

Gabarito Prova I

Amostragem

1. lembramos da definição de probabilidade condicional, i.e. $P(A|B) = P(A \cap B) / P(B)$.

Também lembramos que um planejamento amostral é aleatório simples sem substituição se todas as possíveis amostras de tamanho n tem a mesma probabilidade de serem selecionadas, i.e.,

$$P(S=s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{se } \#S = n \\ 0 & \text{caso contrário} \end{cases} \quad (*)$$

em que $\pi_k = \frac{n}{N}$ para todo $k=1, \dots, N$

Agora, numa amostragem Bernoulli com $\pi_k = \pi \quad \forall k \in U$ temos que

$$P(S=s) = \pi^{n_s} (1-\pi)^{N-n_s} \quad (**)$$

Se consideramos que o tamanho $n_s = n$ (fixado) então

$$P(S=s | n_s = n) = \frac{P(S=s \text{ e } n_s = n)}{P(n_s = n)}$$

Note também que para o plano amostral Bernoulli,

$$n_s = \sum_{k \in U} I_k(s) \quad \text{em que}$$

$$I_k(s) = \begin{cases} 1 & \text{se o elemento } k \in s \\ 0 & \text{se o elemento } k \notin s \end{cases}$$

$$\text{logo } I_k(s) \sim \text{Bernoulli}(\pi_k) = \text{Bernoulli}(\pi)$$

$$\sum_{k \in U} I_k(s) \sim \text{Binomial}(N, \pi_k),$$

" Binomial (N, π) } hipótese

$$\text{logo, } P(n_s = n) = \binom{N}{n} \pi^n (1-\pi)^{N-n}$$

Agora. Note que

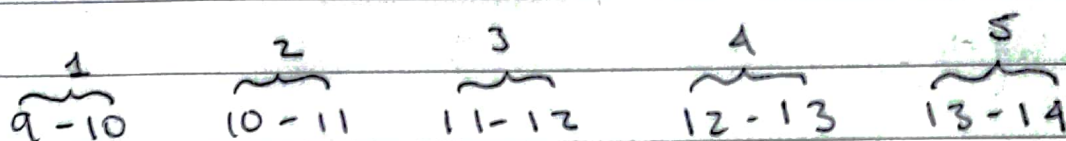
$$P(S=s \text{ e } n_s=n) = P((S=s) \cap (n_s=n))$$

$$= \pi^n (1-\pi)^{N-n}$$

Daí,

$$P(S=s | n_s=n) = \frac{\pi^n (1-\pi)^{N-n}}{\binom{N}{n} \pi^n (1-\pi)^{N-n}} = \frac{1}{\binom{N}{n}}$$

2.



A primeira hora é selecionada com probab. de $\frac{1}{5}$ mas a segunda hora é extraída de forma condicional.

Se a primeira hora foi 1 ou 5 a probabilidade é igual a $\frac{1}{3}$ pela hipótese, e probab. de $\frac{1}{2}$ no resto dos casos.

1^{ha} selecionada

	1	2	3	4	5	
1	0	0	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	} $p(s)$
2	0	0	0	$\frac{1}{10}$	$\frac{1}{10}$	
3	$\frac{1}{10}$	0	0	0	$\frac{1}{10}$	
4	$\frac{1}{10}$	$\frac{1}{10}$	0	0	0	
5	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	0	0	

As amostras são de tamanho 2

Note que $\sum_s p(s) = 1$

$$\text{De fato, } \frac{6}{15} + \frac{6}{10} = \frac{150}{150} = 1$$

k = primeira hora, l = segunda hora.

A amostra $s = (1, 2) \Rightarrow P(k=1, l=2) = 0$
horas contíguas.

$$s = (3, 1) \Rightarrow P(k=3, l=1) = \frac{1}{5} \cdot \frac{1}{3} = \frac{1}{15}$$

$$s = (3, 2) \Rightarrow P(k=4, l=2) = \frac{1}{5} \cdot \frac{1}{2} = \frac{1}{10}$$

Agora lembremos que

$\pi_k = \sum_{s \in S} P(s)$, em que S é o conjunto de
todas as amostras, i.e.,
 $\{1, 2, 3, 4, 5\} \times \{1, 2, 3, 4, 5\}$.

a)

$$\pi_1 = \sum_{\substack{s \in S \\ 1 \in s}} P(s) = \frac{4}{15} + \frac{2}{10} = \frac{8+6}{30} = \frac{12}{30}$$

$$\pi_2 = \sum_{\substack{s \in S \\ 2 \in s}} P(s) = \frac{3}{10} + \frac{1}{15} = \frac{11}{30}$$

$$\pi_3 = \sum_{\substack{s \in S \\ 3 \in s}} P(s) = \frac{2}{15} + \frac{2}{10} = \frac{10}{30}$$

$$\pi_4 = \sum_{\substack{s \in S \\ 4 \in s}} P(s) = \frac{3}{10} + \frac{1}{15} = \frac{11}{30}$$

$$\pi_5 = \sum_{\substack{s \in S \\ 5 \in s}} P(s) = \frac{4}{15} + \frac{2}{10} = \frac{12}{30}$$

Note que

$$E(n_s) = \sum_k \pi_k = \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5$$

$$= \frac{12}{30} + \frac{11}{30} + \frac{10}{30} + \frac{11}{30} + \frac{12}{30} = \frac{56}{30}$$

$$= 1,8667 \approx 2.$$

Agora

$$\pi_{k,l} = \sum_{\substack{S \in \mathcal{S} \\ k \in S, l \in S}} P(S)$$

b)

$\pi_{11} = 0$	$= 0$	$\pi_{23} = 0$	$\pi_{34} = 0$
$\pi_{12} = 0$	$= 0 = \pi_{21}$	$\pi_{24} = \frac{1}{10} + \frac{1}{10} = \frac{2}{10}$	$\pi_{35} = \frac{1}{10} + \frac{1}{15} = \frac{5}{30}$
$\pi_{13} = \frac{1}{15} + \frac{1}{10} = \frac{5}{30} = \pi_{31}$		$\pi_{25} = \frac{1}{10} + \frac{1}{15} = \frac{5}{30}$	$=$
$\pi_{14} = \frac{1}{15} + \frac{1}{10} = \frac{5}{30} = \pi_{41}$			$\pi_{45} = 0$
$\pi_{15} = \frac{1}{15} + \frac{1}{15} = \frac{2}{30} = \pi_{51}$			$\pi_{55} = 0$

$$(\pi)_{k,l} = \begin{bmatrix} 0 & 0 & 5/30 & 5/30 & 2/30 \\ 0 & 0 & 0 & 6/30 & 5/30 \\ 5/30 & 0 & 0 & 0 & 5/30 \\ 5/30 & 6/30 & 0 & 0 & 0 \\ 2/30 & 5/30 & 5/30 & 0 & 0 \end{bmatrix} = \pi$$

Logo o plano amostral é não mensurável pois tem $\pi_{11} = 0$.

c)

$$\underline{\Delta} = C(I_k, I_l) = \begin{cases} \pi_{kl} - \pi_k \pi_l & , k \neq l \\ \pi_k (1 - \pi_k) & , k = l \end{cases}$$

$$k, l \in \{1, 2, 3, 4, 5\}$$

$$\Delta_{11} = \pi_1(1-\pi_1) = \frac{12}{30} \left(1 - \frac{12}{30}\right) = \frac{12}{30} \times \frac{18}{30} = \frac{216}{900}$$

$$\Delta_{12} = \pi_{12} - \pi_1\pi_2 = 0 - \frac{12}{30} \times \frac{11}{30} = -\frac{12}{30} \times \frac{11}{30} = -\frac{132}{900}$$

$$\Delta_{13} = \pi_{13} - \pi_1\pi_3 = \frac{5}{30} - \frac{12}{30} \times \frac{10}{30} = \frac{30}{900}$$

$$\Delta_{14} = \pi_{14} - \pi_1\pi_4 = \frac{5}{30} - \frac{12}{30} \times \frac{11}{30} = \frac{18}{900}$$

$$\Delta_{15} = \pi_{15} - \pi_1\pi_5 = \frac{2}{30} - \frac{12}{30} \times \frac{12}{30} = -\frac{84}{900}$$

$$\Delta_{22} = \pi_2(1-\pi_2) = \frac{11}{30} \left(1 - \frac{11}{30}\right) = \frac{11}{30} \times \frac{19}{30} = \frac{209}{900}$$

$$\Delta_{23} = \pi_{23} - \pi_2\pi_3 = 0 - \frac{11}{30} \left(\frac{10}{30}\right) = -\frac{110}{900}$$

$$\Delta_{24} = \pi_{24} - \pi_2\pi_4 = \frac{2}{30} - \left(\frac{11}{30} \times \frac{11}{30}\right) = \frac{180}{900} - \frac{121}{900} = \frac{59}{900}$$

$$\Delta_{25} = \pi_{25} - \pi_2\pi_5 = \frac{5}{30} - \left(\frac{11}{30} \times \frac{12}{30}\right) = \frac{150}{900} - \frac{132}{900} = \frac{18}{900}$$

$$\Delta_{33} = \pi_3(1-\pi_3) = \frac{10}{30} \times \left(1 - \frac{10}{30}\right) = \frac{10}{30} \times \left(\frac{20}{30}\right) = \frac{200}{900}$$

$$\Delta_{34} = \pi_{34} - \pi_3\pi_4 = 0 - \left(\frac{10}{30} \times \frac{11}{30}\right) = -\frac{110}{900}$$

$$\Delta_{35} = \pi_{35} - \pi_3\pi_5 = \frac{9}{30} - \left(\frac{10}{30} \times \frac{12}{30}\right) = \frac{150}{900} - \frac{120}{900} = \frac{30}{900}$$

$$\Delta_{44} = \pi_4(1-\pi_4) = \frac{11}{30} \left(1 - \frac{11}{30}\right) = \frac{11}{30} \times \left(\frac{19}{30}\right) = \frac{209}{900}$$

$$\Delta_{45} = \pi_{45} - \pi_4\pi_5 = 0 - \frac{11}{30} \times \frac{12}{30} = -\frac{132}{900}$$

$$\Delta_{55} = \pi_5(1-\pi_5) = \frac{12}{30} \left(1 - \frac{12}{30}\right) = \frac{12}{30} \times \frac{18}{30} = \frac{216}{900}$$

$$\underline{\underline{\Delta}} = \frac{1}{900}$$

$$\begin{bmatrix} 216 & -132 & 30 & 18 & -84 \\ -132 & 209 & -110 & 59 & 18 \\ 30 & -110 & 200 & -110 & 30 \\ 18 & 59 & -110 & 209 & -132 \\ -84 & 18 & 30 & -132 & 216 \end{bmatrix}$$

3)

a) Com base nas informações da tabela pode-se obter estimações para o gasto médio da população e I.C.

Seja $N = 20000$ então o gasto médio estimado é

$$\hat{\bar{t}}_{\pi} = \bar{y}_{est} = 0,3(1,2) + 0,2(2,4) + 0,5(0,6)$$

$$= 1,14 \text{ 'salários mínimos'} \quad (\text{média populacional})$$

o gasto total

$$\hat{\hat{t}}_{\pi} = \hat{y}_{est} = N(\bar{y}_{est}) = 20.000(1,14) = 22800$$

salários mínimos

$$\text{Var}(\bar{y}_{est}) = \frac{(0,3)^2(0,36)}{40} + \frac{(0,2)^2(1,21)}{36} + \frac{(0,5)^2(0,04)}{44}$$

$$- \frac{(0,3)(0,36) + (0,2)(1,21) + (0,5)(0,04)}{20000}$$

$$\text{Var}(\bar{y}_{est}) = \text{Var}(\hat{\bar{t}}_{\pi}) = \sum_{h=1}^3 \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^3 \frac{W_h S_h^2}{N}, \text{ logo}$$

$$\text{Var}(\bar{y}_{est}) = 0,001262106$$



$$d.p(\bar{Y}_{est}) = \sqrt{\text{Var}(\bar{Y}_{est})} = \underline{0,03552 \text{ salários mínimos}}$$

logo um intervalo de confiança de 95% para \bar{Y} é

$$\bar{Y}_{est} \pm t_{n-1, \alpha/2} \times d.p(\bar{Y}_{est})$$

$$[1,14 \pm t_{119} \times 0,03552; 1,14 \pm t_{119} \times 0,03552]$$

O valor de t_{119} é 1,98 (que não é muito diferente de $Z_{1-\alpha/2}$ que seria 1,96),

$$1,14 \pm 1,98 \times 0,03552$$

$$(1,06967; 1,21033) \text{ em salários mínimos.}$$

Agora para o total temos que

$$\text{Var}(\hat{t}_n) = \text{Var}(N \bar{Y}_{est}) = N^2 \text{Var}(\bar{Y}_{est})$$

logo

$$d.p(\hat{t}_n) = N \sqrt{\text{Var}(\bar{Y}_{est})} = N \times d.p(\bar{Y}_{est})$$

Dar

$$\hat{t}_{\pi} \pm t_{n-1, 1-\alpha} d_p(\hat{t}_{\pi}) \text{ é um}$$

I.C. de $(1-\alpha)\%$ para o total.

Aqui $d_p(\hat{t}_{\pi}) = 710,5226$ logo o I.C.

$$22800 \pm 1,98 (710,5226) \text{ que é}$$

$$(21393,17 ; 24206,83) \text{ em salários mínimos}$$

b) Se o orçamento para a coleta da informação não pode ser superior a R\$ 2,5 milhões podemos pensar da seguinte forma:

O tamanho de amostra numa amostragem estratificada aleatória simples que minimiza a variância da média (ou total) dada um orçamento fixo pode ser visto como

$$n = \frac{(C - C_0) \sum_{h=1}^3 \left(\frac{W_h S_h}{\sqrt{C_h}} \right)}{\sum_{h=1}^3 (W_h S_h \sqrt{C_h})} \quad (2)$$



lembrando que a função custo pode ser linear, i.e.

$$C = C_0 + \sum_{h=1}^H C_h n_h$$

H = número total de estratos.

C = orçamento total para a coleta de informação
 C_0 = custo fixo que não depende do número de elementos a selecionar e
 C_h = custo de realizar a amostragem no estrato h .

Vimos que o tamanho de amostra ótimo por estrato é proporcional a n e é dado por

$$n_h = n \frac{P_h}{\sum_{h=1}^H P_h}, \text{ e (ii) que, em que}$$

$$P_h = \frac{W_h S_h}{\sqrt{C_h}}$$

P_h é uma variável de proporcionalidade

Utilizando (ii) temos

$$n = 2500000 \left(\frac{4.000 \sqrt{0,36}}{\sqrt{5000}} + \frac{6.000 \sqrt{1,21}}{\sqrt{3000}} + \frac{10.000 \sqrt{0,04}}{\sqrt{1000}} \right)$$

M

$$M = 4000 \sqrt{0,36} \sqrt{5000} + 6000 \sqrt{1,21} \sqrt{6000} + 10000 \sqrt{0,04} \sqrt{1000}$$

$$n = \frac{217,6856 \times 25000}{59448.1}$$

$$n = 915,4948 \quad (\text{tamanho global})$$

$$\underline{n \approx 916}$$

E os tamanhos por estratos devem seguir (:) , i.e

$$\underline{n_1} = 915,4948 \times \frac{\left(\frac{4.000 \sqrt{0,36}}{\sqrt{5000}} \right)}{217,6856} = 142,7422$$

$$\approx \underline{143}$$

$$\underline{n_2} = 915,4948 \times \frac{\left(\frac{6000 \sqrt{1,21}}{\sqrt{3000}} \right)}{217,6856} = 506,7682$$

$$\approx \underline{507}$$

$$\underline{n_3} = 915,4948 \times \frac{\left(\frac{10000 \sqrt{0,04}}{\sqrt{1000}} \right)}{217,6856} = 265,9843$$

$$\approx \underline{266}$$

4) Questão 4 desenvolvida em Th (arquivo adjunto)

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")

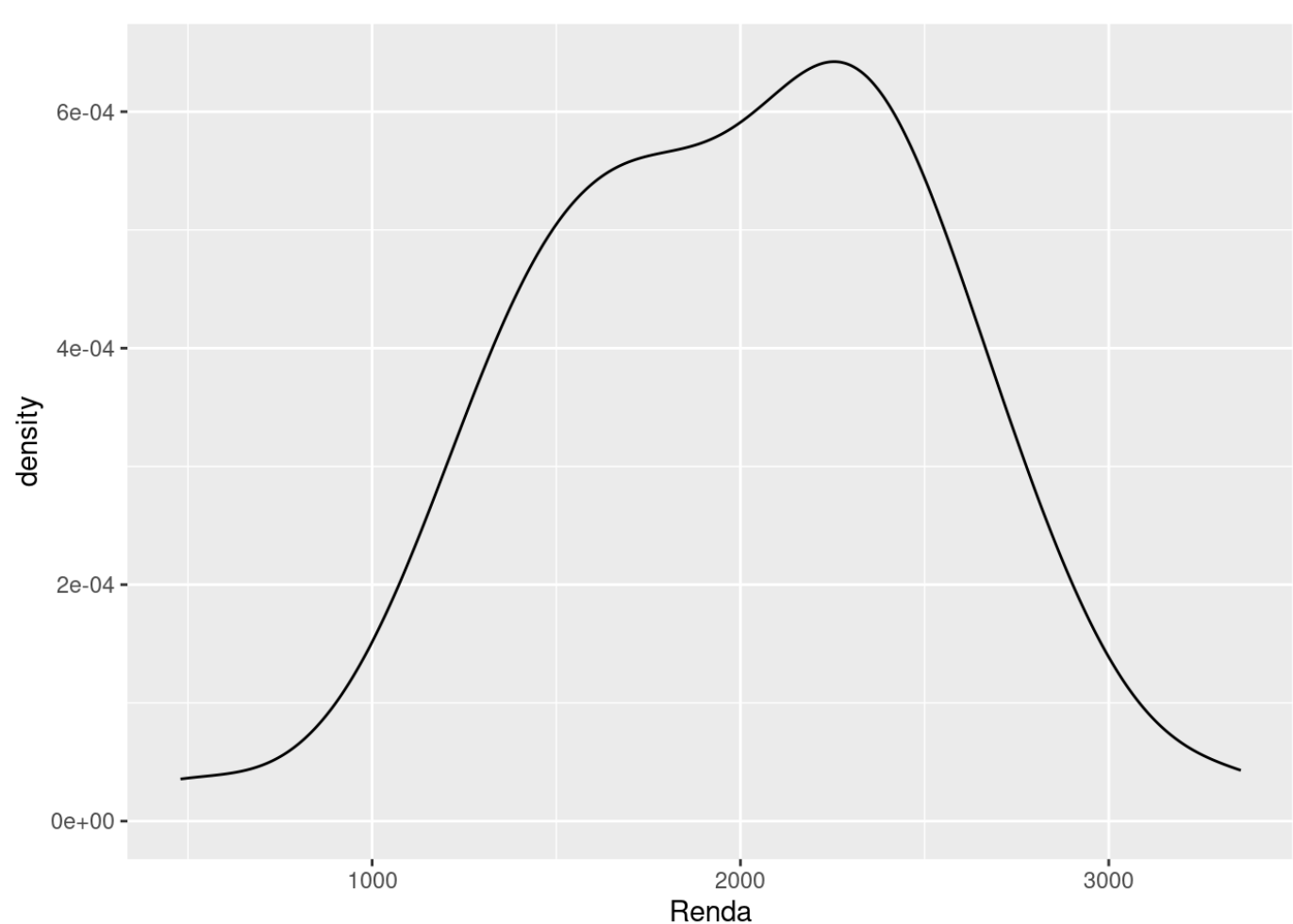
## Intervalo de confiança (95%) para a proporção de mulheres: 0.06 a 0.25 .

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 .
```

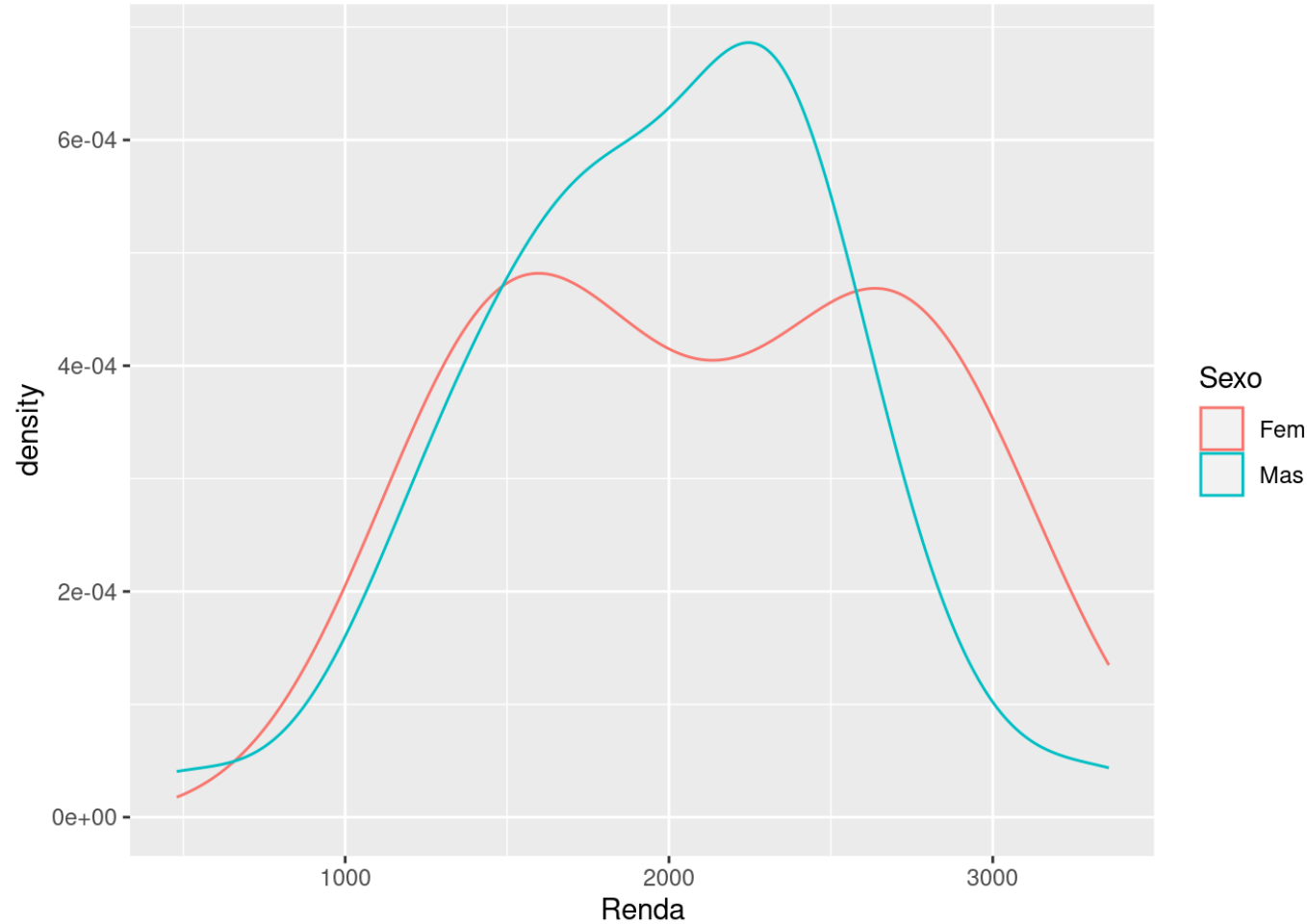
d)

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



The figure is a density plot of the variable 'Renda' (Income). The x-axis is labeled 'Renda' and has major tick marks at 1000, 2000, and 3000. The y-axis is labeled 'density' and has major tick marks at 0e+00, 2e-04, 4e-04, and 6e-04. The plot shows a single, smooth, unimodal curve that starts at a low density near Renda=500, rises to a peak of approximately 6.5e-04 at Renda ≈ 2200, and then falls back to a low density near Renda=3500. The background of the plot area is light gray with white grid lines.

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



```
# 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
```

```
# 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
```

```
# 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
```

```
# 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)

```
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 p  
elos 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

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

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.

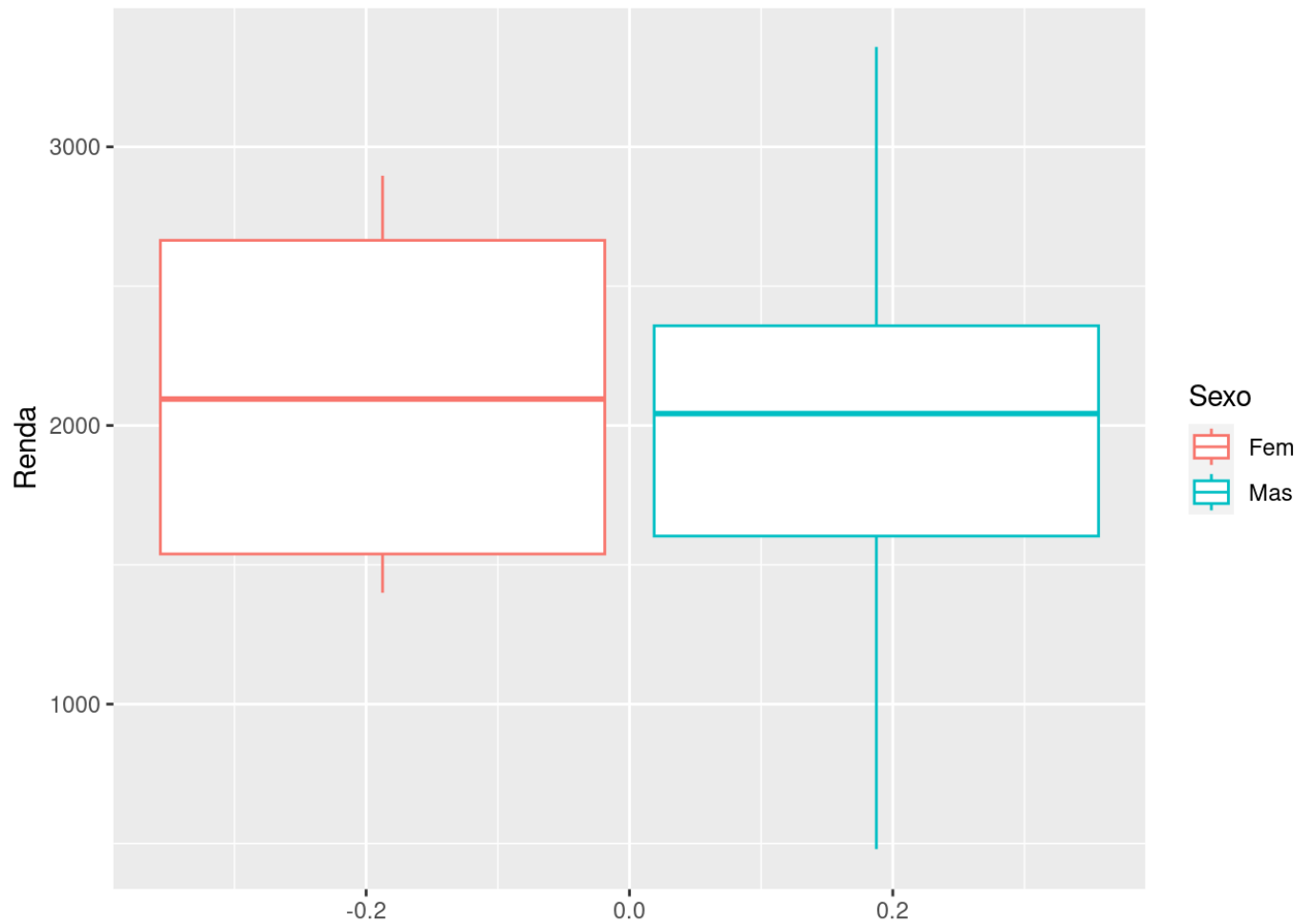
Por outro lado,

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

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

g)

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



```
# 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
```

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

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

```
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
```