AMOSTRAGEM

Unidade 09

Amostragem Estratificada

Amostragem Estratificada (AE)

Processo de amostragem que requer:

(1) Dividir a população U em H grupos disjuntos e exaustivos, geralmente mais **homogêneos**, chamados ***estratos.***

(2) Selecionar amostras dentro de cada um dos estratos, independentemente.

(3) Estimar parâmetros em cada estrato.

(4) Agregar estimativas para o conjunto da população.

Vantagens da AE

- Pode aumentar a precisão das estimativas para o conjunto da população.
- Garante observação de amostras nos estratos formados.
- Permite estimação para subgrupos da população da pesquisa com eficiência e precisão controlada.
- Pode ser operacionalmente e/ou administrativamente mais conveniente.

Desvantagens / Requisitos da AE

- Requer conhecimento das variáveis de estratificação para todas as unidades do cadastro antes da amostragem.
- Requer re-estruturação do cadastro antes da amostragem.
- Apenas uma estratificação possível.
- Dividir a população em muitos estratos pode levar a ter amostras muito pequenas em cada estrato.

Razões para Estratificar uma População

1. Estratos formam domínios "naturais" ou substantivos de interesse. Por exemplo: regiões geográficas; farmácias e lojas de departamentos; homens e mulheres; etc.
2. Para "espalhar" a amostra sobre toda a população, isto é, para fazer a amostra "representativa".
3. Para melhorar a eficiência amostral, isto é, para reduzir a variância dos estimadores.

Tipos de Estratificação

1. **Natural**

   Estratos iguais a subgrupos da população para os quais se requer estimativas com precisão controlada.

2. **Estatística**

   Estratos definidos como subgrupos homogêneos da população, visando aumentar eficiência na estimação para a população como um todo.

   Neste caso, não há interesse específico na estimação de parâmetros dos estratos formados.

Fatores que Influenciam a Eficiência na AE

1. Escolha da(s) variável(is) de estratificação.
2. Número de estratos.
3. Determinação dos limites dos estratos.
4. Alocação da amostra nos estratos.
5. Método de seleção em cada estrato.

O Método Geral

1. Particione (divida) U em H subconjuntos (grupos) ***mutuamente exclusivos e exaustivos***, chamados ***estratos***, e denotados U_1, ..., U_h, ..., U_H, de modo que:
   - $U = U_1 \cup U_2 \cup \ldots \cup U_H = \bigcup_{h=1}^{H} U_h$ e
   - $U_h \cap U_k = \varnothing$, $h \neq k$.

Então U_h = { i : unidade i pertence ao estrato h}, para h=1,2,...,H.

Seja N_h o tamanho de U_h . Então $N_1 + N_2 + \ldots + N_H = N$.

Amostragem Estratificada

2. Selecione uma amostra s_h de tamanho n_h, com $n_h>0$, segundo um plano amostral $p_h(s_h)$ ***independentemente dentro de cada estrato*** h, onde h=1,2,...,H, e $\sum_{h=1}^{H} n_h = n$.

Assim, fica assegurado que cada estrato tem sua população representada na amostra completa dada por:

- $s = s_1 \cup s_2 \cup \ldots \cup s_H$.

Amostragem Estratificada

A independência da amostragem nos estratos consiste em tratar cada estrato como se fosse uma população separada, para fins de sorteio da amostra.

Devido à independência da seleção nos estratos, temos:

$$p(s) = p_1(s_1) \times p_2(s_2) \times \ldots \times p_H(s_H) .$$

Diferentes planos amostrais podem ser empregados nos diversos estratos, embora isso seja pouco comum na prática.

O mais comum é usar um mesmo tipo de sorteio nos vários estratos.

Exemplos

Exemplo 9.1: População dividida em 2 estratos. AAS usada no estrato 1, com Amostragem Binomial usada no estrato 2.

Exemplo 9.2: Amostragem Estratificada por Corte (AEC)

População dividida em dois estratos. Num se faz um censo, isto é, se pesquisa o conjunto completo de unidades ali existentes, e no outro se faz AAS.

Exemplo 9.3: Amostragem Estratificada Simples (AES)

Amostras aleatórias simples selecionadas em cada um dos estratos definidos.

Critério de Eficiência

Para conseguir ganhar eficiência com o uso da estratificação, a idéia é tornar os valores da(s) variável(is) de estudo dentro de cada estrato o mais similares / homogêneos possíveis, isto é, **minimizar a variância dentro dos estratos**.

Para isso é fundamental ter acesso a cadastro com variáveis auxiliares que possam ser usadas para estratificar a população de forma eficiente.

Amostragem Estratificada Simples (AES)

Trata-se do caso especial em que AAS é empregada em todos os estratos.

Neste caso, os tamanhos N_h de cada um dos estratos U_h são considerados conhecidos.

O cadastro deve permitir separar as unidades da população nos H estratos definidos.

Esquema de Seleção

Selecione uma AAS de tamanho n_h>0 das N_h unidades do estrato U_h, h=1,2,...,H.

Então:

$$p_h(s_h) = 1 / \binom{N_h}{n_h} = \binom{N_h}{n_h}^{-1}, h = 1, \ldots, H, \text{ e}$$

$$p(s) = \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1}.$$

Tamanhos total da amostra: $n_1 + n_2 + \ldots + n_H = n$

Notação

Para facilitar a apresentação das fórmulas, é costume re-identificar as unidades populacionais usando dois rótulos.

- Um rótulo h ($h=1,...,H$) é usado para indicar o estrato a que pertence a unidade; e
- Um rótulo i ($i=1,...,N_h$) para indicar a unidade dentro do estrato.

Assim, um valor típico da variável de pesquisa é y_{hi}, para $i=1,...,N_h$ e $h=1,...,H$.

Dados Populacionais

Estrato	Tamanho do Estrato: N_h	Dados
1	N_1	$y_{11}, \ldots, y_{1N_1}$
$\vdots$	$\vdots$	$\vdots$
h	N_h	$y_{h1}, \ldots, y_{hN_h}$
$\vdots$	$\vdots$	$\vdots$
H	N_H	$y_{H1}, \ldots, y_{HN_H}$

Parâmetros nos Estratos

Tamanhos populacionais: $N_1 + N_2 + \ldots + N_H = N$

(1) Total $$Y_h = \sum_{i=1}^{N_h} y_{hi} = \sum_{i \in U_h} y_{hi}$$

(2) Média $$\overline{Y}_h = Y_h / N_h$$

(3) Variância $$S_h^2 = \sum_{i=1}^{N_h} (y_{hi} - \overline{Y}_h)^2 / (N_h - 1)$$

Parâmetros Populacionais (Globais)

(1) Total $$Y = \sum_{h=1}^{H} Y_h = \sum_{h=1}^{H} N_h \overline{Y}_h$$

(2) Média $$\overline{Y} = Y / N = \sum_{h=1}^{H} N_h \overline{Y}_h / N \Rightarrow$$

$$\overline{Y} = \sum_{h=1}^{H} W_h \overline{Y}_h \quad \text{, com } W_h = N_h / N .$$

Parâmetros Populacionais (Globais)

(3) Variância

$$S_y^2 = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\left(y_{hi}-\overline{Y}\right)^2 / (N-1)$$

$$= \sum_{h=1}^{H}\sum_{i=1}^{N_h}\left[\left(y_{hi}-\overline{Y}_h\right)+\left(\overline{Y}_h-\overline{Y}\right)\right]^2 / (N-1)$$

$$= \sum_{h=1}^{H}\left[(N_h-1)S_h^2 + N_h\left(\overline{Y}_h-\overline{Y}\right)^2\right] / (N-1)$$

Isto é:

Variância **Total** = Variância **Dentro** + Variância **Entre**

Nota

Para S_y^2 fixado, maximizar a variância ENTRE

$$\sum_{h=1}^{H} N_h\left(\overline{Y}_h-\overline{Y}\right)^2$$

minimiza a variância DENTRO

$$\sum_{h=1}^{H}(N_h-1)\,S_h^2 .$$

Dados Amostrais

Estrato	Tamanho amostral n_h	Dados amostrais
1	n_1	$y_{11}, \ldots, y_{1n_1}$
$\vdots$	$\vdots$	$\vdots$
h	n_h	$y_{h1}, \ldots, y_{hn_h}$
$\vdots$	$\vdots$	$\vdots$
H	n_H	$y_{H1}, \ldots, y_{Hn_H}$

Estimação

Como a amostragem é feita independentemente por estrato, podemos estimar separadamente os parâmetros de cada estrato.

Sob AES, os estimadores usuais dos parâmetros nos estratos são:

(1) Total $$\hat{Y}_h = \sum_{i=1}^{n_h} w_{hi} y_{hi} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = N_h \bar{y}_h$$

Nota: o peso $w_{hi} = w_h = N_h/n_h$ é o inverso da probabilidade de inclusão para unidades dentro de cada estrato h sob AES.

Estimação

(2) Média $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$.

(3) Variância $s_h^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 / (n_h - 1)$

Propriedades:

$E_{AES}(\bar{y}_h) = \bar{Y}_h$; $E_{AES}(\hat{Y}_h) = Y_h$ e $E_{AES}(s_h^2) = S_h^2$.

Prova: temos AAS de n_h unidades dentro do estrato h.

Estimação de Parâmetros Populacionais (Globais)

(1) Total $\hat{Y}_{AES} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} N_h \bar{y}_h$

(2) Média $\bar{y}_{AES} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$.

Nota: Raramente é necessário estimar a variância global S_y^2.

Se fosse necessário, como você faria isso?

Exercício 9.1

Mostre que a média amostral global

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} y_{hi}$$

pode ser escrita como

$$\bar{y} = \sum_{h=1}^{H} \frac{n_h}{n}\bar{y}_h \neq \sum_{h=1}^{H} \frac{N_h}{N}\bar{y}_h = \bar{y}_{AES},$$

a menos que $\frac{n_h}{n} = \frac{N_h}{N}, \forall h = 1, \ldots, H.$

Nota

Um plano com $\frac{n_h}{n} = \frac{N_h}{N}\ \forall h$ é chamado de plano estratificado proporcional ou auto-ponderado.

Exercício 9.2

Quais são as probabilidades de inclusão de primeira e segunda ordem para unidades na população sob AES?

Que valores estas probabilidades assumem em caso de um plano AES estratificado proporcional ou auto-ponderado?

Propriedades de $\bar{y}_{AES}$ Sob AES

O estimador de média é não viciado sob AES, isto é:

$$E_{AES}(\bar{y}_{AES}) = \bar{Y} .$$

Isto segue porque $E_{AES}(\bar{y}_h) = \bar{Y}_h$, para h=1,...,H, e

$$E_{AES}\left(\sum_{h=1}^{H} W_h \bar{y}_h\right) = \sum_{h=1}^{H} W_h E_{AES}(\bar{y}_h) = \sum_{h=1}^{H} W_h \bar{Y}_h = \bar{Y}.$$

Propriedades de $\bar{y}_{AES}$ Sob AES

A variância do estimador $\bar{y}_{AES}$ pode ser obtida notando que

$$V_{AES}\left(\sum_{h=1}^{H} W_h \bar{y}_h\right) = \sum_{h=1}^{H} W_h^2 V_{AES}(\bar{y}_h).$$

Isto segue devido à independência da amostragem nos estratos, que implica em $COV_{AES}(\bar{y}_h, \bar{y}_k) = 0$, $h \neq k$.

Propriedades de $\overline{y}_{AES}$ Sob AES

$$V_{AES}(\overline{y}_{AES}) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$
$$= \sum_h W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$$

Um estimador não viciado da variância é dado por

$$\hat{V}_{AES}(\overline{y}_{AES}) = \sum_h W_h^2 \hat{V}_{AES}(\overline{y}_h)$$
$$= \sum_h W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2$$

Resumo da Estimação de Totais Dentro de Estratos

1. $\hat{Y}_h = N_h \overline{y}_h$ estima $Y_h = N_h \overline{Y}_h$.
2. $V_{AES}(\hat{Y}_h) = N_h^2 V_{AES}(\overline{y}_h)$.
3. $\hat{V}_{AES}(\hat{Y}_h) = N_h^2 \hat{V}_{AES}(\overline{y}_h)$.

Estimação de Totais p/ o Conjunto da População: Resumo

$$\hat{Y}_{AES} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} N_h \bar{y}_h$$

$$V_{AES}\left(\hat{Y}_{AES}\right) = N^2 \sum_h W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_h^2$$

$$\hat{V}_{AES}\left(\hat{Y}_{AES}\right) = N^2 \sum_h W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s_h^2$$

Intervalos de Confiança

1. Se $n = \sum_h n_h$ for grande, então o Teorema Central do Limite também se aplica. Portanto:

$$z = \frac{\bar{y}_{AES} - \bar{Y}}{\sqrt{\hat{V}_{AES}\left(\bar{y}_{AES}\right)}} \approx N(0;1)$$

O intervalo de confiança de nível 1-α para $\bar{Y}$ é dado por

$$\bar{y}_{AES} \mp z_{\alpha/2} \sqrt{\hat{V}_{AES}\left(\bar{y}_{AES}\right)}$$

Intervalos de Confiança

2. Para médias dentro de estratos, $\bar{y}_h$, os tamanhos de amostras por estratos n_h devem ser grandes. Nesse caso:

$$z = \frac{\bar{y}_h - \bar{Y}_h}{\sqrt{\hat{V}_{AES}(\bar{y}_h)}} \approx N(0,1)$$

e então um intervalo de confiança de nível 1-α para $\bar{Y}_h$ é dado por

$$\bar{y}_h \mp z_{\alpha/2}\sqrt{\hat{V}_{AES}(\bar{y}_h)}$$