Resolução da Prova

Questão 1: Estimação por conglomerados

Uma amostra de 5 fazendas na Bahia (conglomerados) foi selecionada por amostragem aleatória simples sem reposição de uma população de 100 fazendas para estimar a produção média de soja por hectare. Os dados da amostra são:

- Fazenda 1: 80 hectares, produção total 400 toneladas
- Fazenda 2: 120 hectares, produção total 540 toneladas
- Fazenda 3: 100 hectares, produção total 480 toneladas
- Fazenda 4: 90 hectares, produção total 450 toneladas
- Fazenda 5: 110 hectares, produção total 495 toneladas

Estime a produção total de soja por hectare ($\hat{t}$). Forneça uma expressão para a estimativa da variância.

Solução

A questão pede para estimar a "produção total de soja por hectare", o que pode ser interpretado de duas maneiras: a produção **total** de soja ($\hat{t}_y$) ou a produção **média** por hectare ($\hat{R}$). Vamos calcular ambos.

1. Estimativa do Total de Produção de Soja ($\hat{t}_y$): O estimador do total para amostragem de conglomerados em um estágio (com AAS) é dado por:

$$\hat{t}_y = \frac{N}{n} \sum_{i \in S} y_i$$

Onde $N = 100$ (total de fazendas), $n = 5$ (fazendas na amostra) e y_i é a produção da fazenda i.

Primeiro, somamos a produção das fazendas na amostra:

$$\sum_{i \in S} y_i = 400 + 540 + 480 + 450 + 495 = 2365 \text{ toneladas}$$

Agora, aplicamos a fórmula do estimador:

$$\hat{t}_y = \frac{100}{5} \times 2365 = 20 \times 2365 = \mathbf{47300\ toneladas}$$

2. Estimativa da Produção Média por Hectare ($\hat{R}$): Este é um estimador de razão, onde estimamos a produção total e dividimos pela área total.

$$\hat{R} = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} M_i}$$

Onde M_i é a área (em hectares) da fazenda i.

Somamos a área das fazendas na amostra:

$$\sum_{i \in S} M_i = 80 + 120 + 100 + 90 + 110 = 500 \text{ hectares}$$

Calculamos a razão:
$$\hat{R} = \frac{2365}{500} = \mathbf{4.73 \text{ toneladas por hectare}}$$

Expressão para a Estimativa da Variância do Total ($\hat{V}(\hat{t}_y)$): A expressão para a variância estimada do total é:
$$\hat{V}(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}$$
Onde s_y^2 é a variância amostral dos totais dos conglomerados:
$$s_y^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y}_s)^2$$
e $\bar{y}_s$ é a média dos totais dos conglomerados na amostra ($\bar{y}_s = 2365/5 = 473$).

Questão 2: Eficiência na estimação por conglomerados

No contexto da amostragem por conglomerados, como o coeficiente de correlação intraclasse impacta a eficiência do estimador de Horvitz-Thompson para a média populacional?

Solução

O coeficiente de correlação intraclasse (CCI ou ρ) mede o grau de homogeneidade dos elementos dentro dos conglomerados. Seu impacto na eficiência do estimador da média é fundamental:

- **CCI Alto (próximo de +1):** Indica que os elementos dentro de um mesmo conglomerado são muito semelhantes entre si. Isso **diminui a eficiência** (aumenta a variância) do estimador. A intuição é que, após observar uma unidade no conglomerado, as outras unidades do mesmo conglomerado fornecem pouca informação nova, tornando a amostra menos informativa para um dado número de elementos observados.
- **CCI Baixo (próximo de 0):** Indica que os conglomerados são internamente heterogêneos, ou seja, cada conglomerado tende a ser um ”mini-retrato” da população. Isso **aumenta a eficiência** (diminui a variância) do estimador. Neste cenário ideal, cada conglomerado fornece uma representação fiel da variabilidade populacional.

Em resumo, a eficiência do estimador da média em amostragem por conglomerados é **inversamente proporcional** ao valor do coeficiente de correlação intraclasse. Para maximizar a eficiência, deve-se buscar conglomerados que sejam tão heterogêneos internamente quanto possível.

Questão 3: Amostragem com probabilidade proporcional ao tamanho (PPT)

Explique por que a amostragem com probabilidade proporcional ao tamanho (PPT) é frequentemente mais eficiente do que a amostragem aleatória simples (AAS) sem reposição para estimar totais populacionais quando existe uma correlação positiva entre a variável de interesse (y) e uma variável auxiliar de tamanho x.

Solução

A maior eficiência da amostragem com PPT em relação à AAS, sob correlação positiva entre y e x, deve-se à forma como o estimador de Horvitz-Thompson lida com a variabilidade das unidades.

O estimador de total de Horvitz-Thompson é $\hat{t}_{HT} = \sum_{i \in S} \frac{y_i}{\pi_i}$. A variância deste estimador depende da variabilidade dos valores $\frac{y_i}{\pi_i}$ para as unidades da população.

1. **Na Amostragem Aleatória Simples (AAS):** A probabilidade de inclusão é a mesma para todas as unidades ($\pi_i = n/N$). Se uma unidade com um valor y_i muito grande for selecionada, o termo $\frac{y_i}{n/N}$ será muito grande, introduzindo uma alta variabilidade na estimação. A seleção ou não de uma dessas unidades "gigantes" causa grandes flutuações no valor da estimativa total, resultando em uma variância elevada.
2. **Na Amostragem com PPT:** A probabilidade de inclusão π_i é feita proporcionalmente ao tamanho x_i. Como y_i e x_i têm correlação positiva, unidades com y_i grande também terão x_i grande e, consequentemente, uma π_i grande. Isso tem um efeito estabilizador no termo $\frac{y_i}{\pi_i}$. Se a relação for quase linear ($y_i \approx c \cdot x_i$), o termo $\frac{y_i}{\pi_i}$ torna-se aproximadamente constante para todas as unidades, pois a π_i "compensa" o tamanho de y_i.

Conclusão: Ao tornar os valores ponderados ($\frac{y_i}{\pi_i}$) mais homogêneos entre si, a amostragem PPT reduz drasticamente a variância do estimador do total em comparação com a AAS. Essa redução na variância significa um aumento direto na eficiência.

Questão 4: Estimação de parâmetros lineares de totais

Uma pesquisa foi feita para estimar a diferença no número de horas semanais gastas em redes sociais por jovens de dois cursos diferentes na UFBA (A e B). Os resultados das amostragens independentes são:

- Curso A: $\widehat{t}_A = 50.000$ horas; $\widehat{V}(\widehat{t}_A) = 1.200.000$
- Curso B: $\widehat{t}_B = 42.000$ horas; $\widehat{V}(\widehat{t}_B) = 900.000$.

Estime o contraste entre os totais ($D = 3t_A - 2t_B$) e a variância dessa estimativa.

Solução

1. Estimativa do Contraste ($\hat{D}$): O estimador de uma combinação linear de parâmetros é a mesma combinação linear dos estimadores.

$$\hat{D} = 3\hat{t}_A - 2\hat{t}_B$$

Substituindo os valores dados:

$$\hat{D} = 3 \times (50.000) - 2 \times (42.000)$$

$$\hat{D} = 150.000 - 84.000 = \mathbf{66.000 \text{ horas}}$$

2. Estimativa da Variância do Contraste ($\hat{V}(\hat{D})$): Para uma combinação linear de estimadores **independentes**, a variância da combinação é a soma ponderada das variâncias, onde os pesos são os coeficientes ao quadrado.

$$\hat{V}(\hat{D}) = \hat{V}(3\hat{t}_A - 2\hat{t}_B) = 3^2\hat{V}(\hat{t}_A) + (-2)^2\hat{V}(\hat{t}_B)$$

Substituindo os valores dados:

$$\hat{V}(\hat{D}) = 9 \times (1.200.000) + 4 \times (900.000)$$
$$\hat{V}(\hat{D}) = 10.800.000 + 3.600.000 = \mathbf{14.400.000}$$

Questão 5: Estimação de parâmetros não-lineares de totais

Estamos interessados em estimar a variância de um estimador do parâmetro populacional não-linear (produto de totais) $\theta = f(t_y, t_z) = t_y \cdot t_z$, através do produto de dois estimadores de Horvitz-Thompson de totais $\widehat{\theta} = \widehat{t}_{\pi y} \cdot \widehat{t}_{\pi z}$. Descreva passo a passo como poderia aproximar a variância de $\hat{\theta}$.

Solução

Para aproximar a variância de um estimador não-linear como $\hat{\theta} = \hat{t}_y \cdot \hat{t}_z$, utilizamos o **Método da Linearização de Taylor** (ou Método Delta). O procedimento é o seguinte:

Passo 1: Definir a função e o estimador A função dos parâmetros é $f(t_y, t_z) = t_y \cdot t_z$. O estimador é $\hat{\theta} = f(\hat{t}_y, \hat{t}_z) = \hat{t}_y \cdot \hat{t}_z$.

Passo 2: Linearizar a função Aproximamos a função f usando uma expansão de Taylor de primeira ordem em torno dos verdadeiros valores dos totais (t_y, t_z):

$$\hat{\theta} \approx f(t_y, t_z) + (\hat{t}_y - t_y) \frac{\partial f}{\partial t_y}\bigg|_{(t_y, t_z)} + (\hat{t}_z - t_z) \frac{\partial f}{\partial t_z}\bigg|_{(t_y, t_z)}$$

Passo 3: Calcular as derivadas parciais Calculamos as derivadas parciais da função $f(t_y, t_z) = t_y \cdot t_z$ e as avaliamos nos pontos (t_y, t_z):

$$\frac{\partial f}{\partial t_y} = t_z$$
$$\frac{\partial f}{\partial t_z} = t_y$$

Passo 4: Construir a aproximação linear Substituímos as derivadas na expansão de Taylor:

$$\hat{\theta} \approx t_y t_z + (\hat{t}_y - t_y) t_z + (\hat{t}_z - t_z) t_y$$

Isso pode ser reorganizado como:

$$\hat{\theta} - t_y t_z \approx t_z(\hat{t}_y - t_y) + t_y(\hat{t}_z - t_z)$$

Passo 5: Calcular a variância da aproximação A variância de $\hat{\theta}$ é aproximada pela variância do lado direito da expressão linearizada. Como t_y e t_z são constantes, temos:

$$V(\hat{\theta}) \approx V\left(t_z \hat{t}_y + t_y \hat{t}_z\right)$$

Aplicando as propriedades da variância:

$$V(\hat{\theta}) \approx t_z^2 V(\hat{t}_y) + t_y^2 V(\hat{t}_z) + 2 t_y t_z \mathrm{Cov}(\hat{t}_y, \hat{t}_z)$$

Passo 6: Estimar a variância Para obter um estimador da variância, substituímos todos os parâmetros populacionais desconhecidos (t_y, t_z, $V(\hat{t}_y)$, $V(\hat{t}_z)$, $\mathrm{Cov}(\hat{t}_y, \hat{t}_z)$) por seus respectivos estimadores amostrais:

$$\hat{V}(\hat{\theta}) \approx \hat{t}_z^2 \hat{V}(\hat{t}_y) + \hat{t}_y^2 \hat{V}(\hat{t}_z) + 2 \hat{t}_y \hat{t}_z \widehat{\mathrm{Cov}}(\hat{t}_y, \hat{t}_z)$$

Os termos $\hat{V}(\hat{t}_y)$, $\hat{V}(\hat{t}_z)$ e $\widehat{\mathrm{Cov}}(\hat{t}_y, \hat{t}_z)$ devem ser calculados usando as fórmulas apropriadas para o plano amostral utilizado (por exemplo, as fórmulas de variância e covariância de Horvitz-Thompson ou Yates-Grundy).