Primeira Tarefa - Introdução a Probabilidade e Estatística II - 2024

Luigi Carvalho Euzebio

2024-09-01

Soluções dos exercícios 5, 6 e 7 da lista presentes na lista

Exercício 5.

Na linha de produção de uma grande montadora de veículos, existem 7 verificações do controle de qualidade. Sorteamos alguns dias do mês e anotamos o número de OKs recebidos pelos veículos produzidos nesses dias, i.e., em quantos dos controles mencionados o automóvel foi aprovado. Os resultados foram ((x, y), x =número de aprovações, y =frequência): (4, 126), (5, 359), (6, 1685), (7, 4764).

(i) Determine a média, moda e mediana do número de aprovações por automóvel produzido.

(ii) Calcule a variância da amostra.

(iii) Crie uma nova variável "reprovações", indicando o número de verificações não OKs no veículo. Determine média, moda, mediana e variância dessa variável.

(iv) Cada reprovação implica em custos adicionais para a montadora, tendo em vista a necessidade de corrigir o defeito apontado. Admitindo um valor básico de R$ 200,00 por cada item reprovado num veículo, calcule a média e a variância da despesa adicional por automóvel produzido.

Resolução (utilizando o R):

A ideia do código é criar uma função `calcular_medidas` que calcule a média, mediana, moda e variância com base nos dados que são inseridos. Neste código, foi utilizado a função `rep()` do R, para replicar valores, criando repetições de elementos - no caso, para conserguirmos calular os dados considerando as suas respectivas frequências.

Então, após criar a função, as informações necessárias, variáveis fornecidas pelo enunciado, denominadas de `aprov` (aprovações) e `reprov` (reprovações) em conjunto com as suas respectivas frequências. Vale destacar o uso da função criada chamada `moda` responsável por calcular a moda da amostra com base nos valores que mais aparecem no data frame.

Assim, quando chamamos a função `calcular_medidas`, recebemos as medias, modas, medianas e variâncias tanto das aprovações quanto das reprovações.

Por fim, é atribuido a variável `custo_por_reprovação` o valor de 200 e calculado a media e variância do custo, com base na quantidade de reprovações multiplicado pelo custo por reprovação, gerando uma nova variável: `custo_total_por_veiculo`

Observação: Quando calculamos `variancia_custo`, é permitido fazer esse calculo, pois o `custo_por_reprovação` é uma constante (200). Então, podemos "passar" para fora da variância, desde que a constante seja elevada ao quadrado.

```r
moda <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

calcular_medidas <- function(aprov, freq) {
  valores_repetidos <- rep(aprov, freq)

  media <- sum(aprov * freq) / sum(freq)
  moda <- moda(valores_repetidos)
  mediana <- median(valores_repetidos)
  variancia <- var(valores_repetidos)

  return(c(media = media, moda = moda, mediana = mediana, variancia = variancia))
}

aprov <- c(4, 5, 6, 7)
freq_aprov <- c(126, 359, 1685, 4764)

reprov <- c(0, 1, 2, 3)
freq_reprov <- c(4764, 1685, 359, 126)

resultados_aprov <- calcular_medidas(aprov, freq_aprov)
resultados_reprov <- calcular_medidas(reprov, freq_reprov)

cat("Resultados para aprovações:\n")
```

```
## Resultados para aprovações:
```

```r
print(resultados_aprov)
```

```
##     media      moda   mediana variancia
## 6.5989328 7.0000000 7.0000000 0.4528533
```

```r
cat("\nResultados para reprovações:\n")
```

```
##
## Resultados para reprovações:
```

```r
print(resultados_reprov)
```

```
##     media      moda   mediana variancia
## 0.4010672 0.0000000 0.0000000 0.4528533
```

```r
custo_por_reprovacao <- 200
custo_total_por_veiculo <- reprov * custo_por_reprovacao

media_custo <- sum(custo_total_por_veiculo * freq_reprov) / sum(freq_reprov)
variancia_custo <- custo_por_reprovacao^2 * resultados_reprov[['variancia']]

cat("\nMédia do custo adicional por veículo: (em R$)\n")
```

```
##
## Média do custo adicional por veículo: (em R$)
```

```
print(media_custo)
```

```
## [1] 80.21344
```

```
cat("\nVariância do custo adicional por veículo: (em R$)\n")
```

```
##
## Variância do custo adicional por veículo: (em R$)
```

```
print(variancia_custo)
```

```
## [1] 18114.13
```

Exercício 6.

Um hospital maternidade está planejando a ampliação dos leitos para recém nascidos. Um levantamento do número de dias que os bebes permaneceram no hospital forneceu os seguintes dados: 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 7, 7, e 8.Determine a tabela de frequência e calcule a média, moda e mediana destes dados.Determine o desvio padrão.Dentre as medidas de posição calculadas em (i), quais delas seriam mais apropriadas para resumir o conjunto de dados?

Resolução (Utilizando o R):

O código utiliza funções prontas que o próprio software R fornece aos usuários. Primeiro, inseri os dados fornecidos e os denominei de `permancencia_em_dias`. A partir disto, pude criar a `tabela_frequencia` e a converti para um data frame (chamado `tabela_df`), uma estrutura de dados para armazenar dados. Além disso, adicionei uma função que calcula a moda, a mesma utilizada no exercício anterior. Por fim, foram aplicadas as funções `mean`, `median` e `sd`, padrões do R.

```
permanencia_em_dias <- c(1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3

tabela_frequencia <- table(permanencia_em_dias)
tabela_df <- as.data.frame(tabela_frequencia)
names(tabela_df) <- c("Permanencia em dias", "Frequencias")

moda <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

media <- mean(permanencia_em_dias)
moda <- moda(permanencia_em_dias)
mediana <- median(permanencia_em_dias)
desvio_padrao <- sd(permanencia_em_dias)

cat("\nTabela de Frequência:\n")
```

```
##
## Tabela de Frequência:
```

```
print(tabela_df)
```

```
##   Permanencia em dias Frequencias
## 1                   1           3
## 2                   2          11
## 3                   3          15
## 4                   4           9
## 5                   5           6
## 6                   6           1
## 7                   7           2
## 8                   8           1
```

```
cat("\nMédia:", media, "\n")
```

```
##
## Média: 3.416667
```

```
cat("Mediana:", mediana, "\n")
```

```
## Mediana: 3
```

```
cat("Moda:", moda, "\n")
```

```
## Moda: 3
```

```
cat("Desvio Padrão:", desvio_padrao, "\n")
```

```
## Desvio Padrão: 1.555133
```

Resposta item (iii):

Neste caso, a mediana seria a medida de posição que resumiria melhor os dados exibidos anteriormente, pois, como temos diferenças consideráveis de dias — de 1 a 8 dias de permanência — não seria interessante utilizar a média, já que ela tende a ser mais influenciada por valores extremos presentes no conjunto. Além disso, como a mediana é igual à moda e também está próxima da média, temos um indicativo consistente do valor central dos dados.

Exercício 7.

Com os dados do Exercício 4:

i. Obtenha as medidas de posição e variabilidade para as variáveis Idade e Glicose (GL)

ii. Repita o item (i) para cada tipo de diagnóstico. Compare as respostas obtidas.

iii. Faça um boxplot para as idades dos pacientes falso positivos e outro para as idades dos falso negativos. Em base a estes gráficos, é possível dizer que existe diferença entre os dois grupos?

iv. Escolha os dados da AKP para o grupo 2 e o grupo 3. Faça um boxplot para cada conjunto de dados. Existe diferença entre estes grupos? Qual seria a importância de termos diferença neste caso?

Resolução (Utilizando o R):

A ideia deste código é a mesma dos anteriores. São utilizadas funções prórpias do R para calcular as medidas solicitadas e novamente a criação de uma função que calcula a moda. As únicas diferenças desse código para os demais são os filtros desenvolvidos para captar uma única coluna ou um grupo específico de dados que estão presente em uma coluna.

```
dt <- read.table("http://dcm.ffclrp.usp.br/~rrosales/aulas/cancer.txt", header=TRUE)

moda <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

media_idade <- mean(dt$Idade)
mediana_idade <- median(dt$Idade)
moda_idade <- moda(dt$Idade)
desvio_padrao_idade <- sd(dt$Idade)
variancia_idade <- var(dt$Idade)

media_gl <- mean(dt$GL)
mediana_gl <- median(dt$GL)
moda_gl <- moda(dt$GL)
desvio_padrao_gl <- sd(dt$GL)
variancia_gl <- var(dt$GL)

idade_medidas <- c("Média" = media_idade, "Mediana" = mediana_idade, "Moda" = moda_idade,"Desv Pad" = d

gl_medidas <- c("Média" = media_gl, "Mediana" = mediana_gl, "Moda" = moda_gl, "Desv Pad" = desvio_padra

cat("Medidas para Idade:\n")
```

```
## Medidas para Idade:
```

```
print(idade_medidas)
```

```
##     Média   Mediana      Moda  Desv Pad Variância
##  51.21271  54.00000  54.00000  19.11437 365.35906
```

```
cat("\n")
```

```
cat("Medidas para Glicose:\n")
```

```
## Medidas para Glicose:
```

```r
print(gl_medidas)
```

```
##     Média   Mediana      Moda  Desv Pad Variância
## 104.32320  99.00000  93.00000  24.81015 615.54345
```

```r
moda <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

grupos <- unique(dt$Grupo)

calc_medidas <- function(dados) {
  c(
    Media = mean(dados),
    Mediana = median(dados),
    Moda = moda(dados),
    Desv_Pad = sd(dados),
    Variância = var(dados)
  )
}

idade_medidas_df <- do.call(rbind, lapply(grupos, function(grupo) {
  dados_idade <- dt$Idade[dt$Grupo == grupo]
  calc_medidas(dados_idade)
}))

gl_medidas_df <- do.call(rbind, lapply(grupos, function(grupo) {
  dados_gl <- dt$GL[dt$Grupo == grupo]
  calc_medidas(dados_gl)
}))

idade_medidas_df
```

```
##         Media Mediana Moda Desv_Pad Variância
## [1,] 53.26786      55   63 18.80086  353.4724
## [2,] 45.67808      44   24 19.26653  371.1991
## [3,] 58.69474      60   59 16.77262  281.3207
## [4,] 50.93846      51   51 18.60517  346.1524
```

```r
cat("\n")
```

```r
gl_medidas_df
```

```
##         Media Mediana Moda Desv_Pad Variância
## [1,] 100.0357      96   93 16.11094  259.5623
## [2,] 100.9041      97   89 21.85069  477.4528
## [3,] 111.1684     105   97 27.24501  742.2905
## [4,] 105.6923     100   99 31.02918  962.8101
```

Resposta item (ii):

Em relação as medidas calculadas referentes as idades dos grupos, temos:

- A média e a mediana do grupo 3 (Positivo) são as mais altas em relação aos demais grupos, seguido pelo grupo 1 (Falso Negativo);
- Já em relação a moda, o grupo 2 (Negativo) tem esta medida mais baixa que os outros grupos;
- O Grupo 2 (Negativo) tem o maior desvio padrão e variância, o que sugere maior dispersão dos dados em relação à média. O Grupo 3 (Positivo) tem o menor desvio padrão e variância, indicando que as idades estão mais agrupadas em torno da média.

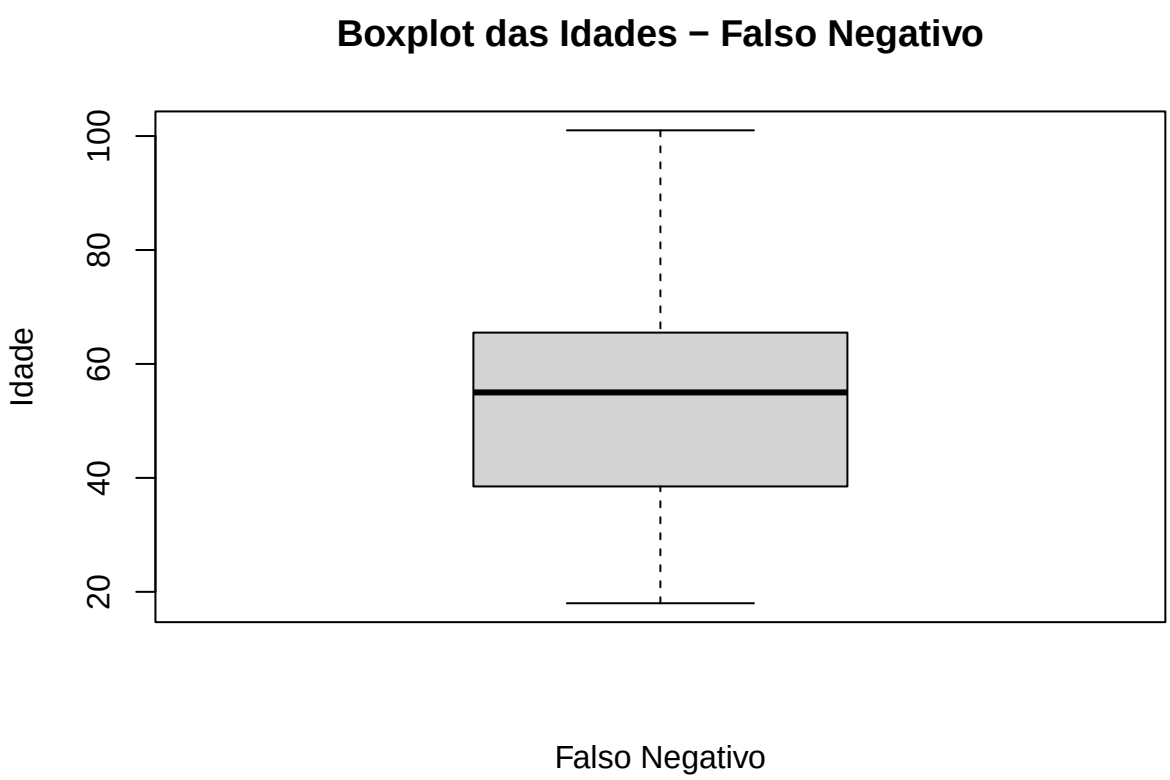
Já em relação a glicose, temos:

- O grupo 3 (Positivo) tem a maior media e mediana de glicose, indicando níveis de glicose mais altos quando comparados a outros grupos;
- O Grupo 4 (Falso Positivo) tem o maior desvio padrão e variância, o que sugere uma maior dispersão dos dados de glicose em relação à média. O Grupo 3 (Positivo) também apresenta alta variância e desvio padrão, indicando ampla variabilidade nos níveis de glicose.

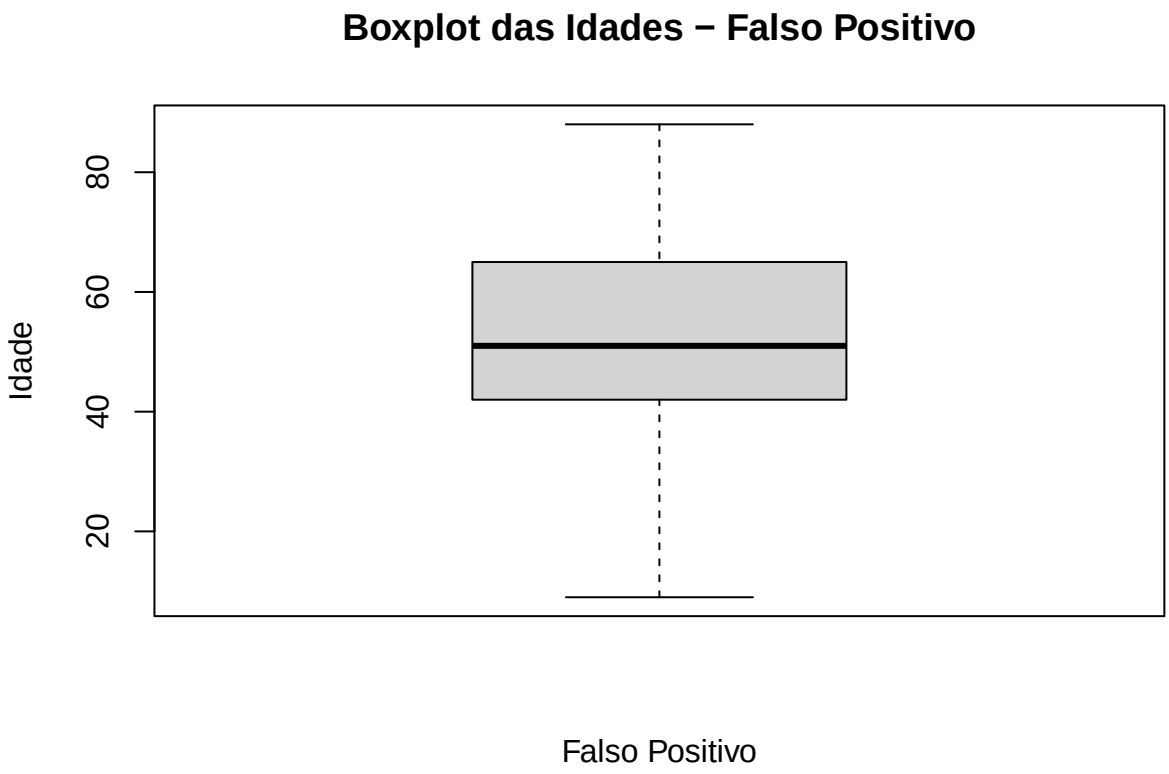
Pode-se concluir, que o Grupo 3 (Positivo) é o grupo com as médias de idade e de glicose mais elevadas em comparação com os outros grupos. Ademais, o Grupo 2 (Negativos) possuem médias de idade e de glicose mais baixas que os outros grupos, tendo apenas a média de glicose maior que o Grupo 1 (Falso Negativo). Esse último indicativo, pode ser um dos motivos que leva ao Grupo 1 a ser denominado de falso negativo, já que os níveis de glicose, tanto na média como na mediana são menores que 100, e por isso esse seja um possível fator que tenha mascarado exames e resultados em falso, porém eram pacientes que tinham cancer.

```
falso_negativo <- dt[dt$Grupo == 1, "Idade"]
falso_positivo <- dt[dt$Grupo == 4, "Idade"]

boxplot(falso_negativo,
        main = "Boxplot das Idades - Falso Negativo",
        xlab = "Falso Negativo",
        ylab = "Idade")
```




```r
boxplot(falso_positivo,
        main = "Boxplot das Idades - Falso Positivo",
        xlab = "Falso Positivo",
        ylab = "Idade")
```

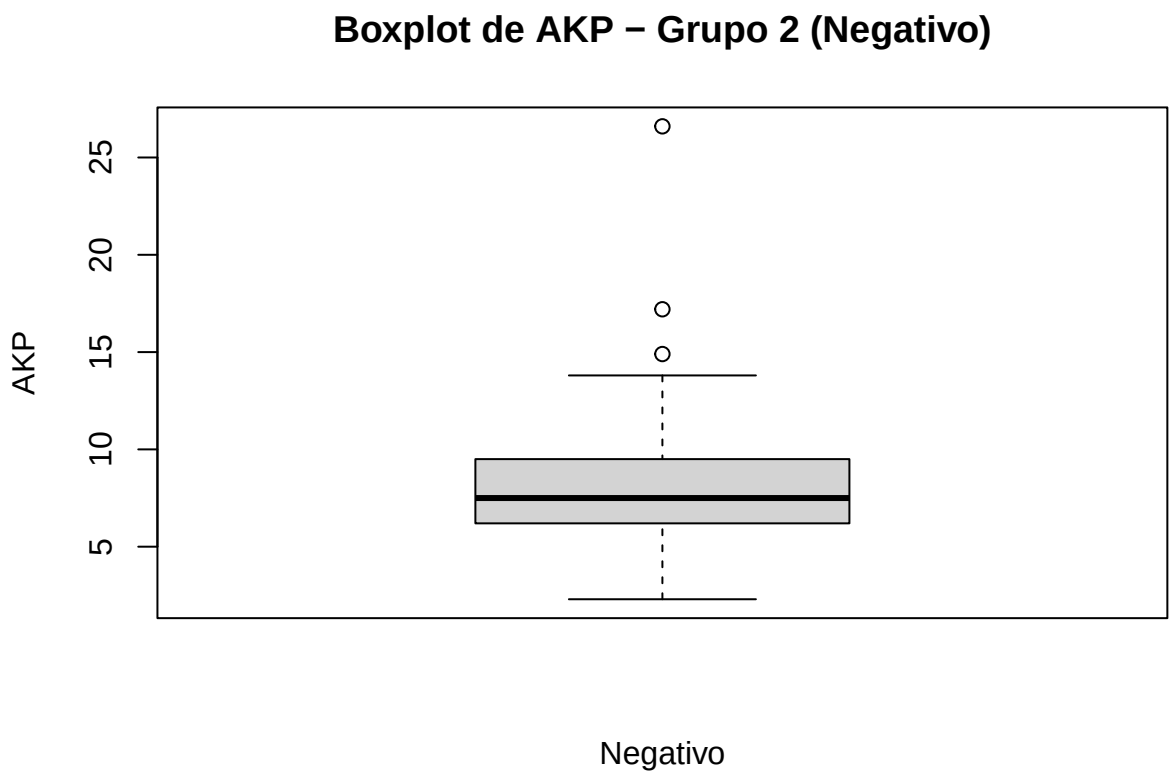


Resposta item (iii):

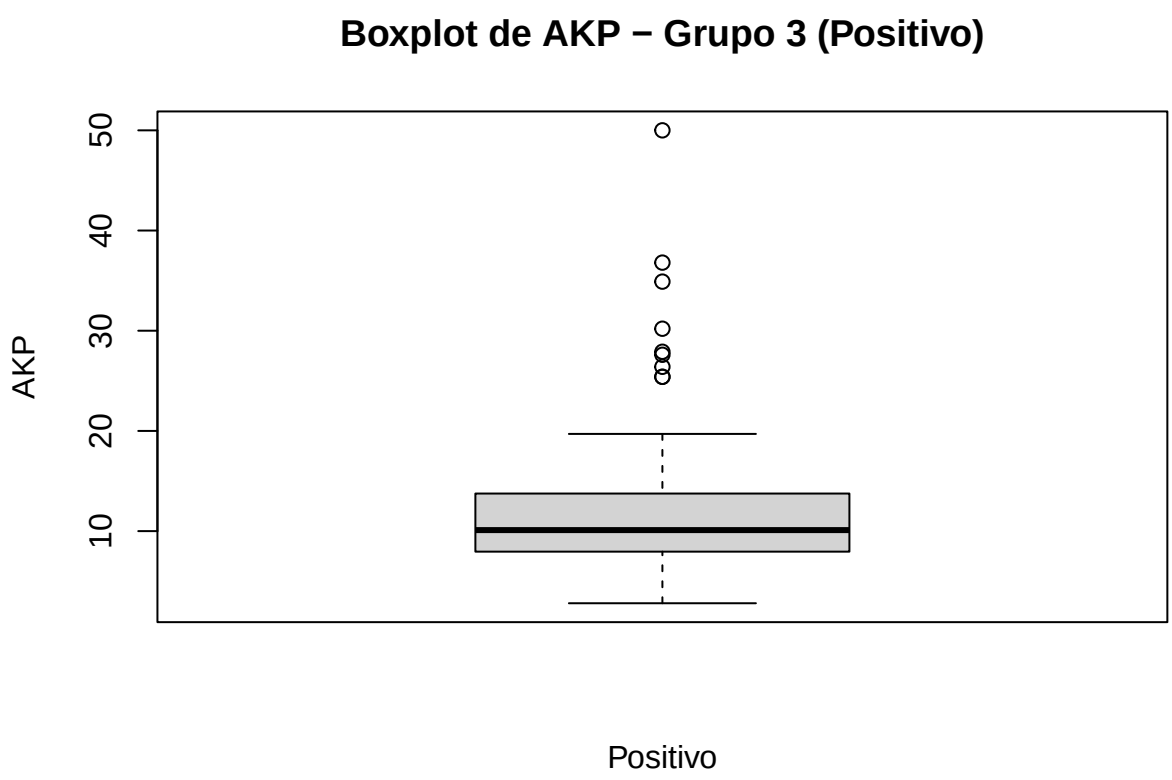
Podemos observar que a linha que representa a mediana das idades de ambos os *Boxplots* estão próximas, isto ocorre pois a mediana da idade do grupo 1 (Falso Negativo) é 55 e a do grupo 4 (Falso Positivo) é de 51. Considerando que o conjunto de idades dos grupos - diferença entre o limite inferior e superior - é relativamente grande, temos então medianas próximas. Além disso, não observamos nenhum *outlier* - valor descrepante - em nenhum dos gráficos.

```
akp_grupo2 <- dt[dt$Grupo == 2, "AKP"]
akp_grupo3 <- dt[dt$Grupo == 3, "AKP"]

boxplot(akp_grupo2,
        main = "Boxplot de AKP - Grupo 2 (Negativo)",
        xlab = "Negativo",
        ylab = "AKP")
```




```
boxplot(akp_grupo3,
        main = "Boxplot de AKP - Grupo 3 (Positivo)",
        xlab = "Positivo",
        ylab = "AKP")
```



Resposta item (iv):

Neste caso, vemos diferenças significativas entre os *boxplots* dos grupos 2 (Negativo) e 3 (Positivo) . Isso porque identifica-se *outliers* em ambos os gráficos, porém em menor quantidade nos dados do AKP do Grupo 2 e muitos nos dados do Grupo 3. Além disso, vemos os limites superiores e inferiores distintos, sendo o limite superior do Grupo 3 maior do que o do Grupo 2. Isso é esperado, pois o AKP é um indicador importante para identificar doenças, tanto quando está alto como quando está baixo. Espera-se, com base nesses dados, que paciêntes que estão com cancer tenham o AKP mais elevado. Por último, vale destacar a mediana - indicada pela linha horizontal contida no retângulo - que é maior no Grupo 3 do que no Grupo 2, provavelmente compreendido pela explicação feita anteriormente.