Amostragem (MATD44)

Prova - 01 (gabarito) - Questão 4

Raydonal Ospina (mailto:raydonal@castlab.org)

a)

```
dados <- read.table("~/Github/matd44/Scripts/dadosTabela.txt", quote="\"", comment.char="")

colnames(dados) <- c("ID", "Sexo", "Renda")

# Código para calcular a renda média e intervalo de confiança
media_renda <- mean(dados$Renda)
desvio_padrao <- sd(dados$Renda)
n <- nrow(dados)
erro_padrao <- desvio_padrao / sqrt(n)

# Intervalo de confiança de 95% para a média
intervalo_confianca_media <- qt(c(0.025, 0.975), df = n - 1) * erro_padrao + media_renda

cat("A renda média dos trabalhadores é:", round(media_renda, 2), "mil reais.\n")
```

```
## A renda média dos trabalhadores é: 1994.54 mil reais.
```

```
cat("Intervalo de confiança (95%) para a renda média:", round(intervalo_confianca_media, 2), "a", round
(intervalo_confianca_media[2], 2), "mil reais.\n")
```

```
## Intervalo de confiança (95%) para a renda média: 1845.68 2143.4 a 2143.4 mil reais.
```

b)

```
# Código para calcular a renda total e intervalo de confiança
renda_total <- sum(dados$Renda)

# Intervalo de confiança de 95% para a renda total
intervalo_confianca_renda_total <- c(renda_total - qt(0.975, df = n - 1) * desvio_padrao * sqrt(n),
                                     renda_total + qt(0.975, df = n - 1) * desvio_padrao * sqrt(n))

cat("A renda total dos trabalhadores é:", round(renda_total, 2), "mil reais.\n")
```

```
## A renda total dos trabalhadores é: 111694.1 mil reais.
```

```
cat("Intervalo de confiança (95%) para a renda total:", round(intervalo_confianca_renda_total[1], 2), "
a", round(intervalo_confianca_renda_total[2], 2), "mil reais.\n")
```

```
## Intervalo de confiança (95%) para a renda total: 103358.1 a 120030.2 mil reais.
```

c)

```
# Código para calcular a proporção e número total de mulheres
proporcao_mulheres <- sum(dados$Sexo == "Fem") / n
numero_total_mulheres <- round(proporcao_mulheres * 1000)

# Intervalo de confiança de 95% para a proporção de mulheres
erro_padrao_proporcao <- sqrt(proporcao_mulheres * (1 - proporcao_mulheres) / n)
intervalo_confianca_proporcao <- prop.test(sum(dados$Sexo == "Fem"), n)$conf.int

# Intervalo de confiança de 95% para o número total de mulheres
intervalo_confianca_numero_mulheres <- round(intervalo_confianca_proporcao * 1000)

cat("A proporção de mulheres na empresa é:", round(proporcao_mulheres, 2), ".\n")
```

```
## A proporção de mulheres na empresa é: 0.12 .
```

```
cat("Intervalo de confiança (95%) para a proporção de mulheres:", round(intervalo_confianca_proporcao
[1], 2), "a", round(intervalo_confianca_proporcao[2], 2), ".\n")
```

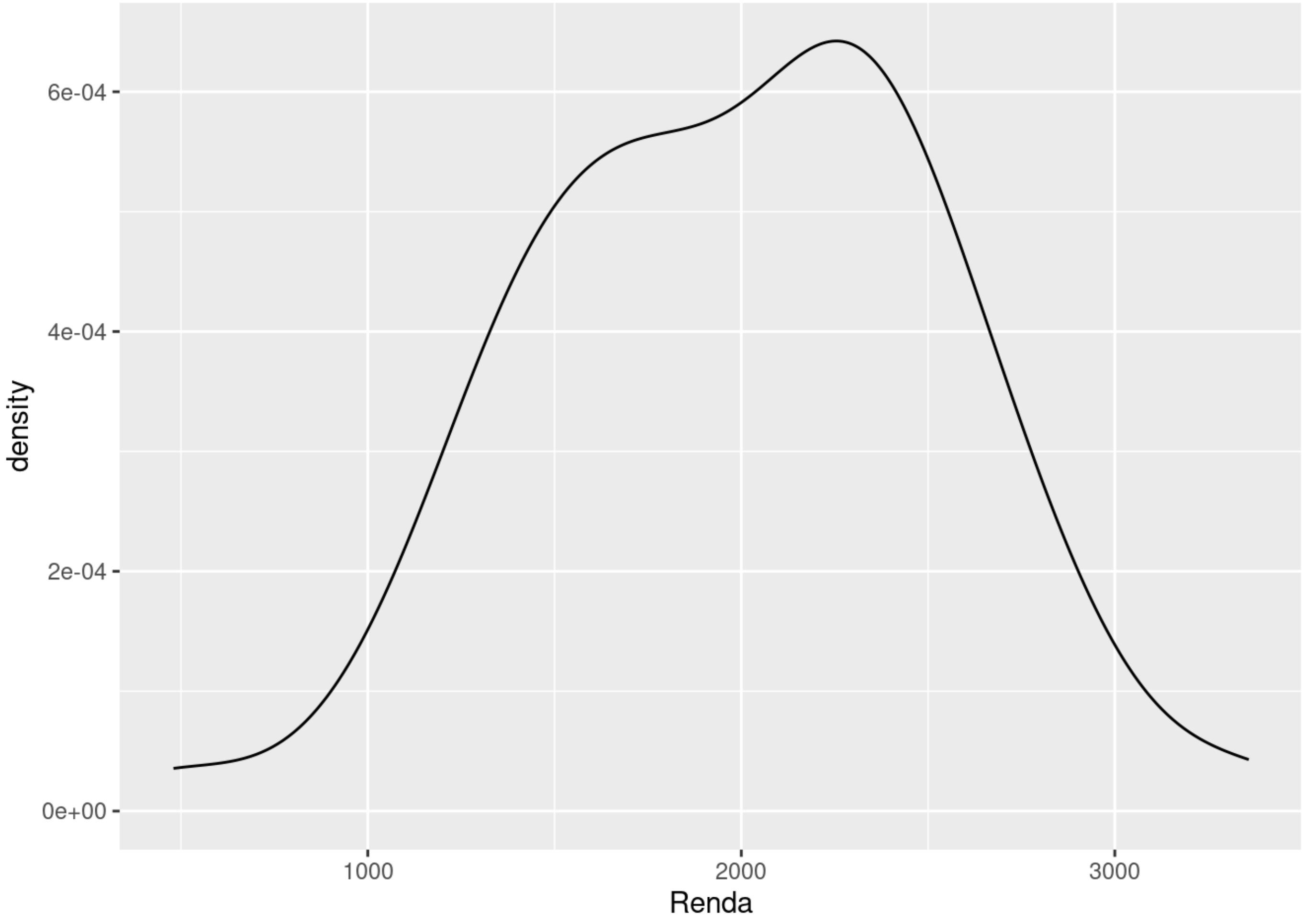
```
## Intervalo de confiança (95%) para a proporção de mulheres: 0.06 a 0.25 .
```

```
cat("O número total estimado de mulheres na empresa é:", round(numero_total_mulheres), "com intervalo de
confiança (95%):", round(intervalo_confianca_numero_mulheres[1]), "a", round(intervalo_confianca_numero_
mulheres[2]), ".\n")
```

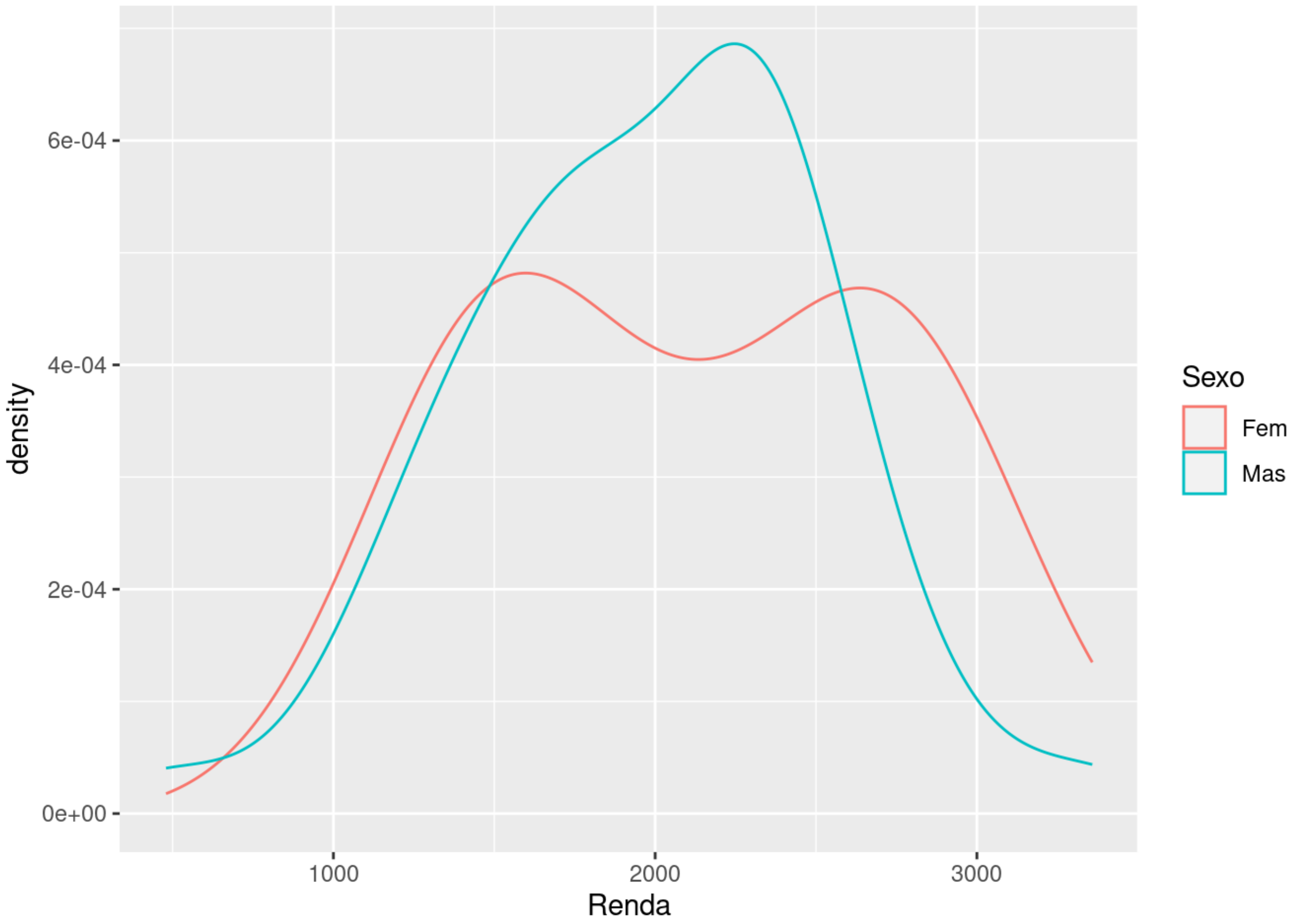
```
## O número total estimado de mulheres na empresa é: 125 com intervalo de confiança (95%): 56 a 247 .
```

d)

```
# Toda a a mostra independente do Sexo
library(ggplot2)
ggplot(dados, aes(x=Renda)) +
  geom_density()
```




```
# Segmentado por subpopulação
ggplot(dados, aes(x=Renda, color=Sexo)) +
  geom_density()
```




```r
# teste de normalidade não paramétrico de Shapiro-Wilk
# Global
shapiro.test(dados$Renda)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$Renda
## W = 0.99134, p-value = 0.9587
```

```r
# teste de normalidade não paramétrico de Shapiro-Wilk
# Subpopulação de mulheres
shapiro.test(dados$Renda[dados$Sexo=="Fem"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$Renda[dados$Sexo == "Fem"]
## W = 0.88161, p-value = 0.2337
```

```r
# teste de normalidade não paramétrico de Shapiro-Wilk
# Subpopulação de homens
shapiro.test(dados$Renda[dados$Sexo=="Mas"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$Renda[dados$Sexo == "Mas"]
## W = 0.98807, p-value = 0.8971
```

```r
# Como n é grande (n = 56), podemos considerar a aproximação pela distribuição normal.
cat("Sim, podemos considerar aproximações pela distribuição normal, pois a amostra é grande (n = 56).\n
e os testes de Shapiro não rejeitam a hipótese nula ao níveis usuais de significância estatística")
```

```
## Sim, podemos considerar aproximações pela distribuição normal, pois a amostra é grande (n = 56).
##  e os testes de Shapiro não rejeitam a hipótese nula ao níveis usuais de significância estatística
```

e)

```r
cat("Sim, as amostras podem ser consideradas como amostras aleatórias simples, pois foram selecionadas n
ão há argumentos para se pensar que foram selecionadas por uma mecanismo mais sofisticado, adicionalment
e pelos gráficos de densidade as distribuiç oes apresentam caudas semelhantes e simetria próxima o que é
um bom indicativo de que não houve mecanismo que favoreça mais um grupo ou outro.\n")
```

```
## Sim, as amostras podem ser consideradas como amostras aleatórias simples, pois foram selecionadas não há argumentos para se pensar que foram selecionadas por uma mecanismo mais sofisticado, adicionalmente pelos gráficos de densidade as distribuiç oes apresentam caudas semelhantes e simetria próxima o que é um bom indicativo de que não houve mecanismo que favoreça mais um grupo ou outro.
```

f)

```
# Código para calcular a renda média e o total das mulheres
media_renda_mulheres <- mean(dados$Renda[dados$Sexo == "Fem"])
total_renda_mulheres <- sum(dados$Renda[dados$Sexo == "Fem"])

cat("A renda média das mulheres na empresa é:", round(media_renda_mulheres, 2), "mil reais.\n")
```

```
## A renda média das mulheres na empresa é: 2113.95 mil reais.
```

```
cat("O total estimado da renda das mulheres na empresa é:", round(total_renda_mulheres, 2), "mil reais.\n")
```

```
## O total estimado da renda das mulheres na empresa é: 14797.64 mil reais.
```

A questão aqui não tem problemas em termos do estimador pontual. Contudo o verdadeiro problema está na variância do estimador.

Neste sentido pode se pensar em estimadores (condicionais), i.e

$$\mathrm{Var}(\bar{y}_k) = \frac{N_k - n_k}{N_k n_k} s_k^2,$$

em que N_k (Número total de elementos na subpopulação é conhecido), com n_k o número de elementos na amostra pertencendo a subpopulação k e s_k^2 a variância amostral.

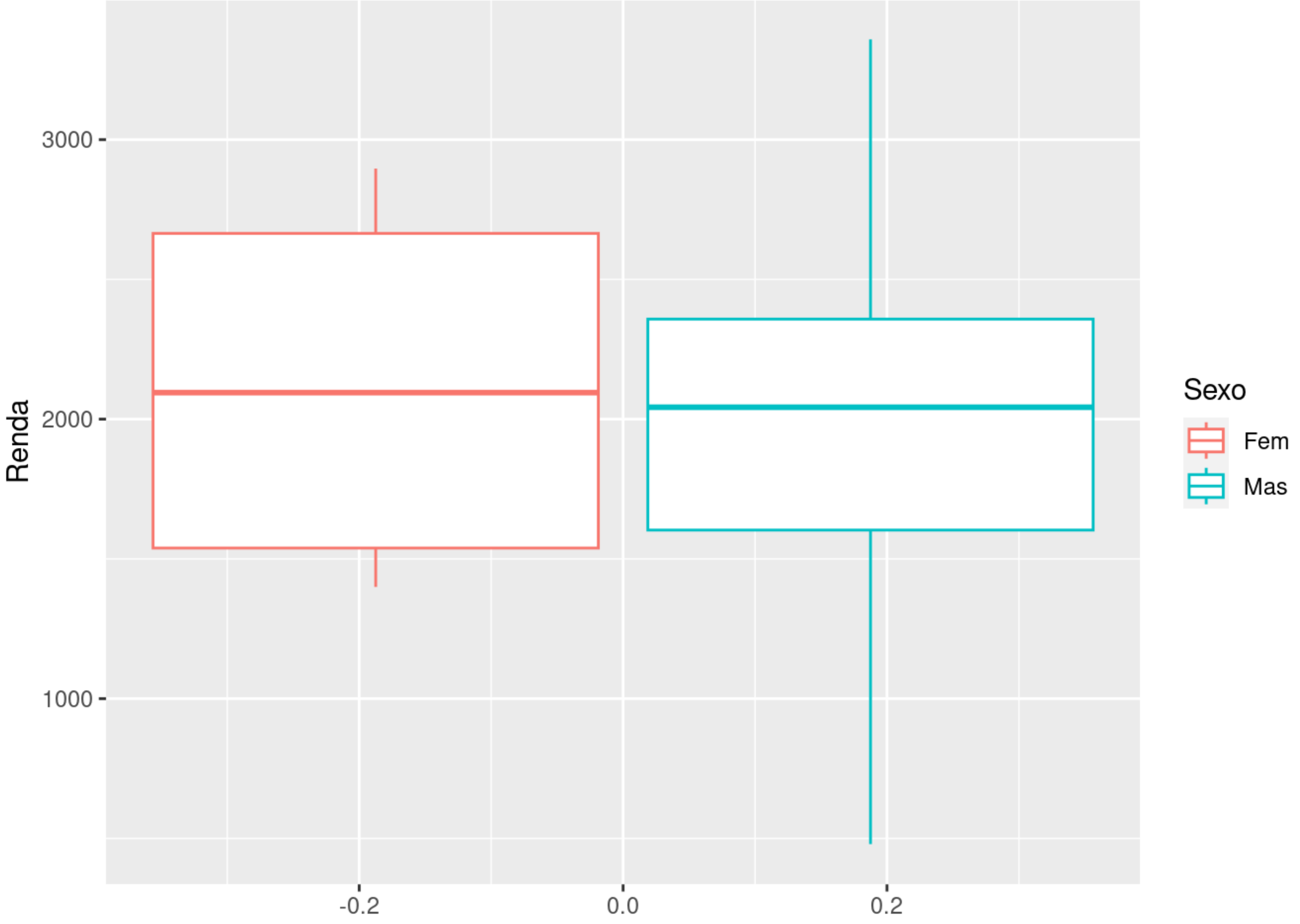
Por outro lado,

$$\mathrm{Var}(\bar{y}_k) = \frac{N - n}{N n_k} s_k^2,$$

se N_k é desconhecido, sendo n o tamanho total da amostra

g)

```
##
ggplot(dados, aes(y=Renda, color=Sexo)) +
  geom_boxplot()
```



```r
# Código para calcular os coeficientes de variação
cv_homens <- sd(dados$Renda[dados$Sexo == "Mas"]) / mean(dados$Renda[dados$Sexo == "Mas"])
cv_mulheres <- sd(dados$Renda[dados$Sexo == "Fem"]) / mean(dados$Renda[dados$Sexo == "Fem"])

# Verificar qual subpopulação tem o menor coeficiente de variação
subpopulacao_mais_homogenea <- ifelse(cv_homens < cv_mulheres, "Homens", "Mulheres")

cat("O coeficiente de variação para homens é:", round(cv_homens, 4), "\n")
```

```
## O coeficiente de variação para homens é: 0.2775
```

```r
cat("O coeficiente de variação para mulheres é:", round(cv_mulheres, 4), "\n")
```

```
## O coeficiente de variação para mulheres é: 0.3007
```

```r
cat("Portanto, a subpopulação mais homogênea em relação à renda é:", subpopulacao_mais_homogenea, "\n")
```

```
## Portanto, a subpopulação mais homogênea em relação à renda é: Homens
```