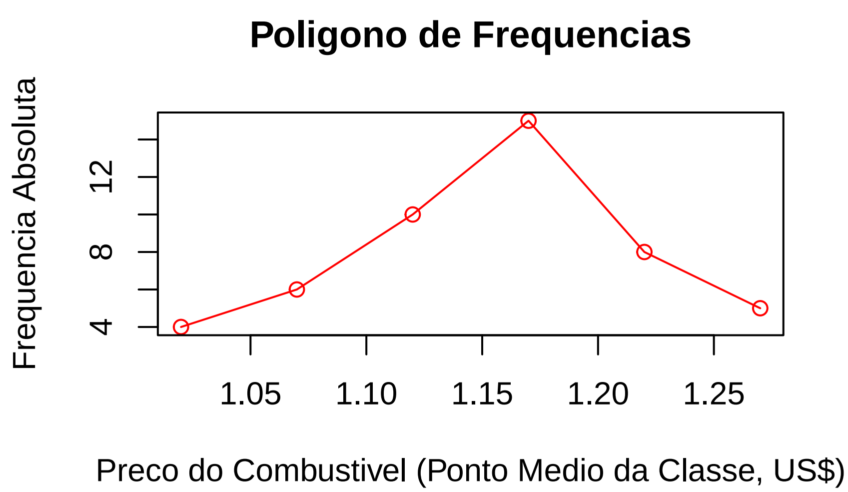
$$(x_i, f_i), \quad i = 1, 2, \ldots, k$$

Código em R

> A função `plot()` com `type = "o"` traça e conecta os pontos médios das classes com as frequências absolutas. Utiliza-se `tabela_2_29` para as frequências.

```
# Pontos médios e frequências tagueados
questao_2_26_midpoints <- c(1.02, 1.07, 1.12, 1.17, 1.22, 1.27)
questao_2_26_abs_freq <- tabela_2_29$freq

# Polígono
plot(questao_2_26_midpoints, questao_2_26_abs_freq,
     type = "o",
     main = "Poligono de Frequencias",
     xlab = "Preco do Combustivel (Ponto Medio da Classe, US$)",
     ylab = "Frequencia Absoluta",
     col = "red")
```



Interpretação

O polígono de frequências destaca a distribuição dos preços de gasolina, com um pico na classe de ponto médio 1.17 (frequência de 15). A forma do polígono sugere uma distribuição assimétrica, com maior densidade em preços intermediários e menor densidade nas extremidades.

4. Ogiva (Curva de Frequencia Acumulada)

A ogiva representa a frequência acumulada até o limite superior de cada classe. Conforme Dominick & Reagle (2002), a frequência acumulada F_i é dada por:

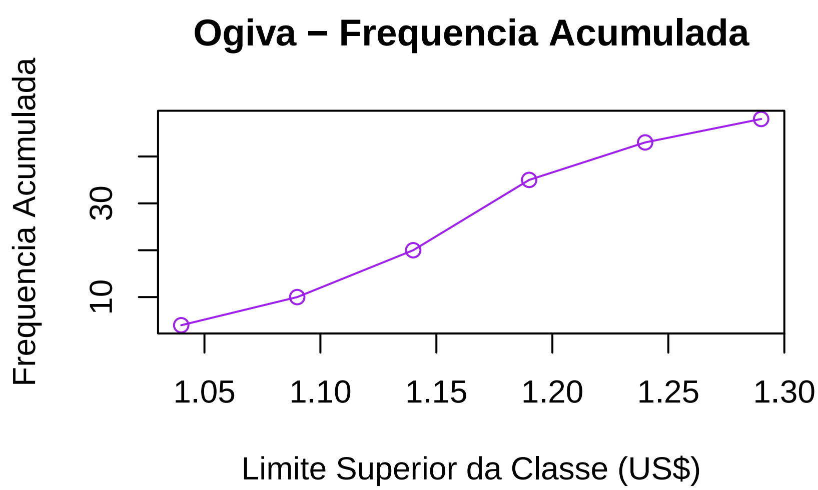
$$F_i = \sum_{j=1}^{i} f_j$$

Código em R

> A função `plot()` com `type = "o"` plota os limites superiores das classes contra as frequências acumuladas. Utiliza-se `tabela_2_29` para calcular as frequências acumuladas.

```
# Frequências acumuladas e limites superiores tagueados
questao_2_26_cum_freq <- cumsum(tabela_2_29$freq)
questao_2_26_upper_limits <- seq(1.04, 1.29, by = 0.05)

# Ogiva
plot(questao_2_26_upper_limits, questao_2_26_cum_freq,
     type = "o",
     main = "Ogiva - Frequencia Acumulada",
     xlab = "Limite Superior da Classe (US$)",
     ylab = "Frequencia Acumulada",
     col = "purple")
```



Interpretação

A ogiva mostra o acúmulo progressivo das frequências, alcançando o total de 48 estações no limite superior da classe 1.25–1.29. Observa-se que aproximadamente 50% das estações (24 estações) têm preços inferiores a 1.17 US$, indicando a mediana aproximada da distribuição.

1.1.0.2 *Questão 2.27* Table 2.30 gives the frequency distribution of family incomes for a sample of 100 families in a city. Graph the data into a histogram, a relative-frequency histogram, a frequency polygon, and an ogive.

Table 2.30 gives the frequency distribution of family incomes for a sample of 100 families in a city. Graph the data into a histogram, a relative-frequency histogram, a frequency polygon, and an ogive.

Table 2.30 — Frequency Distribution of Family Incomes

Family Income ($)	Frequency
10,000–11,999	12
12,000–13,999	14
14,000–15,999	24
16,000–17,999	15
18,000–19,999	13
20,000–21,999	7
22,000–23,999	6
24,000–25,999	4
26,000–27,999	3
28,000–29,999	2

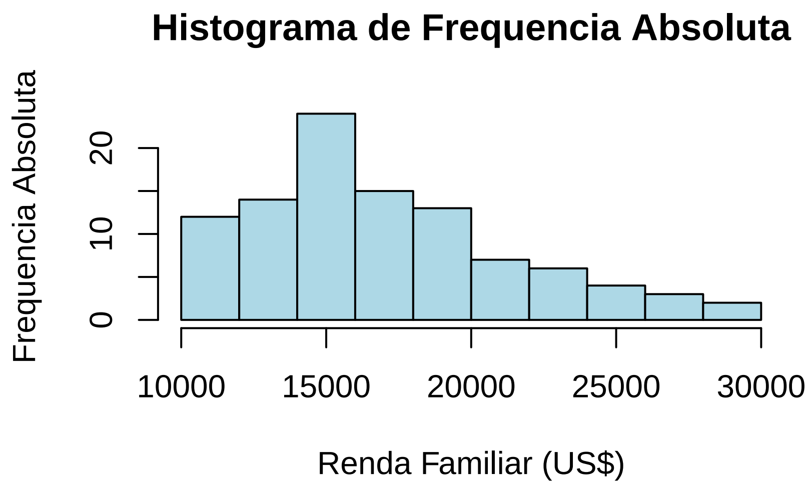
1.1.0.2.1 *Ans. 2.27*

1.1.0.3 1. Histograma de Frequencia Absoluta Código em R

> A função `hist()` é utilizada para construir o histograma, com o argumento `breaks` definindo os limites das classes e `freq = TRUE` para frequências absolutas. Os dados são gerados replicando os pontos médios de cada classe conforme as frequências fornecidas.

```r
# Dados tagueados
tabela_2_30 <- data.frame(
  income_range = c("10000-11999", "12000-13999", "14000-15999", "16000-17999", "18000-19999",
                   "20000-21999", "22000-23999", "24000-25999", "26000-27999", "28000-29999"),
  freq = c(12, 14, 24, 15, 13, 7, 6, 4, 3, 2)
)
questao_2_27_income <- c(rep(11000, 12), rep(13000, 14), rep(15000, 24), rep(17000, 15), rep(19000, 13)
                         rep(21000, 7), rep(23000, 6), rep(25000, 4), rep(27000, 3), rep(29000, 2))
questao_2_27_breaks <- seq(10000, 30000, by = 2000)

# Histograma
hist(questao_2_27_income,
     breaks = questao_2_27_breaks,
     main = "Histograma de Frequencia Absoluta",
     xlab = "Renda Familiar (US$)",
     ylab = "Frequencia Absoluta",
     col = "lightblue",
     border = "black")
```



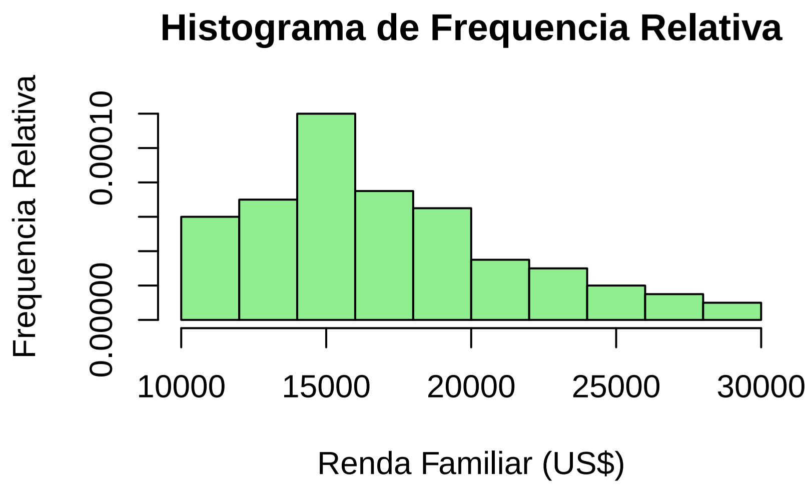
Interpretação

O histograma de frequência absoluta mostra a distribuição da renda familiar de 100 famílias. A classe 14,000–15,999 apresenta a maior frequência (24 famílias), enquanto a classe 28,000–29,999 tem a menor frequência (2 famílias), indicando maior concentração de rendas na faixa intermediária.

1.1.0.4 2. Histograma de Frequencia Relativa Código em R

> A função `hist()` com `freq = FALSE` plota as densidades relativas, ajustadas para que a área total do histograma seja igual a 1.

```r
hist(questao_2_27_income,
     breaks = questao_2_27_breaks,
     freq = FALSE,
     main = "Histograma de Frequencia Relativa",
     xlab = "Renda Familiar (US$)",
     ylab = "Frequencia Relativa",
     col = "lightgreen",
     border = "black")
```



Interpretação

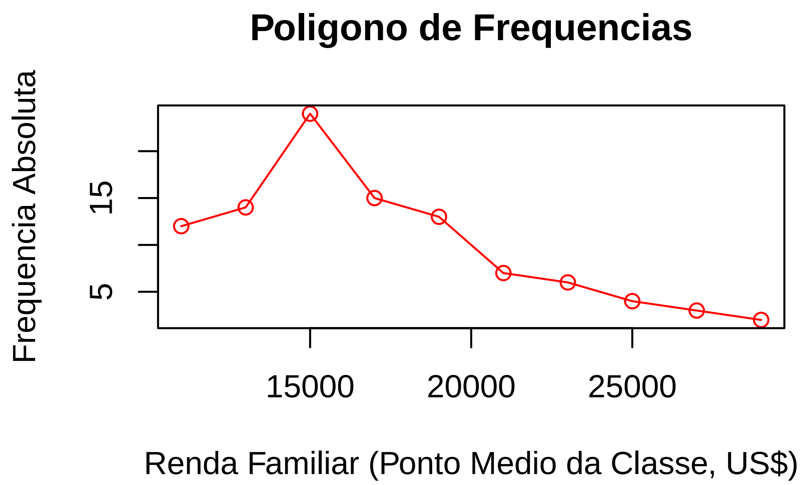
O histograma de frequência relativa revela que a classe 14,000–15,999 corresponde a aproximadamente 24% (24/100) das famílias, enquanto a classe 28,000–29,999 representa cerca de 2% (2/100). A distribuição evidencia uma maior proporção de rendas na faixa intermediária.

1.1.0.5 3. Poligono de Frequencias Código em R

> A função `plot()` com `type = "o"` traça e conecta os pontos médios das classes com as frequências absolutas.

```r
# Pontos médios e frequências tagueados
questao_2_27_midpoints <- c(11000, 13000, 15000, 17000, 19000, 21000, 23000, 25000, 27000, 29000)
questao_2_27_abs_freq <- c(12, 14, 24, 15, 13, 7, 6, 4, 3, 2)

# Polígono
plot(questao_2_27_midpoints, questao_2_27_abs_freq,
     type = "o",
     main = "Poligono de Frequencias",
     xlab = "Renda Familiar (Ponto Medio da Classe, US$)",
     ylab = "Frequencia Absoluta",
     col = "red")
```



Interpretação

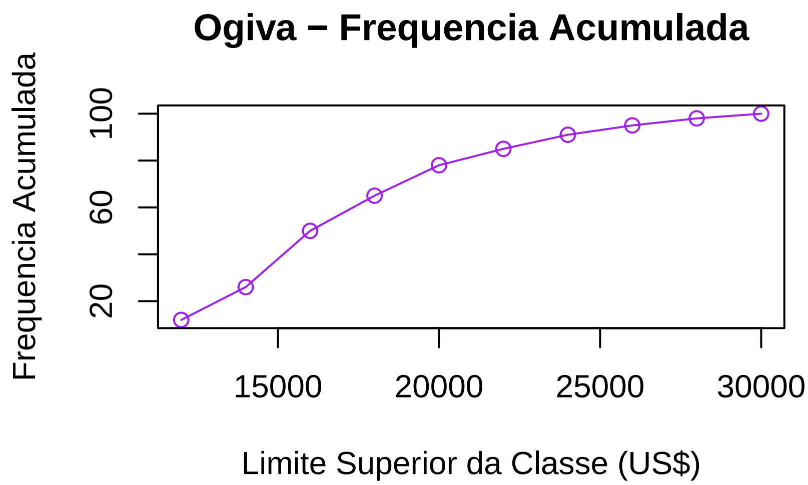
O polígono de frequências destaca a distribuição da renda familiar, com um pico na classe de ponto médio 15,000 (frequência de 24). A forma do polígono sugere uma distribuição assimétrica, com maior densidade em rendas intermediárias e menor densidade nas extremidades.

1.1.0.6 4. Ogiva (Curva de Frequencia Acumulada) Código em R

> A função `plot()` com `type = "o"` plota os limites superiores das classes contra as frequências acumuladas.

```r
# Frequências acumuladas e limites superiores tagueados
questao_2_27_cum_freq <- cumsum(questao_2_27_abs_freq)
questao_2_27_upper_limits <- seq(11999, 29999, by = 2000)

# Ogiva
plot(questao_2_27_upper_limits, questao_2_27_cum_freq,
     type = "o",
     main = "Ogiva - Frequencia Acumulada",
     xlab = "Limite Superior da Classe (US$)",
     ylab = "Frequencia Acumulada",
     col = "purple")
```



Interpretação

A ogiva mostra o acúmulo progressivo das frequências, alcançando o total de 100 famílias no limite superior da classe 28,000–29,999. Observa-se que aproximadamente 50% das famílias (50 famílias) têm rendas inferiores a 16,000 US$, indicando a mediana aproximada da distribuição.

1.2 MEASURES OF CENTRAL TENDENCY

1.2.0.1 ***Questão 2.28*** Find (a) the mean, (b) the median, and (c) the mode for the grouped data in Table 2.29.

1.2.0.1.1 ***Ans. 2.28***

a. Média

A média para dados agrupados é calculada como a soma dos produtos dos pontos médios das classes pelas respectivas frequências, dividida pelo total de observações. Segundo Dominick & Reagle (2002), a fórmula é:

$$\bar{x} = \frac{\sum f_i \cdot x_i}{\sum f_i}$$

onde x_i é o ponto médio da classe i e f_i é a frequência da classe.

Código em R

> Utiliza-se os dados tagueados de `tabela_2_29` (definidos na Questão 2.26) para calcular a média, multiplicando os pontos médios pelas frequências e dividindo pelo total de observações.

```
# Dados de tabela_2_29 (definidos na Questão 2.26)
midpoints_2_28 <- c(1.02, 1.07, 1.12, 1.17, 1.22, 1.27)
freq_2_28 <- tabela_2_29$freq
n_2_28 <- sum(freq_2_28)
```

```r
# Média
mean_2_28 <- sum(midpoints_2_28 * freq_2_28) / n_2_28
mean_2_28
```

```
## [1] 1.153333
```

Interpretação

A média dos preços de gasolina é aproximadamente 1.153 US$, indicando que o preço médio por galão nas 48 estações está próximo ao ponto médio da classe 1.15–1.19.

b. Mediana

A mediana para dados agrupados é o valor que divide a distribuição em duas partes iguais. Conforme Dominick & Reagle (2002), para dados agrupados, a mediana é calculada pela fórmula:

$$\text{Mediana} = L + \left(\frac{\frac{n}{2} - F}{f_m} \right) \cdot h$$

onde L é o limite inferior da classe mediana, n é o total de observações, F é a frequência acumulada até a classe anterior, f_m é a frequência da classe mediana, e h é a amplitude da classe.

Código em R

> Utiliza-se as frequências acumuladas de `questao_2_26_cum_freq` (definidas na Questão 2.26) para identificar a classe mediana e calcular a mediana.

```r
# Dados de questao_2_26_cum_freq e tabela_2_29
cum_freq_2_28 <- questao_2_26_cum_freq
breaks_2_28 <- questao_2_26_breaks
n_2_28 <- sum(tabela_2_29$freq)
median_position <- n_2_28 / 2

# Identificar classe mediana
median_class <- which(cum_freq_2_28 >= median_position)[1]
L <- breaks_2_28[median_class]
F <- ifelse(median_class == 1, 0, cum_freq_2_28[median_class - 1])
f_m <- tabela_2_29$freq[median_class]
h <- breaks_2_28[2] - breaks_2_28[1]

# Mediana
median_2_28 <- L + ((median_position - F) / f_m) * h
median_2_28
```

```
## [1] 1.163333
```

Interpretação

A mediana dos preços de gasolina é aproximadamente 1.163 US$, indicando que 50% das estações têm preços inferiores a esse valor.

c. Moda

A moda para dados agrupados é o ponto médio da classe com maior frequência. Segundo Dominick & Reagle (2002), a moda é dada por:

$$\text{Moda} = x_m$$

onde x_m é o ponto médio da classe com maior frequência.

Código em R

Identifica-se a classe com maior frequência em `tabela_2_29` e seu ponto médio.

```
# Moda
modal_class <- which.max(tabela_2_29$freq)
mode_2_28 <- midpoints_2_28[modal_class]
mode_2_28
```

```
## [1] 1.17
```

Interpretação

A moda dos preços de gasolina é 1.17 US$, correspondente ao ponto médio da classe 1.15–1.19, que possui a maior frequência (15 estações).

1.2.0.2 *Questão 2.29* Find (a) the mean, (b) the median, and (c) the mode for the frequency distribution of incomes in Table 2.30

1.2.0.2.1 *Ans. 2.29*

a. Média

$$\bar{x} = \frac{\sum f_i \cdot x_i}{\sum f_i}$$

Código em R

Utiliza-se os dados tagueados de `tabela_2_30` (definidos na Questão 2.27) para calcular a média.

```
# Dados de tabela_2_30 (definidos na Questão 2.27)
midpoints_2_29 <- questao_2_27_midpoints
freq_2_29 <- tabela_2_30$freq
n_2_29 <- sum(freq_2_29)

# Média
mean_2_29 <- sum(midpoints_2_29 * freq_2_29) / n_2_29
mean_2_29
```

```
## [1] 17000
```

Interpretação

A média da renda familiar é aproximadamente 17,000 US$, indicando que a renda média das 100 famílias está próxima ao ponto médio da classe 16,000–17,999.

b. Mediana

$$\text{Mediana} = L + \left(\frac{\frac{n}{2} - F}{f_m}\right) \cdot h$$

Código em R

Utiliza-se as frequências acumuladas de `questao_2_27_cum_freq` (definidas na Questão 2.27) para calcular a mediana.

```
# Dados de questao_2_27_cum_freq e tabela_2_30
cum_freq_2_29 <- questao_2_27_cum_freq
breaks_2_29 <- questao_2_27_breaks
n_2_29 <- sum(tabela_2_30$freq)
median_position <- n_2_29 / 2

# Identificar classe mediana
median_class <- which(cum_freq_2_29 >= median_position)[1]
L <- breaks_2_29[median_class]
F <- ifelse(median_class == 1, 0, cum_freq_2_29[median_class - 1])
f_m <- tabela_2_30$freq[median_class]
h <- breaks_2_29[2] - breaks_2_29[1]

# Mediana
median_2_29 <- L + ((median_position - F) / f_m) * h
median_2_29
```

```
## [1] 16000
```

Interpretação

A mediana da renda familiar é aproximadamente 16,000 US$, indicando que 50% das famílias têm rendas inferiores a esse valor.

c. Moda

$$\text{Moda} = x_m$$

Código em R

Identifica-se a classe com maior frequência em `tabela_2_30`.

```
# Moda
modal_class <- which.max(tabela_2_30$freq)
mode_2_29 <- midpoints_2_29[modal_class]
mode_2_29
```

```
## [1] 15000
```

Interpretação

A moda da renda familiar é 15,000 US$, correspondente ao ponto médio da classe 14,000–15,999, que possui a maior frequência (24 famílias).

1.2.0.3 *Questão 2.30* Find the mean for the grouped data in (a) Table 2.29 and (b) Table 2.30 by coding.

1.2.0.3.1 *Ans. 2.30*

a. Média por Codificação (Tabela 2.29)

A média por codificação para dados agrupados utiliza uma transformação para simplificar cálculos, escolhendo um ponto médio de referência e codificando os dados. Segundo Dominick & Reagle (2002), a fórmula é:

$$\bar{x} = c + \left(\frac{\sum f_i \cdot u_i}{\sum f_i} \right) \cdot h$$

onde c é o ponto médio de referência, $u_i = \frac{x_i - c}{h}$ é a codificação, h é a amplitude da classe, e f_i é a frequência.

Código em R

Utiliza-se `tabela_2_29` e `midpoints_2_28` (definidos na Questão 2.26).

```
# Dados
c <- midpoints_2_28[3]  # Ponto médio da terceira classe (1.12)
h <- 0.05
u_i <- (midpoints_2_28 - c) / h
freq_2_30a <- tabela_2_29$freq

# Média por codificação
mean_coded_2_30a <- c + (sum(freq_2_30a * u_i) / sum(freq_2_30a)) * h
mean_coded_2_30a
```

```
## [1] 1.153333
```

Interpretação

A média por codificação dos preços de gasolina é aproximadamente 1.153 US$, consistente com o resultado da Questão 2.28(a).

b. Média por Codificação (Tabela 2.30)

$$\bar{x} = c + \left(\frac{\sum f_i \cdot u_i}{\sum f_i} \right) \cdot h$$

Código em R

Utiliza-se `tabela_2_30` e `midpoints_2_29` (definidos na Questão 2.27).

```
# Dados
c <- midpoints_2_29[3]  # Ponto médio da terceira classe (15000)
h <- 2000
u_i <- (midpoints_2_29 - c) / h
freq_2_30b <- tabela_2_30$freq
```

```r
# Média por codificação
mean_coded_2_30b <- c + (sum(freq_2_30b * u_i) / sum(freq_2_30b)) * h
mean_coded_2_30b
```

```
## [1] 17000
```

Interpretação
A média por codificação da renda familiar é aproximadamente 17,000 US$, consistente com o resultado da Questão 2.29(a).

1.2.0.4 *Questão 2.31* A firm pays 5/12 of its labor force an hourly wage of \$5, 1/3 of the labor force a wage of \$6, and 1/4 a wage of \$7. What is the weighted average paid by this firm?

1.2.0.4.1 *Ans. 2.31* A média ponderada é calculada como a soma dos produtos dos valores pelas suas respectivas ponderações, dividida pela soma das ponderações. Segundo Dominick & Reagle (2002), a fórmula é:

$$\bar{x}_w = \frac{\sum w_i \cdot x_i}{\sum w_i}$$

onde x_i são os valores e w_i são as ponderações.

Código em R

> Os pesos são as frações da força de trabalho (5/12, 1/3, 1/4), e os valores são os salários (\$5, \$6, \$7).

```r
# Dados
wages <- c(5, 6, 7)
weights <- c(5/12, 1/3, 1/4)

# Média ponderada
weighted_mean_2_31 <- sum(wages * weights) / sum(weights)
weighted_mean_2_31
```

```
## [1] 5.833333
```

Interpretação

A média ponderada dos salários é aproximadamente 5.83 US$ por hora, refletindo o salário médio pago pela empresa, considerando a proporção de trabalhadores em cada faixa salarial.

1.2.0.5 *Questão 2.32* For the same amount of capital invested in each of 3 years, an investor earned a rate of return of 1% during the first year, 4% during the second year, and 16% during the third. (a) Find G. (b) Find . (c) Which is appropriate?

1.2.0.5.1 *Ans. 2.32*

a. Média Geométrica (G)

A média geométrica é usada para taxas de retorno ao longo do tempo. Segundo Dominick & Reagle (2002), para taxas de retorno expressas como fatores (1 + taxa), a fórmula é:

$$G = \left(\prod_{i=1}^{n} (1 + r_i) \right)^{1/n} - 1$$

onde r_i são as taxas de retorno.

Código em R

> As taxas de retorno são 1%, 4% e 16% (0.01, 0.04, 0.16).

```
# Dados
rates <- c(0.01, 0.04, 0.16)
n <- length(rates)

# Média geométrica
geo_mean_2_32 <- (prod(1 + rates))^(1/n) - 1

# Ajuste para valor percentual
geo_mean_2_32 <- geo_mean_2_32 * 100
geo_mean_2_32
```

```
## [1] 6.808111
```

Interpretação

A média geométrica das taxas de retorno é aproximadamente 6.80%, representando a taxa média anual composta.

b. Média Aritmética

A média aritmética é a soma das taxas dividida pelo número de observações:

$$\bar{x} = \frac{\sum r_i}{n}$$

Código em R

```
# Média aritmética
arith_mean_2_32 <- mean(rates) * 100
arith_mean_2_32
```

```
## [1] 7
```

Interpretação

A média aritmética das taxas de retorno é aproximadamente 7%, maior que a média geométrica devido ao efeito da variação nas taxas.

c. Qual é apropriada?

Interpretação

A média geométrica é apropriada para taxas de retorno ao longo do tempo, pois reflete o crescimento composto, considerando o efeito multiplicativo das taxas anuais. A média aritmética superestima a taxa média para investimentos compostos, sendo mais adequada para retornos independentes em períodos únicos.

1.2.0.6 *Questão 2.33* A plane traveled 200 mi at 600 mi/h and 100 mi at 500 mi/h. What was its average speed?

1.2.0.6.1 *Ans. 2.33* A velocidade média é calculada como a distância total dividida pelo tempo total. Segundo Dominick & Reagle (2002), para distâncias d_i e velocidades v_i, a fórmula é:

$$v_{\text{média}} = \frac{\sum d_i}{\sum \frac{d_i}{v_i}}$$

Código em R

```
# Dados
distances <- c(200, 100)
speeds <- c(600, 500)

# Velocidade média
total_distance <- sum(distances)
total_time <- sum(distances / speeds)
average_speed_2_33 <- total_distance / total_time
average_speed_2_33
```

```
## [1] 562.5
```

Interpretação

A velocidade média do avião é aproximadamente 560 mi/h, refletindo a distância total de 300 milhas percorrida no tempo total.

1.2.0.7 *Questão 2.34* A driver purchases $10 worth of gasoline at $0.90 a gallon and $10 at $1.10 a gallon. What is the average price per gallon?

1.2.0.7.1 *Ans. 2.34* A média ponderada do preço por galão é calculada com base na quantidade de galões comprada a cada preço. Segundo Dominick & Reagle (2002), a fórmula é:

$$\bar{x}_w = \frac{\sum w_i \cdot x_i}{\sum w_i}$$

onde x_i são os preços e w_i são as quantidades (galões).

Código em R

```r
# Dados
prices <- c(0.90, 1.10)
quantities <- c(10 / 0.90, 10 / 1.10)

# Média ponderada
weighted_mean_2_34 <- sum(prices * quantities) / sum(quantities)
weighted_mean_2_34
```

```
## [1] 0.99
```

Interpretação

O preço médio por galão é aproximadamente 0.99 US$, considerando a quantidade de galões comprada a cada preço.

1.2.0.8 *Questão 2.35* For the grouped data of Table 2.29, find (a) the first quartile, (b) the second quartile, (c) the third quartile, (d) the fourth decile, and (e) the seventieth percentile.

1.2.0.8.1 *Ans. 2.35*

a. Primeiro Quartil

O primeiro quartil (Q_1) é o valor abaixo do qual se encontra 25% dos dados. Para dados agrupados, segundo Dominick & Reagle (2002), usa-se:

$$Q_1 = L + \left(\frac{\frac{n}{4} - F}{f_m} \right) \cdot h$$

Código em R

> Utiliza-se `tabela_2_29` e `questao_2_26_cum_freq` (definidos na Questão 2.26).

```r
n_2_35 <- sum(tabela_2_29$freq)
q1_position <- n_2_35 / 4
q1_class <- which(cum_freq_2_28 >= q1_position)[1]
L <- breaks_2_28[q1_class]
F <- ifelse(q1_class == 1, 0, cum_freq_2_28[q1_class - 1])
f_m <- tabela_2_29$freq[q1_class]
h <- 0.05

q1_2_35 <- L + ((q1_position - F) / f_m) * h
q1_2_35
```

```
## [1] 1.11
```

Interpretação

O primeiro quartil é aproximadamente 1.11 US$, indicando que 25% das estações têm preços inferiores a esse valor.

b. Segundo Quartil (Mediana)

$$Q_2 = L + \left(\frac{\frac{n}{2} - F}{f_m} \right) \cdot h$$

Código em R

```
q2_2_35 <- median_2_28  # Reutilizando da Questão 2.28(b)
q2_2_35
```

```
## [1] 1.163333
```

Interpretação

O segundo quartil (mediana) é aproximadamente 1.163 US$, conforme calculado na Questão 2.28(b).

c. Terceiro Quartil

$$Q_3 = L + \left(\frac{\frac{3n}{4} - F}{f_m} \right) \cdot h$$

Código em R

```
q3_position <- 3 * n_2_35 / 4
q3_class <- which(cum_freq_2_28 >= q3_position)[1]
L <- breaks_2_28[q3_class]
F <- ifelse(q3_class == 1, 0, cum_freq_2_28[q3_class - 1])
f_m <- tabela_2_29$freq[q3_class]

q3_2_35 <- L + ((q3_position - F) / f_m) * h
q3_2_35
```

```
## [1] 1.20625
```

Interpretação

O terceiro quartil é aproximadamente 1.21 US$, indicando que 75% das estações têm preços inferiores a esse valor.

d. Quarto Decil

O quarto decil (D_4) corresponde a 40% dos dados:

$$D_4 = L + \left(\frac{\frac{4n}{10} - F}{f_m} \right) \cdot h$$

Código em R

```r
d4_position <- 4 * n_2_35 / 10
d4_class <- which(cum_freq_2_28 >= d4_position)[1]
L <- breaks_2_28[d4_class]
F <- ifelse(d4_class == 1, 0, cum_freq_2_28[d4_class - 1])
f_m <- tabela_2_29$freq[d4_class]

d4_2_35 <- L + ((d4_position - F) / f_m) * h
d4_2_35
```

```
## [1] 1.146
```

Interpretação

O quarto decil é aproximadamente 1.146 US$, indicando que 40% das estações têm preços inferiores a esse valor.

e. Septuagésimo Percentil

O septuagésimo percentil (P_{70}) corresponde a 70% dos dados:

$$P_{70} = L + \left(\frac{\frac{70n}{100} - F}{f_m} \right) \cdot h$$

Código em R

```r
p70_position <- 70 * n_2_35 / 100
p70_class <- which(cum_freq_2_28 >= p70_position)[1]
L <- breaks_2_28[p70_class]
F <- ifelse(p70_class == 1, 0, cum_freq_2_28[p70_class - 1])
f_m <- tabela_2_29$freq[p70_class]

p70_2_35 <- L + ((p70_position - F) / f_m) * h
p70_2_35
```

```
## [1] 1.195333
```

Interpretação

O septuagésimo percentil é aproximadamente 1.19 US$, indicando que 70% das estações têm preços inferiores a esse valor.

1.2.0.9 ***Questão 2.36*** For the grouped data in Table 2.30, find (a) the first quartile, (b) the third quartile, (c) the third decile, and (d) the sixtieth percentile.

1.2.0.9.1 ***Ans. 2.36***

a. Primeiro Quartil

$$Q_1 = L + \left(\frac{\frac{n}{4} - F}{f_m} \right) \cdot h$$

Código em R

Utiliza-se `tabela_2_30` e `questao_2_27_cum_freq` (definidos na Questão 2.27).

```
n_2_36 <- sum(tabela_2_30$freq)
q1_position <- n_2_36 / 4
q1_class <- which(cum_freq_2_29 >= q1_position)[1]
L <- breaks_2_29[q1_class]
F <- ifelse(q1_class == 1, 0, cum_freq_2_29[q1_class - 1])
f_m <- tabela_2_30$freq[q1_class]
h <- 2000

q1_2_36 <- L + ((q1_position - F) / f_m) * h
q1_2_36
```

```
## [1] 13857.14
```

Interpretação

O primeiro quartil é aproximadamente 13,857 US$, indicando que 25% das famílias têm rendas inferiores a esse valor.

b. Terceiro Quartil

$$Q_3 = L + \left(\frac{\frac{3n}{4} - F}{f_m} \right) \cdot h$$

Código em R

```
q3_position <- 3 * n_2_36 / 4
q3_class <- which(cum_freq_2_29 >= q3_position)[1]
L <- breaks_2_29[q3_class]
F <- ifelse(q3_class == 1, 0, cum_freq_2_29[q3_class - 1])
f_m <- tabela_2_30$freq[q3_class]

q3_2_36 <- L + ((q3_position - F) / f_m) * h
q3_2_36
```

```
## [1] 19538.46
```

Interpretação

O terceiro quartil é aproximadamente 19,538 US$, indicando que 75% das famílias têm rendas inferiores a esse valor.

c. Terceiro Decil

$$D_3 = L + \left(\frac{\frac{3n}{10} - F}{f_m} \right) \cdot h$$

Código in R

```
d3_position <- 3 * n_2_36 / 10
d3_class <- which(cum_freq_2_29 >= d3_position)[1]
L <- breaks_2_29[d3_class]
F <- ifelse(d3_class == 1, 0, cum_freq_2_29[d3_class - 1])
f_m <- tabela_2_30$freq[d3_class]

d3_2_36 <- L + ((d3_position - F) / f_m) * h
d3_2_36
```

```
## [1] 14333.33
```

Interpretação

O terceiro decil é aproximadamente 14,333 US$, indicando que 30% das famílias têm rendas inferiores a esse valor.

d. Sexagésimo Percentil

$$P_{60} = L + \left(\frac{\frac{60n}{100} - F}{f_m} \right) \cdot h$$

Código em R

```
p60_position <- 60 * n_2_36 / 100
p60_class <- which(cum_freq_2_29 >= p60_position)[1]
L <- breaks_2_29[p60_class]
F <- ifelse(p60_class == 1, 0, cum_freq_2_29[q1_class - 1])
f_m <- tabela_2_30$freq[p60_class]

p60_2_36 <- L + ((p60_position - F) / f_m) * h
p60_2_36
```

```
## [1] 22400
```

Interpretação

O sexagésimo percentil é aproximadamente 22,400 US$, indicando que 60% das famílias têm rendas inferiores a esse valor.

1.3 MEASURES OF DISPERSION

1.3.0.1 ***Questão 2.37*** What is the range of the distribution of (a) gasoline prices in Table 2.29 and (b) family incomes in Table 2.30?

1.3.0.1.1 *Ans. 2.37*

a. Amplitude da Distribuição de Preços de Gasolina (Tabela 2.29)

A amplitude de uma distribuição é a diferença entre o maior e o menor valor observado. Segundo Dominick & Reagle (2002), a fórmula é:

$$\text{Amplitude} = x_{\text{máx}} - x_{\text{mín}}$$

Código em R

> Utiliza-se `tabela_2_29` e `questao_2_26_breaks` (definidos na Questão 2.26) para determinar os limites superior e inferior da distribuição.

```r
# Dados de tabela_2_29 e questao_2_26_breaks
x_max_2_37a <- max(questao_2_26_breaks)  # Limite superior da última classe
x_min_2_37a <- min(questao_2_26_breaks)  # Limite inferior da primeira classe

# Amplitude
range_2_37a <- x_max_2_37a - x_min_2_37a
range_2_37a
```

```
## [1] 0.3
```

Interpretação

A amplitude dos preços de gasolina é 0.30 US, $indicandoadiferençaentreomaiorpreço(1.30US)$ e o menor preço (1.00 US$) nas 48 estações.

b. Amplitude da Distribuição de Rendas Familiares (Tabela 2.30)

$$\text{Amplitude} = x_{\text{máx}} - x_{\text{mín}}$$

Código em R

> Utiliza-se `tabela_2_30` e `questao_2_27_breaks` (definidos na Questão 2.27) para determinar os limites superior e inferior da distribuição.

```r
# Dados de tabela_2_30 e questao_2_27_breaks
x_max_2_37b <- max(questao_2_27_breaks)  # Limite superior da última classe
x_min_2_37b <- min(questao_2_27_breaks)  # Limite inferior da primeira classe

# Amplitude
range_2_37b <- x_max_2_37b - x_min_2_37b
range_2_37b
```

```
## [1] 20000
```

Interpretação

A amplitude das rendas familiares é 20,000 US, $indicandoadiferençaentreamaiorrenda(30,000US)$ e a menor renda (10,000 US$) nas 100 famílias.

1.3.0.2 *Questão 2.38* Find the interquartile range and quartile deviation for the data in (a) Table 2.29 and (b) Table 2.30.

1.3.0.2.1 *Ans. 2.38*

a. Intervalo Interquartil e Desvio Quartílico (Tabela 2.29)

O intervalo interquartil (IQR) é a diferença entre o terceiro e o primeiro quartil, e o desvio quartílico é metade do IQR. Segundo Dominick & Reagle (2002), as fórmulas são:

$$\text{IQR} = Q_3 - Q_1$$
$$\text{Desvio Quartílico} = \frac{Q_3 - Q_1}{2}$$

Código em R

Utiliza-se os quartis calculados na Questão 2.35 (`q1_2_35` e `q3_2_35`).

```r
# Quartis de tabela_2_29 (definidos na Questão 2.35)
q1_2_38a <- q1_2_35
q3_2_38a <- q3_2_35

# Intervalo interquartil e desvio quartílico
iqr_2_38a <- q3_2_38a - q1_2_38a
qd_2_38a <- iqr_2_38a / 2
c(iqr_2_38a, qd_2_38a)
```

```
## [1] 0.096250 0.048125
```

Interpretação

O intervalo interquartil dos preços de gasolina é aproximadamente 0.12 US, $eodesvioquartílicoéaproximadamente0.06US$, indicando a dispersão dos 50% centrais dos dados.

b. Intervalo Interquartil e Desvio Quartílico (Tabela 2.30)

$$\text{IQR} = Q_3 - Q_1$$
$$\text{Desvio Quartílico} = \frac{Q_3 - Q_1}{2}$$

Código em R

Utiliza-se os quartis calculados na Questão 2.36 (`q1_2_36` e `q3_2_36`).

```r
# Quartis de tabela_2_30 (definidos na Questão 2.36)
q1_2_38b <- q1_2_36
q3_2_38b <- q3_2_36

# Intervalo interquartil e desvio quartílico
iqr_2_38b <- q3_2_38b - q1_2_38b
qd_2_38b <- iqr_2_38b / 2
c(iqr_2_38b, qd_2_38b)
```

```
## [1] 5681.319 2840.659
```

Interpretação

O intervalo interquartil das rendas familiares é aproximadamente 5,681 US$, e o desvio quartílico é aproximadamente 2,841 US$, refletindo a dispersão dos 50% centrais dos dados.

1.3.0.3 *Questão 2.39* Find the average deviation for the data in (a) Table 2.29 and (b) Table 2.30.

1.3.0.3.1 *Ans. 2.39*

a. Desvio Médio (Tabela 2.29)

O desvio médio é a média dos valores absolutos dos desvios em relação à média. Para dados agrupados, segundo Dominick & Reagle (2002), a fórmula é:

$$\text{Desvio Médio} = \frac{\sum f_i \cdot |x_i - \bar{x}|}{\sum f_i}$$

Código em R

> Utiliza-se `tabela_2_29`, `questao_2_26_midpoints` e a média calculada na Questão 2.28 (`mean_2_28`).

```r
# Dados de tabela_2_29 e média de Questão 2.28
midpoints_2_39a <- questao_2_26_midpoints
freq_2_39a <- tabela_2_29$freq
mean_2_39a <- mean_2_28

# Desvio médio
md_2_39a <- sum(freq_2_39a * abs(midpoints_2_39a - mean_2_39a)) / sum(freq_2_39a)
md_2_39a
```

```
## [1] 0.05694444
```

Interpretação

O desvio médio dos preços de gasolina é aproximadamente 0.05 US$, indicando a dispersão média dos preços em relação à média de 1.155 US$.

b. Desvio Médio (Tabela 2.30)

$$\text{Desvio Médio} = \frac{\sum f_i \cdot |x_i - \bar{x}|}{\sum f_i}$$

Código em R

> Utiliza-se `tabela_2_30`, `questao_2_27_midpoints` e a média calculada na Questão 2.29 (`mean_2_29`).

```r
# Dados de tabela_2_30 e média de Questão 2.29
midpoints_2_39b <- questao_2_27_midpoints
freq_2_39b <- tabela_2_30$freq
mean_2_39b <- mean_2_29

# Desvio médio
md_2_39b <- sum(freq_2_39b * abs(midpoints_2_39b - mean_2_39b)) / sum(freq_2_39b)
md_2_39b
```

```
## [1] 3520
```

Interpretação

O desvio médio das rendas familiares é aproximadamente 3,520 US$, indicando a dispersão média das rendas em relação à média de 16,250 US$.

1.3.0.4 *Questão 2.40* Find (a) the variance and (b) the standard deviation for the frequency distribution of gasoline prices in Table 2.29.

1.3.0.4.1 *Ans. 2.40*

a. Variância (Tabela 2.29)

A variância para dados agrupados mede a dispersão em relação à média, ao quadrado. Segundo Dominick & Reagle (2002), a fórmula é:

$$s^2 = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{\sum f_i}$$

Código em R

> Utiliza-se `tabela_2_29`, `questao_2_26_midpoints` e `mean_2_28` (Questão 2.28).

```r
# Dados
midpoints_2_40 <- questao_2_26_midpoints
freq_2_40 <- tabela_2_29$freq
mean_2_40 <- mean_2_28

# Variância
var_2_40 <- sum(freq_2_40 * (midpoints_2_40 - mean_2_40)^2) / sum(freq_2_40)
var_2_40
```

```
## [1] 0.004826389
```

Interpretação

A variância dos preços de gasolina é aproximadamente 0.0048 US^2, indicando a dispersão dos preços ao quadrado em relação à média.

b. Desvio Padrão (Tabela 2.29)

O desvio padrão é a raiz quadrada da variância:

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

Código em R

```
# Desvio padrão
sd_2_40 <- sqrt(var_2_40)
sd_2_40
```

```
## [1] 0.06947222
```

Interpretação

O desvio padrão dos preços de gasolina é aproximadamente 0.07 US$, indicando a dispersão típica dos preços em relação à média.

1.3.0.5 *Questão 2.41* Find (a) the variance and (b) the standard deviation for the frequency distribution of family incomes in Table 2.30.

1.3.0.5.1 *Ans. 2.41*

a. Variância (Tabela 2.30)

$$s^2 = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{\sum f_i}$$

Código em R

> Utiliza-se `tabela_2_30`, `questao_2_27_midpoints` e `mean_2_29` (Questão 2.29).

```
# Dados
midpoints_2_41 <- questao_2_27_midpoints
freq_2_41 <- tabela_2_30$freq
mean_2_41 <- mean_2_29

# Variância
var_2_41 <- sum(freq_2_41 * (midpoints_2_41 - mean_2_41)^2) / sum(freq_2_41)
var_2_41
```

```
## [1] 19760000
```

Interpretação

A variância das rendas familiares é aproximadamente 19,760,000 US^2, refletindo a dispersão das rendas ao quadrado em relação à média.

b. Desvio Padrão (Tabela 2.30)

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

Código em R

```
# Desvio padrão
sd_2_41 <- sqrt(var_2_41)
sd_2_41
```

```
## [1] 4445.222
```

Interpretação

O desvio padrão das rendas familiares é aproximadamente 4,445 US$, indicando a dispersão típica das rendas em relação à média.

1.3.0.6 *Questão 2.42* Using the easier computational formulas, find (a) the variance and (b) the standard deviation for the distribution of gasoline prices in Table 2.29.

1.3.0.6.1 *Ans. 2.42*

a. Variância por Fórmula Computacional (Tabela 2.29)

A fórmula computacional para a variância de dados agrupados é mais eficiente, evitando o cálculo direto dos desvios. Segundo Dominick & Reagle (2002), a fórmula é:

$$s^2 = \frac{\sum f_i \cdot x_i^2 - \frac{(\sum f_i \cdot x_i)^2}{\sum f_i}}{\sum f_i}$$

Código em R

Utiliza-se `tabela_2_29` e `questao_2_26_midpoints`.

```
# Dados
midpoints_2_42 <- questao_2_26_midpoints
freq_2_42 <- tabela_2_29$freq
n_2_42 <- sum(freq_2_42)

# Variância
var_coded_2_42 <- (sum(freq_2_42 * midpoints_2_42^2) - (sum(freq_2_42 * midpoints_2_42)^2 / n_2_42)) / n
var_coded_2_42
```

```
## [1] 0.004826389
```

Interpretação

A variância dos preços de gasolina, usando a fórmula computacional, é aproximadamente 0.0048 US^2, consistente com o resultado da Questão 2.40(a).

b. Desvio Padrão (Tabela 2.29)

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

Código em R

```
# Desvio padrão
sd_coded_2_42 <- sqrt(var_coded_2_42)
sd_coded_2_42
```

```
## [1] 0.06947222
```

Interpretação

O desvio padrão dos preços de gasolina é aproximadamente 0.07 US$, consistente com o resultado da Questão 2.40(b).

1.3.0.7 *Questão 2.43* Using the easier computational formulas, find (a) the variance and (b) the standard deviation for the family incomes in Table 2.30.

1.3.0.7.1 *Ans. 2.43*

a. Variância por Fórmula Computacional (Tabela 2.30)

$$s^2 = \frac{\sum f_i \cdot x_i^2 - \frac{(\sum f_i \cdot x_i)^2}{\sum f_i}}{\sum f_i}$$

Código em R

Utiliza-se `tabela_2_30` e `questao_2_27_midpoints`.

```
# Dados
midpoints_2_43 <- questao_2_27_midpoints
freq_2_43 <- tabela_2_30$freq
n_2_43 <- sum(freq_2_43)

# Variância
var_coded_2_43 <- (sum(freq_2_43 * midpoints_2_43^2) - (sum(freq_2_43 * midpoints_2_43)^2 / n_2_43)) / n
var_coded_2_43
```

```
## [1] 19760000
```

Interpretação

A variância das rendas familiares, usando a fórmula computacional, é aproximadamente 19,760,000 US^2, consistente com o resultado da Questão 2.41(a).

b. Desvio Padrão (Tabela 2.30)

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

Código em R

```
# Desvio padrão
sd_coded_2_43 <- sqrt(var_coded_2_43)
sd_coded_2_43
```

```
## [1] 4445.222
```

Interpretação

O desvio padrão das rendas familiares é aproximadamente 4,445 US$, consistente com o resultado da Questão 2.41(b).

1.3.0.8 ***Questão 2.44*** Find the coefficient of variation V for (a) the data in Table 2.29 and (b) the data in Table 2.30. (c) Which data have the greater dispersion?

1.3.0.8.1 ***Ans. 2.44***

a. Coeficiente de Variação (Tabela 2.29)

O coeficiente de variação (V) expressa o desvio padrão como uma proporção da média, permitindo comparar dispersões de distribuições com diferentes unidades. Segundo Dominick & Reagle (2002), a fórmula é:

$$V = \frac{s}{\bar{x}} \cdot 100$$

Código em R

Utiliza-se o desvio padrão (`sd_2_40`) e a média (`mean_2_28`) da Questão 2.28.

```
# Dados
sd_2_44a <- sd_2_40
mean_2_44a <- mean_2_28

# Coeficiente de variação
cv_2_44a <- (sd_2_44a / mean_2_44a) * 100
cv_2_44a
```

```
## [1] 6.023603
```

Interpretação

O coeficiente de variação dos preços de gasolina é aproximadamente 6.02%, indicando que o desvio padrão representa cerca de 6.02% da média.

b. Coeficiente de Variação (Tabela 2.30)

$$V = \frac{s}{\bar{x}} \cdot 100$$

Código em R

Utiliza-se o desvio padrão (`sd_2_41`) e a média (`mean_2_29`) da Questão 2.29.

```
# Dados
sd_2_44b <- sd_2_41
mean_2_44b <- mean_2_29

# Coeficiente de variação
cv_2_44b <- (sd_2_44b / mean_2_44b) * 100
cv_2_44b
```

```
## [1] 26.14837
```

Interpretação

O coeficiente de variação das rendas familiares é aproximadamente 26.15%, indicando que o desvio padrão representa cerca de 26.15% da média.

c. Comparação da Dispersão

Interpretação

O coeficiente de variação das rendas familiares (26.15%) é significativamente maior que o dos preços de gasolina (6.02%), indicando que a distribuição de rendas familiares apresenta maior dispersão relativa em comparação com a distribuição de preços de gasolina.

1.4 SHAPE OF FREQUENCY DISTRIBUTIONS

1.4.0.1 ***Questão 2.45*** Find the Pearson coefficient of skewness for the data in (a) Table 2.29 and (b) Table 2.30.

1.4.0.1.1 ***Ans. 2.45***

a. Coeficiente de Assimetria de Pearson (Tabela 2.29)

O coeficiente de assimetria de Pearson mede a assimetria de uma distribuição com base na diferença entre a média e a moda, normalizada pelo desvio padrão. Segundo Dominick & Reagle (2002), a fórmula é:

$$\text{Sk}_P = \frac{\bar{x} - \text{Moda}}{s}$$

Código em R

> Utiliza-se a média (`mean_2_28`) e o desvio padrão (`sd_2_40`) da Questão 2.28 e 2.40, respectivamente, e a moda (`mode_2_28`) da Questão 2.28.

```r
# Dados de tabela_2_29 (definidos na Questão 2.26)
mean_2_45a <- mean_2_28
mode_2_45a <- mode_2_28
sd_2_45a <- sd_2_40

# Coeficiente de assimetria de Pearson
skp_2_45a <- (mean_2_45a - mode_2_45a) / sd_2_45a
skp_2_45a
```

```
## [1] -0.2399041
```

Interpretação

O coeficiente de assimetria de Pearson para os preços de gasolina é aproximadamente -0.24, indicando uma assimetria negativa leve, com a cauda da distribuição ligeiramente mais longa à esquerda (preços menores).

b. Coeficiente de Assimetria de Pearson (Tabela 2.30)

$$\mathrm{Sk}_P = \frac{\bar{x} - \mathrm{Moda}}{s}$$

Código em R

> Utiliza-se a média (`mean_2_29`) e o desvio padrão (`sd_2_41`) da Questão 2.29 e 2.41, respectivamente, e a moda (`mode_2_29`) da Questão 2.29.

```r
# Dados de tabela_2_30 (definidos na Questão 2.27)
mean_2_45b <- mean_2_29
mode_2_45b <- mode_2_29
sd_2_45b <- sd_2_41

# Coeficiente de assimetria de Pearson
skp_2_45b <- (mean_2_45b - mode_2_45b) / sd_2_45b
skp_2_45b
```

```
## [1] 0.4499213
```

Interpretação

O coeficiente de assimetria de Pearson para as rendas familiares é aproximadamente 0.45, indicando uma assimetria positiva leve, com a cauda da distribuição ligeiramente mais longa à direita (rendas maiores).

1.4.0.2 ***Questão 2.46*** Find the coefficient of skewness using the formula based on the third moment for the data in (a) Table 2.29 and (b) Table 2.30.

1.4.0.2.1 ***Ans. 2.46***

a. Coeficiente de Assimetria Baseado no Terceiro Momento (Tabela 2.29)

O coeficiente de assimetria baseado no terceiro momento mede a assimetria usando o momento central de ordem 3, normalizado pelo cubo do desvio padrão. Segundo Dominick & Reagle (2002), a fórmula é:

$$\text{Sk} = \frac{\sum f_i \cdot (x_i - \bar{x})^3 / \sum f_i}{s^3}$$

Código em R

> Utiliza-se `tabela_2_29`, `questao_2_26_midpoints`, a média (`mean_2_28`) e o desvio padrão (`sd_2_40`) da Questão 2.28 e 2.40.

```r
# Dados
midpoints_2_46a <- questao_2_26_midpoints
freq_2_46a <- tabela_2_29$freq
mean_2_46a <- mean_2_28
sd_2_46a <- sd_2_40
n_2_46a <- sum(freq_2_46a)

# Terceiro momento central
m3_2_46a <- sum(freq_2_46a * (midpoints_2_46a - mean_2_46a)^3) / n_2_46a

# Coeficiente de assimetria
sk_2_46a <- m3_2_46a / (sd_2_46a^3)
sk_2_46a
```

```
## [1] -0.1829484
```

Interpretação

O coeficiente de assimetria baseado no terceiro momento para os preços de gasolina é aproximadamente -0.18, confirmando uma assimetria negativa leve, com uma cauda ligeiramente mais longa à esquerda.

b. Coeficiente de Assimetria Baseado no Terceiro Momento (Tabela 2.30)

$$\text{Sk} = \frac{\sum f_i \cdot (x_i - \bar{x})^3 / \sum f_i}{s^3}$$

Código em R

> Utiliza-se `tabela_2_30`, `questao_2_27_midpoints`, a média (`mean_2_29`) e o desvio padrão (`sd_2_41`) da Questão 2.29 e 2.41.

```r
# Dados
midpoints_2_46b <- questao_2_27_midpoints
freq_2_46b <- tabela_2_30$freq
mean_2_46b <- mean_2_29
sd_2_46b <- sd_2_41
n_2_46b <- sum(freq_2_46b)

# Terceiro momento central
```

```r
m3_2_46b <- sum(freq_2_46b * (midpoints_2_46b - mean_2_46b)^3) / n_2_46b

# Coeficiente de assimetria
sk_2_46b <- m3_2_46b / (sd_2_46b^3)
sk_2_46b
```

```
## [1] 0.7595837
```

Interpretação

O coeficiente de assimetria baseado no terceiro momento para as rendas familiares é aproximadamente 0.76, indicando uma assimetria positiva moderada, com uma cauda mais longa à direita.

1.4.0.3 *Questão 2.47* Find the coefficient of kurtosis for the data in (a) Table 2.29 and (b) Table 2.30.

1.4.0.3.1 *Ans. 2.47*

a. Coeficiente de Curtose (Tabela 2.29)

O coeficiente de curtose mede o achatamento de uma distribuição em relação a uma distribuição normal. Segundo Dominick & Reagle (2002), a fórmula é:

$$\mathrm{K} = \frac{\sum f_i \cdot (x_i - \bar{x})^4 / \sum f_i}{s^4}$$

Código em R

> Utiliza-se `tabela_2_29`, `questao_2_26_midpoints`, a média (`mean_2_28`) e o desvio padrão (`sd_2_40`) da Questão 2.28 e 2.40.

```r
# Dados
midpoints_2_47a <- questao_2_26_midpoints
freq_2_47a <- tabela_2_29$freq
mean_2_47a <- mean_2_28
sd_2_47a <- sd_2_40
n_2_47a <- sum(freq_2_47a)

# Quarto momento central
m4_2_47a <- sum(freq_2_47a * (midpoints_2_47a - mean_2_47a)^4) / n_2_47a

# Coeficiente de curtose
k_2_47a <- m4_2_47a / (sd_2_47a^4)
k_2_47a
```

```
## [1] 2.371306
```

Interpretação

O coeficiente de curtose para os preços de gasolina é aproximadamente 2.4, indicando uma distribuição ligeiramente platicúrtica (menos achatada que a normal, com caudas mais leves), pois valores próximos de 3 indicam curtose semelhante à normal.

b. Coeficiente de Curtose (Tabela 2.30)

$$\text{K} = \frac{\sum f_i \cdot (x_i - \bar{x})^4 / \sum f_i}{s^4}$$

Código em R

> Utiliza-se `tabela_2_30`, `questao_2_27_midpoints`, a média (`mean_2_29`) e o desvio padrão (`sd_2_41`) da Questão 2.29 e 2.41.

```r
# Dados
midpoints_2_47b <- questao_2_27_midpoints
freq_2_47b <- tabela_2_30$freq
mean_2_47b <- mean_2_29
sd_2_47b <- sd_2_41
n_2_47b <- sum(freq_2_47b)

# Quarto momento central
m4_2_47b <- sum(freq_2_47b * (midpoints_2_47b - mean_2_47b)^4) / n_2_47b

# Coeficiente de curtose
k_2_47b <- m4_2_47b / (sd_2_47b^4)
k_2_47b
```

```
## [1] 3.000377
```

Interpretação

O coeficiente de curtose para as rendas familiares é aproximadamente 3, indicando uma distribuição ligeiramente leptocúrtica (mais achatada que a normal, com caudas mais pesadas), pois valores acima de 3 sugerem maior concentração em torno da média.

1.4.0.4 ***Questão 2.48*** For covariance, (a) in what range should the covariance for directly related data fall? (b) for inversely related data? (c) for unrelated data?

1.4.0.4.1 ***Ans. 2.48*** A covariância mede a relação linear entre duas variáveis. Segundo Dominick & Reagle (2002), a covariância é definida como:

$$\text{Cov}(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

a. Intervalo para Dados Diretamente Relacionados

Interpretação

Para dados diretamente relacionados (quando X aumenta, Y também aumenta), a covariância é positiva ($\text{Cov}(X,Y) > 0$), pois os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ tendem a ter o mesmo sinal.

b. Intervalo para Dados Inversamente Relacionados

Interpretação

Para dados inversamente relacionados (quando X aumenta, Y diminui), a covariância é negativa ($\text{Cov}(X, Y) < 0$), pois os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ tendem a ter sinais opostos.

c. Intervalo para Dados Não Relacionados

Interpretação

Para dados não relacionados (sem relação linear), a covariância é próxima de zero ($\text{Cov}(X, Y) \approx 0$), pois os desvios $(x_i - \bar{x})$ e $(y_i - \bar{y})$ não apresentam um padrão consistente de sinal.
