UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA
PROVA II — MATD44 — GABARITO

Professor: Raydonal Ospina

E-mail: raydonal@castlab.org

Regras: Leia com atenção as perguntas. **Todas as questões devem ser detalhadas.** A prova deve ser claramente resolvida. Seja claro e organizado.

Atenção: Você deverá encaminhar a solução da sua prova digitalizada em formato **PDF** no e-mail acima no dia 20/08/2023 até as 15:00h (GMT-3 - Horário de Brasília)

I ► (**Amostragem estratificada**) A Tabela 1 contem as informações do gasto mensal em serviços públicos de uma amostra aleatória estratificada de 120 famílias na cidade de Salvador a qual foi geograficamente dividida em três estratos: Norte, Centro e Sul.

	Estratos		
Estatísticas	Norte (1)	Centro (2)	Sul (3)
N_h	4000	6000	10000
W_h	0,3	0,2	0,5
n_h	40	36	44
$\bar{y}_h$	1,2	2,4	0,6
$\bar{Y}_h$	9600	7200	6000
s_h^2	0,36	1,21	0,04
$\text{Var}(\bar{y}_h)$	0,000993	0,004404	0,000226

Tabela 1: Informações do gasto familiar mensal em serviços públicos (em salários mínimos) a partir de uma amostra aleatória simples estratificada na cidade de Salvador.

a) Estime o gasto médio e o gasto total de toda a população. Estabeleça um intervalo de confiança de 95% para essas medidas.

b) Suponhamos que os custos de coletar a informação por família para cada um dos estratos são: C_1 = R\$5.000, C_2 = R\$3.000 e C_3 = R\$1.000, respectivamente. Assuma que que existem informações previas em relação a variabilidade de cada estrato e que estas informações correspondem as estimativas s_h^2 da tabela 1. Se o orçamento da coleta de informação não pode ser superior a R\$2, 5 milhões qual deve ser o tamanho da amostra global e como ele deve ser dividido entre os diferentes estratos? Suponha que a função de custo é linear na forma

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

em que C é o orçamento total para a coleta de informação, C_0 é o custo fixo que não depende do número de unidades amostrais a serem selecionadas e C_h é o custo de amostrar uma unidade amostral no estrato h.

a) Res: O gasto médio da população (em salários mínimos), $\bar{y}$, é dado por:

$$\bar{y} = \sum_{h=1}^{H} W_h \bar{y}_h = 0,3 \times 1,2 + 0,2 \times 2,4 + 0,5 \times 0,6 = 1,14$$

O gasto total estimado da população (em salários mínimos), $\hat{Y}$, é:

$$\hat{Y} = N\bar{y} = (4000 + 6000 + 10000) \times 1,14 = 20.000 \times 1,14 = 22.800$$

O erro padrão da média é dado pela fórmula ajustada, incluindo o fator de correção finita:

$$\text{Var}(\bar{y}) = \sum_{h=1}^{H} W_h^2 \frac{s_h^2}{n_h} \left(\frac{N_h - n_h}{N_h} \right).$$

Substituindo os valores dos estratos:

$$\text{Var}(\bar{y}) = (0,3)^2 \times \frac{0,36}{40} \times \frac{4000-40}{4000} + (0,2)^2 \times \frac{1,21}{36} \times \frac{6000-36}{6000}$$

$$+(0,5)^2 \times \frac{0,04}{44} \times \frac{10000-44}{10000}.$$

$$\text{Var}(\bar{y}) \approx 0,002944.$$

Logo,

$$\text{EP}(\bar{y}) = \sqrt{0,002944} = 0,0543.$$

Para construir o intervalo de confiança usando a distribuição t-Student, devemos calcular os graus de liberdade aproximados. Como o plano é estratificado, podemos usar a fórmula aproximada para graus de liberdade em amostragem estratificada:

$$df \approx \frac{\left(\sum_{h=1}^{H} W_h^2 \frac{s_h^2}{n_h} \right)^2}{\sum_{h=1}^{H} \left(\frac{W_h^4 s_h^4}{n_h^2 (n_h - 1)} \right)}.$$

Calculando este valor para os dados fornecidos, obtemos:

$$df \approx 78.$$

O valor crítico da distribuição t para $df = 78$ e um nível de confiança de 95% é aproximadamente $t_{0,025} \approx 1,99$. O intervalo de confiança de 95% para o gasto médio da população é dado por:

$$\bar{y} \pm t_{0,025} \times \text{EP}(\bar{y}) = 1,14 \pm 1,99 \times 0,0543 = [1,032, 1,248] \text{ salários mínimos.}$$

Para o gasto total da população, o intervalo de confiança de 95% é:

$$\hat{Y} \pm t_{0,025} \times \text{EP}(\hat{Y}) = 22.800 \pm 1,99 \times (20.000 \times 0,0543) = [21.724, 23.876]$$

Se usamos uma aproximação apenas normal, supondo que o n é suficientemente grande, temos que o intervalo de confiança de 95% para o gasto médio é:

$$\bar{y} \pm z_{0,025} \times \text{EP}(\bar{y}) = 1,14 \pm 1,96 \times 0,0543 = [1,033, 1,264]$$

Para o gasto total da população, o intervalo de confiança de 95% é:

$$\hat{Y} \pm z_{0,025} \times \text{EP}(\hat{Y}) = 22.800 \pm 1,96 \times (20.000 \times 0,0543) = [20.671, 24.928]$$

b) Res: Orçamento total disponível:

$$C = 2.500.000$$

Os Custos de coleta por estrato:

$$C_1 = 5.000, \quad C_2 = 3.000, \quad C_3 = 1.000$$

Alocação ótima proporcional ao tamanho e à variância ajustada pelo custo, neste caso, o tamanho da amostra em cada estrato n_h é:

$$n_h = n \times \frac{W_h \sqrt{s_h^2 / C_h}}{\sum_{h=1}^{H} W_h \sqrt{s_h^2 / C_h}},$$

onde n é o tamanho total da amostra.

Para o tamanho total da amostra n, a A soma do número de unidades amostrais nos estratos deve satisfazer a restrição orçamentária:

$$\sum_{h=1}^{H} C_h n_h = 2.500.000.$$

Calculemos a fração $\frac{W_h \sqrt{s_h^2 / C_h}}{\sum_{h=1}^{H} W_h \sqrt{s_h^2 / C_h}}$ para cada estrato e depois encontramos os valores de n e dos n_h, i.e., $W_1 = 0,3$, $s_1^2 = 0,36$, $C_1 = 5000$ $W_2 = 0,2$, $s_2^2 = 1,21$, $C_2 = 3000$ $W_3 = 0,5$, $s_3^2 = 0,04$, $C_3 = 1000$ Para cada estrato, calcule $W_h \sqrt{s_h^2 / C_h}$, logo:

$$W_1 \sqrt{\frac{s_1^2}{C_1}} = 0,3 \times \sqrt{\frac{0,36}{5000}} \approx 0,00268$$

$$W_2 \sqrt{\frac{s_2^2}{C_2}} = 0,2 \times \sqrt{\frac{1,21}{3000}} \approx 0,00401$$

$$W_3 \sqrt{\frac{s_3^2}{C_3}} = 0,5 \times \sqrt{\frac{0,04}{1000}} \approx 0,00316$$

Somando todos os valores:

$$\sum_{h=1}^{H} W_h \sqrt{\frac{s_h^2}{C_h}} \approx 0,00268 + 0,00401 + 0,00316 = 0,00985.$$

O tamanho total da amostra n é dado por:

$$n = \frac{C}{\sum_{h=1}^{H} C_h \frac{W_h \sqrt{s_h^2 / C_h}}{\sum_{h=1}^{H} W_h \sqrt{s_h^2 / C_h}}}.$$

Substituindo os valores, calculamos n e depois os n_h.

```
# Código em Python

# Valores fornecidos
C_total = 2500000  # Orçamento total
C_h = [5000, 3000, 1000]  # Custos por estrato
W_h = [0.3, 0.2, 0.5]  # Pesos dos estratos
s2_h = [0.36, 1.21, 0.04]  # Variâncias por estrato

# Calcular W_h * sqrt(s2_h / C_h) para cada estrato
W_h_sqrt_sh_C_h = [W_h[i] * (s2_h[i] / C_h[i])**0.5
for i in range(len(C_h))]

# Somatório de W_h * sqrt(s2_h / C_h)
sum_W_h_sqrt_sh_C_h = sum(W_h_sqrt_sh_C_h)

# Tamanho total da amostra n
n_total = C_total / sum([C_h[i] * (W_h_sqrt_sh_C_h[i] /
sum_W_h_sqrt_sh_C_h) for i in range(len(C_h))])

# Calcular n_h para cada estrato
n_h = [n_total * (W_h_sqrt_sh_C_h[i] /
sum_W_h_sqrt_sh_C_h) for i in range(len(C_h))]

n_total, n_h
```

O tamanho total da amostra necessário é de aproximadamente $n = 870$ unidades. A amostra deve ser distribuída entre os estratos da seguinte forma:

Estrato Norte: $n_1 \approx 228$ unidades.

Estrato Centro: $n_2 \approx 359$ unidades.

Estrato Sul: $n_3 \approx 283$ unidades.

Esses valores garantem que o orçamento total de R$ 2,5 milhões seja respeitado, considerando os custos de coleta e a variabilidade entre os estratos.

2 ▶ **(Estimação em subpopulações finitas (domínios))** Uma amostra aleatória simples e sem substituição de 56 pessoas foi selecionada de uma população de 1000 trabalhadores da empresa LINCATECH. Foram coletadas informações sobre a renda mensal em miles de reais (Renda) e o sexo (Sexo) do trabalhador. Com as informações da Tabela 2 estime:

a) Estime a proporção e o número total de mulheres na empresa. Estabeleça intervalos de confiança de 95% para a proporção e total de mulheres na empresa.

b) Podem ser consideradas válidas aproximações pela distribuição normal no item anterior? Explique.

c) Considera que as amostras, tanto de homens como de mulheres poderiam ser assumidas como amostras aleatórias simples e sem substituição das respectivas subpopulações de homens e mulheres da empresa? Explique.

d) Como poderia ser estimada a renda média e o total das mulheres para toda a empresa se não se conhecesse o número total delas?

ID	Sexo	Renda	ID	Sexo	Renda
1	Fem	2094,90	15	Mas	2939,21
2	Mas	2386,14	16	Mas	1722,62
3	Mas	1562,82	17	Fem	2739,79
4	Mas	1781,41	18	Mas	1821,61
5	Mas	1603,14	19	Mas	1742,40
6	Mas	479,73	20	Mas	1845,22
7	Mas	2196,85	21	Mas	1916,60
8	Mas	2365,21	22	Mas	1329,28
9	Mas	2016,32	23	Mas	2143,04
10	Mas	1322,23	24	Mas	2618,97
11	Fem	2589,08	25	Fem	1399,85
12	Fem	2896,28	26	Mas	1610,41
13	Mas	1370,55	27	Mas	2300,84
14	Mas	975,94	28	Mas	1192,90
29	Mas	2715,62	43	Mas	1802,16
30	Mas	2042,58	44	Mas	2444,68
31	Mas	2235,73	45	Mas	2644,75
32	Mas	2223,33	46	Fem	1431,53
33	Mas	2618,16	47	Mas	1094,02
34	Mas	2206,57	48	Mas	1548,96
35	Mas	2432,08	49	Mas	2410,51
36	Mas	1340,94	50	Mas	2286,37
37	Mas	2321,37	51	Mas	1589,07
38	Mas	1922,91	52	Fem	1646,21
39	Mas	2520,83	53	Mas	3358,63
40	Mas	2063,78	54	Mas	1369,42
41	Mas	2335,66	55	Mas	2047,63
42	Mas	2357,94	56	Mas	1719,34

Tabela 2: Tabela de Informações dos empregados na amostra

e) Qual das duas subpopulações (homens, mulheres) é mais homogênea em relação a renda?

a) Res: Para estimar a proporção de mulheres na empresa, usamos a fórmula da proporção amostral:

$$\hat{p} = \frac{n_{\text{mulheres}}}{n}$$

em que n_{mulheres} é o número de mulheres na amostra e n é o tamanho da amostra. Contando o número de mulheres na tabela, temos 8 mulheres em uma amostra de 56 pessoas, então:

$$\hat{p} = \frac{8}{56} = 0,1429$$

Agora, para estimar o número total de mulheres na empresa

$$\hat{N}_{\text{mulheres}} = \hat{p} \times N$$

em que N é o tamanho total da população, que é 1000. Portanto, a estimativa para o número total de mulheres é:

$$\hat{N}_{\text{mulheres}} = 0,1429 \times 1000 = 142,9$$

O erro padrão da proporção é

$$\text{EP}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \times \frac{N-n}{N-1}}$$

Logo

$$\text{EP}(\hat{p}) = \sqrt{\frac{0,1429 \times (1-0,1429)}{56} \times \frac{1000-56}{1000-1}}$$
$$= \sqrt{0,1224 \times 0,945} = 0,0461$$

O intervalo de confiança de 95% para a proporção de mulheres é dado por:

$$\hat{p} \pm z_{0,025} \times \text{EP}(\hat{p}) = 0,1429 \pm 1,96 \times 0,0461 = [0,0525, 0,2333]$$

O intervalo de confiança para o número total de mulheres é

$$\hat{N}_{\text{mulheres}} \pm z_{0,025} \times N \ \text{EP}(\hat{p}) = 142,9 \pm 1,96 \times 1000 \times 0,0461$$
$$= [50,9, 234,9] \approx [51, 235]$$

b) Res: A aproximação pela distribuição normal pode ser considerada válida se o número de sucessos e fracassos na amostra for suficientemente grande, ou seja, se $n\hat{p} \geq 5$ e $n(1-\hat{p}) \geq 5$. Note que,

$$n\hat{p} = 56 \times 0,1429 = 8 \quad \text{e} \quad n(1-\hat{p}) = 56 \times 0,8571 = 48$$

Logo, ambas as quantidades são maiores que 5, então a aproximação pela distribuição normal é válida.

c) Res: Para que possamos assumir que as amostras de homens e mulheres são amostras aleatórias simples e sem substituição de suas respectivas subpopulações, é necessário que a seleção da amostra tenha sido feita de tal maneira que cada indivíduo de cada subpopulação tivesse a mesma probabilidade de ser selecionado. Na prática, isso significa que i) A amostra original foi selecionada da população total sem diferenciação explícita entre homens e mulheres, ou seja, sem estratificação prévia por sexo. Portanto, os indivíduos foram selecionados de uma única população que inclui homens e mulheres. Isso indica que a amostra não foi diretamente extraída das subpopulações de homens e mulheres, mas sim da população total. ii) Se a amostragem foi verdadeiramente aleatória, a probabilidade de selecionar um homem ou uma mulher depende apenas da proporção de homens e mulheres na população. No entanto, se quisermos tratar as subpopulações de homens e mulheres separadamente, a amostra não pode ser considerada como uma amostra aleatória simples dentro de cada subpopulação, pois a seleção foi feita da população total e não individualmente para cada subgrupo. ii) Para que a amostragem aleatória simples e sem substituição (AAS) seja válida dentro de cada subpopulação, precisaríamos ter selecionado amostras aleatórias simples separadamente para homens e mulheres. Como a seleção foi feita sobre a população total, a amostra atual pode não refletir perfeitamente as características de uma AAS para cada subpopulação, especialmente se houver uma diferença significativa entre a proporção de homens e mulheres na amostra em comparação com a população, como é o caso. Para garantir uma amostragem AAS dentro de cada subpopulação, seria necessário utilizar uma amostragem estratificada, onde a amostra é dividida em estratos (neste caso, por sexo) e selecionada aleatoriamente dentro de cada estrato.

d) Res: Seja $n = 56$: Tamanho total da amostra, $N = 1000$: Tamanho total da população, n_m: Número de mulheres na amostra, $\bar{y}_m$: Renda média das mulheres na amostra, $\hat{N}_m$: Estimativa do número total de mulheres na população - $\hat{Y}_m$: Estimativa da renda total das mulheres na população A proporção de mulheres na amostra, denotada por $\hat{p}_m$, pode ser usada para estimar a proporção de mulheres na população. Se n_f é o número de mulheres na amostra, temos:

$$\hat{p}_m = \frac{n_m}{n}.$$

A estimativa do número total de mulheres na população é dada por:

$$\hat{N}_m = N \times \hat{p}_m.$$

A renda média das mulheres na amostra $\bar{y}_m$ é usada como estimativa da renda média das mulheres na população:

$$\bar{y}_m = \frac{1}{n_m} \sum_{i=1}^{n_m} y_i,$$

UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA
PROVA II — MATD44 — GABARITO

ID da Quadra	Domicílios na Quadra	Renda Total na Quadra
AW45	120	25000
AW02	100	24000
AW31	80	19000
AW28	95	20100
AW44	80	18000

Tabela 3: Tabela das cinco quadras selecionadas

em que y_i é a renda da i-ésima mulher.

A estimativa do total da renda das mulheres na população é dada por:

$$\hat{Y}_m = \hat{N}_m \times \bar{y}_m.$$

Aqui, temos na amostra

$$n_m = 8.$$

$$\bar{y}_m = \frac{2094.90 + \cdots + 1646.21}{8} = \frac{16797.64}{8} = 2099.71.$$

A proporção estimada de mulheres na população é:

$$\hat{p}_m = \frac{8}{56} = 0,1429.$$

Então, a estimativa do número total de mulheres na população é:

$$\hat{N}_m = 1000 \times 0,1429 = 142,9 \approx 143.$$

O total da renda estimado para todas as mulheres na população é:

$$\hat{Y}_m = 143 \times 2099.71 = 300262,53 \text{ (em reais)}.$$

e) Res: Para determinar qual subpopulação é mais homogênea em relação à renda, devemos comparar as variâncias das rendas dentro de cada subpopulação. A subpopulação com menor variância será a mais homogênea. Podemos calcular a variância amostral e o coeficiente de variação para homens e mulheres e compará-las. Separamos os trabalhadores por sexo e calculamos a média, a variância e o coeficiente de variação para cada grupo.

A estimativa do número total de mulheres, $\hat{N}_m$, foi dada como:

$$\hat{N}_m = \frac{\hat{P}_m}{p_m} = \frac{16}{56} \times 1000 = 285,71$$

A estimativa do número total de homens, $\hat{N}_h$, seria:

$$\hat{N}_h = N - \hat{N}_m = 1000 - 285,71 = 714,29$$

Calculamos agora a variância estimada da renda média e o coeficiente de variação para cada subpopulação (homens e mulheres), considerando o Fator de Correção para Populações Finitas (FPC). A fórmula da variância estimada da média para amostras aleatórias simples sem reposição com FPC é:

$$\hat{\text{Var}}(\bar{y}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)s^2$$

onde n é o tamanho da amostra, N é o tamanho da população estimado, e s^2 é a variância amostral.

Variância amostral:

- Mulheres: $s_m^2 = 404.040,65$

- Homens: $s_h^2 = 301.149,13$

ariância corrigida pelo plano amostral (FPC): - Mulheres: $\hat{\text{Var}}(\bar{y}_m) = 56.305,93$ - Homens: $\hat{\text{Var}}(\bar{y}_h) = 5.724,29$

Coeficiente de variação (CV): - Mulheres: $CV_m = 0,112$ (11,2%) - Homens: $CV_h = 0,038$ (3,8%)

Os coeficientes de variação mostram que a renda das mulheres apresenta maior dispersão em relação à média do que a renda dos homens. Isso indica que a subpopulação masculina é mais homogênea em termos de renda do que a feminina. O uso do FPC corrigiu as estimativas de variância, refletindo o impacto do plano amostral, que deve ser considerado em análises amostrais em populações finitas. Aqui observamos que essa correção induz uma acentuação ainda maior desse comportamento indicando de forma mais precisa que a subpopulação mais homogênea em relação à renda é a dos homens.

3 ► **(Estimação por conglomerados)** suponha que o objetivo de uma pesquisa seja estimar a renda média em um bairro da cidade. Assuma que neste bairro existam $N_I = 60$ quadras. (A quadra ou quarteirão é a menor área de espaço urbano delimitada por ruas, rios ou avenidas, caracterizando a unidade básica de formação destes espaços)

É realizado um plano de amostragem aleatória simples de conglomerados e são selecionadas $n_I = 5$ quadras, nas quais todos os domicílios são entrevistados. Os resultados da pesquisa são apresentados na Tabela 3 a seguir:

(a) Estime a renda total dos domicílios no bairro. Reporte o coeficiente de variação estimado.

(b) Estime o número de domicílios no bairro. Reporte o coeficiente de variação estimado.

(c) Assumindo que no bairro existam $N = 2000$ domicílios, estime a renda média dos domicílios no bairro. Reporte o coeficiente de variação estimado.

(d) Estime a renda média utilizando o estimador de Hájek[1]. Explique a diferença em relação à estimativa do ponto anterior.

[1]Lembrete: O estimador de Hájek (1971) é uma forma aprimorada do estimador de razão. O estimador de Hájek para a média populacional pode ser expresso como:

$$\widehat{\overline{Y}}_{\text{H}} = \frac{\sum_{i \in s} y_i/\pi_i}{\sum_{i \in s} \frac{1}{\pi_i}}.$$

O denominador é um estimador do tamanho populacional.

a) Res: Estimativa da renda média das quadras amostradas

$$\overline{y}_s = \frac{1}{n_I}\sum_{i=1}^{n_I} y_i = \frac{25000+\cdots+18000}{5} = \frac{108100}{5} = 21620$$

Estimativa da renda total no bairro

$$\widehat{Y} = N_I \cdot \overline{y}_s = 60 \cdot 21620 = 1297200$$

Estimamos a variância amostral das rendas das quadras

$$S_y^2 = \frac{1}{n_I - 1}\sum_{i=1}^{n_I}(y_i - \overline{y}_s)^2$$

$$\begin{aligned} S_y^2 &= \frac{1}{4}\left[(25000-21620)^2 + \cdots + (18000-21620)^2\right] \\ &= \frac{1}{4}[113256400 + \cdots + 12979600] \\ &= \frac{1}{4} \times 26613000 = 6653250 \end{aligned}$$

A variância da estimativa da renda total:

$$\text{Var}(\widehat{Y}) = N_I^2 \cdot \frac{S_y^2}{n_I} = 60^2 \cdot \frac{6653250}{5} == 4798350000$$

O Coeficiente de variação estimado é

$$\text{CV} = \frac{\sqrt{\text{Var}(\widehat{Y})}}{\widehat{Y}} = \frac{\sqrt{4798350000}}{1297200} \approx \frac{69276.27}{1297200} \approx 0.0534 \text{ ou } 5.34\%$$

b) Res: Estimativa do número total de domicílios no bairro

$$\widehat{N} = N_I \cdot \overline{d}_s$$

em que $\overline{d}_s$ é a média do número de domicílios nas quadras amostradas

$$\overline{d}_s = \frac{120+100+80+95+80}{5} = \frac{475}{5} = 95$$

Logo

$$\widehat{N} = 60 \cdot 95 = 5700$$

A estimativa da variância do número total de domicílios é

$$\text{Var}(\widehat{N}) = N_I^2 \cdot \frac{S_d^2}{n_I}$$

em que S_d^2 é a variância do número de domicílios nas quadras amostradas

$$\begin{aligned} S_d^2 &= \frac{1}{4}\left[(120-95)^2 + \cdots + (80-95)^2\right] \\ &= \frac{1}{4}[625+25+225+0+225] \\ &= \frac{1}{4} \times 1075 = 268.75 \end{aligned}$$

$$\text{Var}(\widehat{N}) = 60^2 \cdot \frac{268.75}{5} = 3600 \cdot 53.75 = 193500$$

Dai, o Coeficiente de variação estimado é

$$\text{CV} = \frac{\sqrt{\text{Var}(\widehat{N})}}{\widehat{N}} = \frac{\sqrt{193500}}{5700} \approx \frac{439.2}{5700} \approx 0.0770 \text{ ou } 7.70\%$$

c) Res: A renda média estimada é:

$$\widehat{\overline{Y}} = \frac{\widehat{Y}}{N} = \frac{1297200}{2000} = 648.60$$

A Estimativa da variância da renda média é

$$\text{Var}(\widehat{\overline{Y}}) = \frac{\text{Var}(\widehat{Y})}{N^2}$$

$$\text{Var}(\widehat{\overline{Y}}) = \frac{4798350000}{2000^2} = \frac{4798350000}{4000000} = 1199.59$$

Loog, o Coeficiente de variação estimado é

$$\text{CV} = \frac{\sqrt{\text{Var}(\widehat{\overline{Y}})}}{\widehat{\overline{Y}}} = \frac{\sqrt{1199.59}}{648.60} \approx \frac{34.64}{648.60} \approx 0.0534 \text{ ou } 5.34\%$$

d) Res: Para o estimador de Hájek, precisamos das probabilidades de inclusão π_i. Assumindo amostragem aleatória simples e a probabilidade de inclusão igual, temos $\pi_i = \frac{n_I}{N_I} = \frac{5}{60} = \frac{1}{12}$.

$$\widehat{\overline{Y}}_{\text{H}} = \frac{\sum_{i\in s}\frac{y_i}{\pi_i}}{\sum_{i\in s}\frac{1}{\pi_i}}$$

Logo

$$\frac{y_i}{\pi_i} = 12 \cdot \{25000, 24000, 19000, 20100, 18000\}$$

$$\sum_{i\in s}\frac{y_i}{\pi_i} = 12\cdot(25000+24000+19000+20100+18000) = 12\cdot 108100 = 1297200$$

$$\sum_{i\in s}\frac{1}{\pi_i} = 5 \cdot 12 = 60$$

$$\widehat{\overline{Y}}_{\text{H}} = \frac{1297200}{60} = 21620$$

A estimativa da renda média usando o estimador de Hájek é igual à estimativa inicial, pois a amostragem foi simples e a probabilidade de inclusão foi constante. O estimador de Hájek é uma forma ajustada que se destaca em amostras onde as probabilidades de inclusão não são iguais.

BOA PROVA