MAE 5905: Introdução à Ciência de Dados Gabarito Lista 1 - Primeiro Semestre de 2025

Nhancololo, A. M.

2025-04-04

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(dplyr,tidyverse,xtable,kableExtra,knitr)
```

Exercício 1

Num conjunto de dados, o primeiro quartil é 10, a mediana é 15 e o terceiro quartil é 20.

Seja $X = \{2.1, 7.2, 10, 11.1, 14, 16, 19, 20, 21.1, 23\}$ o conjunto de dados para uma variável aleatória contínua, ou $Y = \{2, 7, 10, 11, 14, 16, 19, 20, 21, 23\}$ para uma variável aleatória discreta.

Nota: Utilize a Fórmula 3.5 de Morettin e Singer (2023, p. 48). O software R aproxima os números para inteiros e aplica a versão discreta da Fórmula 3.5.

(a) A distância interquartil é 5.

FALSO

A distância interquartil é calculada pela fórmula:

$d_Q = Q_3 - Q_1 = 20 - 10 = 10$ (Morettin e Singer, 2023, p. 50, ponto 3.12)

(b) O valor 32 seria considerado outlier segundo o critério utilizado na construção do boxplot.

FALSO, resposta baseada no critério (2) abaixo.

Possíveis respostas aceitáveis mediante justificativa

(1) O critério utilizado na construção do boxplot considera X_i um outlier se

$X_i > \min[x_{(n)}, Q_3 + 1.5 \times d_Q]$ ou $X_i < \max[x_{(1)}, Q_1 - 1.5 \times d_Q]$

onde $x_{(1)}$ e $x_{(n)}$ são, respectivamente, o valor mínimo e máximo do conjunto de dados (Morettin & Singer, 2023, p. 54, ponto 3.4, e R/RStudio, objeto *out*, do comando `boxplot.stats` na saída abaixo, onde $d_Q = \text{iqr}$ e $\text{coef} = 1.5$).

$Q_1 - 1.5 \times d_Q = 10 - 1.5 \times 10 = -5$, $\lim_{\text{inf}} = \max[x_{(1)}, Q_1 - 1.5 \times d_Q] = \max[x_{(1)}, -5] > 32$?, considerando a descrição do exercício $(Q_1, Q_2 = \text{md}, Q_3)$, **não**.

$Q_3 + 1.5 \times d_Q = 20 + 1.5 \times 10 = 35$, $\lim_{\text{sup}} = \min[x_{(n)}, Q_3 + 1.5 \times d_Q] = \min[x_{(n)}, 35] < 32$? **Depende**.

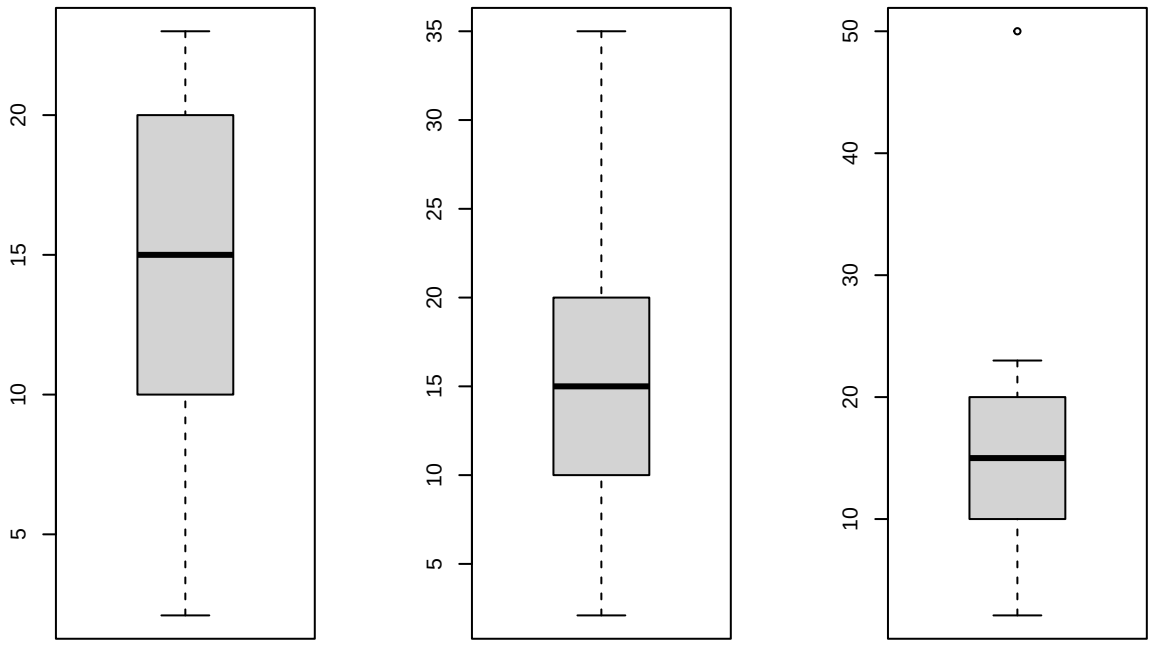
Como o primeiro quartil é 10, temos certeza de que, se 32 for um outlier, será um outlier superior. Usando o critério (1) para a construção do boxplot, para decidir se 32 é outlier ou não, **vai depender do valor de** $x_{(n)}$.

(2) O valor 32 **não é um outlier**, pois está dentro dos limites inferior ($Q_1 - 1.5 \times d_Q = -5$) e superior ($Q_3 + 1.5 \times d_Q = 35$), que são padrão em várias literaturas (Morettin & Singer, 2023, p. 55, Figura 3.16 e diversas outras literaturas).

```
X=c(2.1,7.2,10,11.1,14,16,19,20,21.1,23)
X1=c(2.1,7.2,10,11.1,14,16,19,20,23,35)
X2=c(2.1,7.2,10,11.1,14,16,19,20,23,50)

par(mfrow=c(1,3))
boxplot(X,main=expression("Figura 1: Dados com"~X[n]==23)) ##Aqui 32 seria outlier por (1)
boxplot(X1, main=expression("Figura 2: Dados com"~X[n]==35)) #32 não seria outlier por (1)
boxplot(X2, main=expression("Figura 3:Dados com"~X[n]==50))
```

Figura 1: Dados com $X_n = 23$ Figura 2: Dados com $X_n = 35$ Figura 3:Dados com $X_n = 50$




```
#32 não seria outlier por (1) mas no gráfico mostra que é outlier.
#summary(X);
#summary(X1);
#summary(X2)
#a <- 19.75+1.5*IQR(X);a
#a1<- 19.75+1.5*IQR(X1);a1
#a2<- 19.75 +1.5*IQR(X1);a2

df <- data.frame(
  Statistic = c("Min.", "1st Qu.", "Median", "Mean", "3rd Qu.", "Max.","3rd + 1.5 * IQR(.)"),
  X  = c(2.10, 10.28, 15.00, 14.35, 19.75, 23.00, 33.9625),
  X1 = c(2.10, 10.28, 15.00, 15.74, 19.75, 35.00, 33.9625),
  X2 = c(2.10, 10.28, 15.00, 17.24, 19.75, 50.00, 33.9625)
)

kable(df, booktabs = TRUE)
```

Statistic	X	X1	X2
Min.	2.1000	2.1000	2.1000
1st Qu.	10.2800	10.2800	10.2800
Median	15.0000	15.0000	15.0000
Mean	14.3500	15.7400	17.2400
3rd Qu.	19.7500	19.7500	19.7500
Max.	23.0000	35.0000	50.0000
3rd + 1.5 * IQR(.)	33.9625	33.9625	33.9625

O gráfico mostra que, em caso de presença de outliers nos dados, o R ajusta os *whiskers* (bigodes) para o n-ésimo valor mais próximo so conjunto dados que não é outlier.

(c) A mediana ficaria alterada de 2 unidades se um ponto com valor acima do terceiro quartil fosse substituído por outro 2 vezes maior.

Resolução

FALSO

```
# Alternativamente
X=c(2.1,7.2,10,11.1,14,16,19,20,21.1,23)
Y=c(2,7,10,11,14,16,19,20,21,23)
summary(Y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   10.25   15.00   14.30   19.75   23.00
```

Só lembrando!!

A mediana med(X) de um conjunto ordenado de n elementos $x_1, x_2, \ldots, x_n$ pode ser definida por:

$$md(x) = \begin{cases} x_{(\frac{n+1}{2})}, \text{ese } n \text{ for ímpar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, \text{ese } n \text{ for par} \end{cases}$$

onde x_i representa os valores ordenados do conjunto (ver Morettin e Singer, 2023, p. 47, ponto 3.3) .

Justificativa

(1) O exercício fala sobre substituir um ponto e não aumentar a quantidade de pontos.

(2) A mediana não é influenciada por valores extremos (**ver Figuras 1, 2 e 3**). Nas Figuras 1,2,3 a substituição do 23 por 35 e 50 não alterou a mediana.

(3) O valor a ser retirado não é um ponto de massa (ponto que divide os dados ao meio) e muito menos próximo a este, o que implica que a mediana não seria alterada.

(d) O valor mínimo é maior do que zero.

Resolução

FALSO

```
Xj= c(2,7,9.67,11,14,16,19,20.34,21,23)
summary(Xj)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0    10.0    15.0    14.3    20.0    23.0
```

```r
IQR(Xj) #Distancia interquantílica
```

```
[1] 10.0025
```

```r
Xk= c(-2,7,9.67,11,14,16,19,20.34,21,23)
summary(Xk)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -2.0    10.0    15.0    13.9    20.0    23.0
```

```r
IQR(Xk)
```

```
[1] 10.0025
```

As duas listas de dados satisfazem o descrito no exercício, mas o valor mínimo **não é necessariamente** maior que zero.

Exercício 2

Resolução

Alternativa correta: (B)

Tabela de Frequências e Frequências Acumuladas

Número de vasos	Frequência absoluta	Frequência acumulada (%)
$0-5$	8(12%)	12%
$5-10$	23(35%)	47%
$10-15$	12(18%)	65%
$15-20$	9(14%)	79%
$20-25$	8(12%)	91%
$25-30$	6(9%)	100%

Justificativa

A mediana é o valor que divide os dados em duas partes de igual frequência acumulada, ou seja, 50%.

Observando a tabela:
1) A classe **5 - 10** acumula **47**% dos dados.
2) A classe **10 - 15** acumula **65**% dos dados.

Como **50**% está entre 47% e 65%, a mediana pertence à classe $\mathbf{10-15}$.

Portanto, concluímos que a alternativa correta é **(B)**.

Fórmula caso seja de interesse

Para calcular quantis (quartis, decis e percentis) em uma distribuição de frequências, pode-se utilizar a fórmula:

$$Q(p) = Lim_{inf} + \left(\frac{p \times n - F}{f}\right) \times h$$

onde **Q(p)** é quantil desejado. **p** é Posição relativa/ordem do quantil. Lim_{inf} é o limite inferior da classe onde o quantil se encontra. **n** é número total de observações (soma das frequências absolutas). **F** é afrequência acumulada da classe anterior à classe do quantil. **f** é frequência absoluta da classe do quantil e **h** é amplitude da classe (diferença entre os limites inferior e superior de um intervalo).

Exercício 3

Resolução

(a)

Como os grupos têm médias diferentes, seria errado avaliar a variabilidade com base no desvio padrão. Assim, uma alternativa seria (1) usar o coeficiente de variação, que é uma estatística útil para comparar a variação nos dados em que as unidades de medida diferem ou para os quais as médias diferem substancialmente (Kvalseth, 2017, p. 403, eq. 4 e o parágrafo seguinte). Outra seria (2) comparar as distâncias interquartílicas, como existe representação gráfica dos da distribuição dos dados (Figura 1).

$CV = \frac{S}{\bar{X}} \times 100$

Grupo VO2MAX

$\text{CV}_{\text{Normais}} = 795 \cdot 100/1845 = 43.09\%$

$\text{CV}_{\text{Cardiopatas}} = 434 \cdot 100/1065 = 40.75\%$

$\text{CV}_{\text{DPOC}} = 381 \cdot 100/1065 = 35.78\%$

Grupo VCO2MAX

$\text{CV}_{\text{Normais}} = 918 \cdot 100/2020 = 45.45\%$

$\text{CV}_{\text{Cardiopatas}} = 479 \cdot 100/1206 = 39.72\%$

$\text{CV}_{\text{DPOC}} = 430 \cdot 100/934 = 46.04\%$

Com base nos resultados acima, para o grupo VCO2MAX, o grupo Normais (N) apresenta maior variabilidade, pois possui maior coeficiente de variação. Para o grupo VO2MAX, o grupo DPOC (D) apresenta maior variabilidade, também por apresentar maior coeficiente de variação, conforme o critério (1). Pelo critério (2), como existem outliers e a média, o desvio padrão e o coeficiente de variação são influenciados por valores extremos, considerariam-se os dados dos Normais, em ambos os casos, como os de maior variabilidade pois tem maior distência interquantílica.

(b)

Tanto a média quanto a mediana diminuem nos cardiopatas e nos portadores de DPOC em relação aos grupos normais, sendo que o grupo DPOC ainda apresenta menor média e mediana que o grupo de cardiopatas. Ou seja, pessoas com cardiopatia apresentam menor média e mediana de consumo máximo de O2 e de CO2 em relação ao grupo normal. Além disso, portadores de doença pulmonar obstrutiva crônica apresentam menor média e mediana de consumo máximo de O2 e de CO2 em relação aos cardiopatas.

(c)

Distância interquartílica: Visualmente (Figura 1), tanto em VCO2MAX quanto em VO2MAX, os grupos normais apresentam maior distância interquartílica em relação aos grupos C e D, que, por sua vez, apresentam distância interquartílica aproximadamente igual em ambos os casos.

Uso da distribuição normal: Sem querer ser tão exaustivo, diríamos que seria razoável o uso da distribuição normal nos dados, pois a mediana está "centrada" entre os dois quartis (Q1, Q3). No entanto, sendo mais criterioso, no VCO2MAX, os grupos N e C apresentam a mediana não tão centralizada, o que sugere uma leve assimetria à direita. Nesses casos, a distribuição normal não seria a mais adequada.

(d)

A quantidade de asteriscos representa o número de pessoas que apresentaram esforço cardiopulmonar (doença pulmonar) acima do limite máximo ($\lim_{sup}$) aceitável, conforme definido no Exercício 1. O número

correspondente a cada asterisco representa o valor estimado pelo teste/instrumento utilizado para medir o esforço cardiopulmonar.

Resposta aceitável: Os asteriscos indicam valores atípicos ou discrepantes (outliers).

(e)

Se conseguiu estimar as medidas de tendência central (média, mediana) e de variabilidade (desvio padrão), isso significa que:
(1) A variável dependente (número de esforços cardiopulmonares) é uma variável quantitativa. Assim, tanto o modelo linear (OLS), caso satisfaça as condições de normalidade e homocedasticidade (igualdade de variâncias), quanto o modelo linear ponderado (WLS), para corrigir o impacto de outliers, seriam adequados.

Dessa forma, a relação pode ser expressa como: $y_{ij} = \beta_0 + \beta_1 \times x_{ij}$ onde x_{ij} é a observação do indivíduo i no grupo j, e y_{ij} representa o esforço cardiovascular medido em consumo máximo de O2 e CO2, ambos em ml/min.

(f)

Seria importante verificar se os outliers presentes nos boxplots (Figura 1) ou no gráfico de dispersão podem ser resultado de erros de codificação (digitação) ou de amostragem. Essa análise ajudaria a determinar se esses valores extremos refletem variabilidade real nos dados ou se devem ser corrigidos ou removidos para evitar distorções na modelagem e inferência.

Exercício 4

O gráfico QQ corresponde ao ajuste de um modelo de regressão linear múltipla.

(a)

Alternativa correta: (D)

Possível justificativa: Os resíduos desviam-se nas caudas (extremidades), ficando fora dos limites mínimos e máximos (bandas de confiança) para quantis abaixo de -1 e acima de 1, evidenciando que os resíduos padronizados não são normalmente distribuídos. Como o desvio ocorre em ambos os extremos, entre -1 e 1, isso remete à simetria.

Em síntese, trata-se de uma distribuição com caudas mais pesadas do que a distribuição normal padrão, pois os resíduos inferiores apresentam valores mais baixos do que os respectivos quantis de uma distribuição normal padrão, enquanto os resíduos superiores apresentam valores maiores do que os respectivos quantis dessa mesma distribuição.

Exercício 5

$\log\left\{\frac{\pi_i(x_i,w_i)}{1-\pi_i(x_i,w_i)}\right\} = \alpha + \beta x_i + \gamma(w_i - 5)$, onde x_i representa o gênero da i-ésima criança ($x_i = 1$ para masculino e $x_i = 0$ para feminino), e w_i é a idade da i-ésima criança.

Antes de interpretar cada parâmetro, considerando um nível de significância de 5% (correspondente a um intervalo de confiança de 95%), podemos afirmar que, como o valor-p é menor que 5% para cada parâmetro (α, β e γ), a idade (w_i) e o gênero (x_i) são variáveis importantes para determinar a preferência pelo refrigerante. Consequentemente, os intervalos de confiança (IC) estimados para os parâmetros α, β e γ não incluirão o valor zero, pois ao testar esses parâmetros, estamos verificando se são estatisticamente diferentes de zero. Se o valor-p fosse maior que 5%, não teríamos evidências suficientes para rejeitar a hipótese nula de que os parâmetros são iguais a zero, e, nesse caso, o IC incluiria o valor zero.

(a) Interpretar de cada parâmetro

$\Longrightarrow$ Para α, considere criança do género feminimo $x_i = 0$ e idade $w_i = 5$ anos:

$$\log\left\{\frac{\pi_i(0,5)}{1-\pi_i(0,5)}\right\} = \alpha + \beta \cdot 0 + \gamma(5-5) = \alpha$$
$$\log\left\{\frac{\pi_i(0,5)}{1-\pi_i(0,5)}\right\} = \alpha$$
$$\frac{\pi_i(0,5)}{1-\pi_i(0,5)} = \exp(\alpha)$$
$$\frac{\pi_i(0,5)}{1-\pi_i(0,5)} = \exp(0.69)$$
$$\frac{\pi_i(0,5)}{1-\pi_i(0,5)} \approx 1.99$$

.

Assim, $\alpha = 0,69$ é o logaritmo da chance de uma criança de 5 anos, do gênero feminino, preferir refrigerante Kcola e, $\exp(0.69) = 1.99$ é a chance de uma criança de 5 anos, do gênero feminino, preferir refrigerante Kcola.

$\Longrightarrow$ Para β, considere $x_i = 1$ ou $x_j = 0$ com w fixo:

$$\log\left\{\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right\} = \log\left[\frac{\pi_i(1,w)}{1-\pi_i(1,w)}\right] - \log\left[\frac{\pi_j(0,w)}{1-\pi_j(0,w)}\right]$$
$$\log\left\{\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right\} = \alpha + \beta \cdot 1 + \gamma(w-5) - [\alpha + \beta \cdot 0 + \gamma(w-5)]$$
$$\log\left\{\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right\} = \alpha + \beta \cdot 1 + \gamma(w-5) - \alpha - \beta \cdot 0 - \gamma(w-5)]$$
$$\log\left\{\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right\} = \beta \quad (*)$$
$$\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}} = \exp(\beta) = \exp(0,33) \approx 1.39 \quad (**)$$
$$\frac{\pi_i(1,w)}{1-\pi_i(1,w)} = 1.39\left[\frac{\pi_j(0,w)}{1-\pi_j(0,w)}\right](***)$$

.

Assim, $\beta = 0,33$ $(*)$ representa o logaritmo da razão de chances de preferir o refrigerante Kcola entre crianças do sexo masculino e do sexo feminino, mantendo a idade fixa. $(**)$ representa a razão de chances de preferir o refrigerante Kcola entre crianças do sexo masculino e do sexo feminino, mantendo a idade fixa. $(***)$ indica que, fixada a idade (w), a chance de uma criança do gênero masculino preferir o refrigerante Kcola é 1,39 vezes a chance de uma criança do gênero feminino preferir Kcola.

$\Longrightarrow$ Para γ, considere as idades $w_i = 6$ e $w_j = 5$ com género $x_i = x$ fixo:

$$\log\left\{\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right\} = \log\left(\frac{\pi_i(x,6)}{1-\pi_i(x,6)}\right) - \log\left(\frac{\pi_j(x,5)}{1-\pi_j(x,5)}\right)$$

$$\log\left\{\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right\} = \alpha + \beta x + \gamma(6-5) - [\alpha + \beta x + \gamma(5-5)]$$

$$\log\left\{\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right\} = \alpha + \beta x + \gamma(6-5) - \alpha - \beta x - \gamma(5-5)$$

$$\log\left\{\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right\} = \gamma \quad (a)$$

$$\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}} = \exp(\gamma) = \exp(-0,03) \approx 0.97 \quad (b)$$

$$\frac{\pi_i(x,6)}{1-\pi_i(x,6)} = 0.97 \times \frac{\pi_j(x,5)}{1-\pi_j(x,5)} \quad (c)$$

.

Assim, $\gamma = -0{,}03$ (a) representa o logaritmo da razão de chances de preferência pelo refrigerante Kcola entre crianças de 5 e 6 anos de idade, com o sexo fixado. (b) representa a razão de chances de preferência pelo refrigerante Kcola entre crianças de 5 e 6 anos de idade, com o sexo fixado. (c) informa que a chance de uma criança de 6 anos de idade preferir o refrigerante Kcola, independentemente do gênero, é 0,97 vezes a chance de uma criança de 5 anos de idade preferir o refrigerante Kcola.

Caso geral: Interpretação do parâmetro γ.

Seja idade $w_j = w_i - 1$, com $w_i \geq 6$, e género $x_i = x_j = x$ fixo:

$$\log\left\{\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right\} = \log\left[\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}\right] - \log\left[\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}\right]$$

$$\log\left\{\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right\} = \alpha + \beta \cdot x + \gamma(w_i-5) - [\alpha + \beta \cdot x + \gamma(w_i-1-5)]$$

$$\log\left\{\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right\} = \alpha + \beta \cdot x + \gamma(w_i-5) - \alpha - \beta \cdot x - \gamma(w_i-1-5)$$

$$\log\left\{\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right\} = \gamma(w_i-5) - \gamma(w_i-5) + \gamma(1)$$

$$\log\left\{\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right\} = \gamma = -0.03 \quad (*)$$

$$\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}} = \exp(\gamma) = \exp(-0.03) = 0.97 \quad (**)$$

$$\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)} = 0.97 \times \frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)} \quad (***)$$

Assim, $\gamma = -0{,}03$ $(*)$ representa o logaritmo da razão de chances de preferência por Kcola entre duas crianças do mesmo gênero, mas que diferem na idade em apenas um ano, sendo, respectivamente, a primeira mais velha

e a segunda mais nova. $(**)$ representa a razão de chances de preferência por Kcola entre duas crianças do mesmo gênero, mas que diferem na idade em apenas um ano, sendo, respectivamente, a primeira mais velha e a segunda mais nova. $(***)$ informa que a chance de uma criança mais velha um ano, independentemente do gênero, preferir Kcola é 0,97 vezes a chance de uma criança um ano mais nova, também independentemente do gênero, preferir Kcola.

Nota importante: Dizer que a razão de chances entre duas crianças aumenta em $\exp(\gamma)$ a cada ano, fixado o gênero, remete à ideia de uma relação linear. Ou seja, se a diferença de idade entre as duas crianças é Δw_{ij}, então a razão de chances correta seria $\exp(\Delta w_{ij} \cdot \gamma)$, e **não** $\Delta w_{ij} \cdot \exp(\gamma)$, pois esta última forma está incorreta.

b)

A estimativa da razão de chances de preferência por Kcola correspondente à comparação de crianças do mesmo gênero x_i, com idades de 10 e 15 anos é dada por:

$$\log\left(\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right) = \log\left(\frac{\pi_i(x,10)}{1-\pi_i(x,10)}\right) - \log\left(\frac{\pi_j(x,15)}{1-\pi_j(x,15)}\right)$$

$$\log\left(\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right) = \alpha + \beta x + \gamma(10-5) - [\alpha + \beta x + \gamma(15-5)]$$

$$\log\left(\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right) = \alpha + \beta x + 5\gamma - \alpha - \beta x - 10\gamma$$

$$\log\left(\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right) = -5\gamma$$

$$\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}} = \exp(-5\gamma) = \exp(-5 \times -0.03) = \exp(0.15) \approx 1.1618$$

c)

$\Longrightarrow$ Intervalo de confiança 95% para β:

$\text{IC}_{95\%} = \exp(\hat{\beta} \pm z_{95\%} \times s_{\hat{\beta}})$, em que $s_{\hat{\beta}}$ é o erro padrão de β.

$\text{IC}_{95\%} = \exp(0,33 \pm 1,96 \times 0,10) = [1,14; 1,69]$, que, como afirmado acima não inclui zero.

Interpretação: Com 95% de confiança, o intervalo [1,14; 1,69] contém a verdadeira razão de chances de preferência por Kcola entre uma criança do sexo masculino e uma do sexo feminino, fixadas as idades. Ou seja, se o estudo fosse repetido várias vezes, utilizando o mesmo procedimento de amostragem, 95% dos intervalos de confiança construídos conteriam a verdadeira razão de chances de preferência por Kcola entre crianças do sexo masculino e do sexo feminino, com idades iguais.

$\Longrightarrow$ Intervalo de Confiança 95% para γ:

$\text{IC}_{95\%} = \exp(\hat{\gamma} \pm z_{95\%} \times s_{\hat{\gamma}})$, em que $s_{\hat{\gamma}}$ é o erro padrão.

$\text{IC}_{95\%} = \exp(-0,03 \pm 1,96 \times 0,005) = [0,96; 0,98]$.

Interpretação: Com 95% de confiança, o intervalo [0,96; 0,98] contém a verdadeira razão de chances de preferir Kcola entre duas crianças do mesmo sexo, mas que diferem na idade em apenas um ano, sendo, respectivamente, a primeira mais velha e a segunda mais nova.
Ou seja, se o estudo fosse realizado várias vezes, com o mesmo procedimento de amostragem, 95% dos intervalos de confiança construídos conteriam a verdadeira razão de chances de preferir Kcola entre duas

crianças do mesmo sexo, com diferença de um ano de idade, sendo a primeira mais velha e a segunda mais nova.

d)

A probabilidade de uma criança do sexo x_i e idade w_i preferir k-cola, pode ser estimada por $\pi_i(x_i, w_i)$, que é obtido da seguinte forma:

$$\begin{aligned}
\log\left(\frac{\pi_i(x_i,w_i)}{1-\pi_i(x_i,w_i)}\right) &= \alpha + \beta x_i + \gamma(w_i - 5) \\
\frac{\pi_i(x_i,w_i)}{1-\pi_i(x_i,w_i)} &= \exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right) \\
\pi_i(x_i,w_i) &= [1-\pi_i(x_i,w_i)]\exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right) \\
\pi_i(x_i,w_i) &= \exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right) - \pi_i(x_i,w_i)\exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right) \\
\pi_i(x_i,w_i) + \pi_i(x_i,w_i)\exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right) &= \exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right) \\
\pi_i(x_i,w_i)\left[1 + \exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right)\right] &= \exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right) \\
\pi_i(x_i,w_i) &= \frac{\exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right)}{1 + \exp\left(\alpha + \beta x_i + \gamma(w_i - 5)\right)} \quad (*)
\end{aligned}$$

Assim, pela expressão $(*)$, a probabilidade de meninos com 15 anos preferirem Kcola é dado por:

$\pi_i(1,15) = \frac{\exp[0{,}69+0{,}33\times 1-0{,}03(15-5)]}{1+\exp[0{,}69+0{,}33\times 1-0{,}03(15-5)]} = \frac{\exp(0{,}72)}{1+\exp(0{,}72)} = 0{,}67 \approx 67\%$.

Referências

Kvalseth, Tarald O. **Coefficient of variation: the second-order alternative.** Journal of Applied Statistics 44.3 (2017): 402-415.

Morettin, P. A. e Singer, J. M. **Estatística e Ciências de Dados**. Rio de Janeiro:LTC, 2023. (Já existe segunda edição do recente do livro, caso haja interressado pode comprar pelo link