Capítulo 12 Amostragem Estratificada

Amostragem Estratificada (AE) é um processo de amostragem que usa alguma informação relevante que deve estar disponível para todos os elementos da população para *dividir a população* U *em* H *grupos* disjuntos e exaustivos, geralmente mais *homogêneos*, chamados *estratos* e com a *seleção de amostras dentro de cada um dos estratos*, independentemente.

As variáveis de estratificação podem ser geográficas ou não-geográficas (rendimento, idade, sexo, número de empregados,etc.), sendo limitadas àquelas informações que estão disponíveis no cadastro.

Os parâmetros são estimados em cada estrato, de acordo com o esquema de seleção adotado e as estimativas são agregadas para o conjunto da população.

Dentre as vantagens da AE destacam-se:

- a possibilidade de aumentar a precisão das estimativas para o conjunto da população;
- a garantia de observação de amostras em todos os estratos formados;
- a possibilidade de estimação para subgrupos da população da pesquisa com eficiência e precisão controladas; e
- a viabilidade de ser operacionalmente e/ou administrativamente mais conveniente.

Por outro lado, a AE requer conhecimento das variáveis de estratificação para todas as unidades do cadastro antes da amostragem e a re-estruturação do cadastro antes da amostragem. Apenas uma estratificação é possível e dividir a população em muitos estratos pode levar a ter amostras muito pequenas em cada estrato.

Seguem as 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”e assegurar que todas as partes (estratos) da população estão representados na amostra.
3. Para melhorar a eficiência amostral, isto é, para reduzir a variância dos estimadores quanto maior for a homogeneidade dentro dos estratos.

4. Quando for necessário usar métodos diferentes de coleta em diferentes subgrupos da população.

Podemos distinguir dois tipos de estratificação:

1. *Natural* - quando os estratos são iguais a subgrupos da população para os quais se requer estimativas com precisão controlada.

2. *Estatística* - quando os estratos são 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.

Há diversos fatores que influenciam a eficiência na AE: a escolha da(s) variável(is) de estratificação; o número de estratos; a determinação dos limites dos estratos; a alocação da amostra nos estratos; e o método de seleção da amostra em cada estrato.

Para *estratificação natural* a escolha da(s) variável(is) de estratificação se dá considerando TODAS as variáveis disponíveis com as quais são definidos os estratos naturais ou domínios de interesse da pesquisa.

No caso da *estratificação estatística*, escolha entre as variáveis disponíveis as que são *melhores preditoras* das variáveis de interesse da pesquisa.

Para conseguir ganhar eficiência com o uso da estratificação, a ideia é tornar os valores da(s) variável(is) de estudo dentro de cada estrato os 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.

12.1 O método geral

Particione (divida) U em H subconjuntos (grupos) *mutuamente exclusivos e exaustivos*, chamados *estratos*, e denotados por $U_1, \ldots, U_h, \ldots, U_H$, de modo que:

$$U = U_1 \cup U_2 \cup \cdots \cup U_H = \bigcup_{h=1}^{H} U_h \text{ e } U_h \cap U_k = \emptyset,\ h \neq k.$$

Então $U_h = \{\text{unidades pertencentes ao estrato h}\}$, para $h = 1, 2, \ldots, H$.

Seja N_h o tamanho de U_h. Então $N_1 + N - 2 + \cdots + N_H = N$.

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

Assim, fica assegurado que cada estrato tem sua população representada na amostra completa dada por: $s = a_1 \cup a_2 \cup \cdots \cup a_H$.

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: $p(s) = p_1(a_1) \times p_2(a_2) \times \cdots \times p_H(a_H)$.

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.

Exemplo 12.1 População dividida em 2 estratos. AAS usada no estrato 1, com Amostragem Binomial usada no estrato 2.

Exemplo 12.2 Amostragem Estratificada por Corte (AEC)

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

Exemplo 12.3 Amostragem Estratificada Simples (AES)

Amostras aleatórias simples selecionadas em cada um dos estratos definidos.

12.2 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.

12.2.1 Esquema de seleção

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

Então:

$$p_h(a_h) = 1 \Big/ \binom{N_h}{n_h} = \binom{N_h}{n_h}^{-1}, \quad h = 1, 2, \ldots, H$$

e

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

O *tamanho total da amostra* é: $n_1 + n_2 + \cdots + n_H = n$.

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

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

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

A Tabela 12.1 representa uma população estratificada com os tamanhos dos estratos e os *dados populacionais* da variável y.

Tabela 12.1: Tamanho do estrato e dados populacionais da variável y por estrato

Estrato	Tamanho do estrato	Dados populacionais
1	N_1	$y_{11}, \cdots, y_{1N_1}$
⋮	⋮	⋮
h	N_h	$y_{h1}, \cdots, y_{hN_h}$
⋮	⋮	⋮
H	N_H	$y_{H1}, \cdots, y_{HN_H}$

12.2.2 Parâmetros populacionais

Seguem alguns parâmetros populacionais nos estratos e em toda população.

Parâmetros nos estratos:

A seguir são apresentados, respectivamente, o total do estrato h, a média do estrato h e a variância do estrato h da variável y:

$$Y_h = \sum_{i=1}^{N_H} y_{hi} = \sum_{i \in U_h} y_{hi}$$

$$\bar{Y}_h = Y_h / N_h$$

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

Parâmetros populacionais (globais):

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

O total populacional é dado por:

$$Y = \sum_{h=1}^{H} Y_h = \sum_{h=1}^{H} N_h \bar{Y}_h$$

A média do total populacional é definida como:

$$\bar{Y} = Y/N = \sum_{h=1}^{H} N_h \bar{Y}_h / N = \sum_{h=1}^{H} W_h \bar{Y}_h$$

com $W_h = N_h / N$

e a variância do total populacional pode ser escrita como:

$$\begin{aligned} S_y^2 &= \sum_{h=1}^{H} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2 / (N-1) \\ &= \sum_{h=1}^{H} \sum_{i=1}^{N_h} [(y_{hi} - \bar{Y}_h) + (\bar{Y}_h - \bar{Y})]^2 / (N-1) \\ &= \sum_{h=1}^{H} \frac{N_h - 1}{N-1} S_h^2(y) + \sum_{h=1}^{H} \frac{N_h}{N-1} (\bar{Y}_h - \bar{Y})^2 \end{aligned}$$

Isto é: Variância *Total* = Variância *Dentro* + Variância *Entre*,

sendo $\text{Variância Dentro} = \sum_{h=1}^{H} \frac{N_h - 1}{N-1} S_h^2(y)$ e $\text{Variância Entre} = \sum_{h=1}^{H} \frac{N_h}{N-1} (\bar{Y}_h - \bar{Y})^2$.

Para S_y^2 fixado, maximizar a variância *Entre* minimiza a variância *Dentro*.

12.2.3 Dados amostrais

A notação para os dados amostrais é semelhante à usada para denotar os dados populacionais. A Tabela 12.2 representa o tamanho amostral e os *dados amostrais* da variável y de cada estrato de uma AES.

Tabela 12.2: Tamanho amostral e dados amostrais da variável y por estrato

Estrato	Tamanho amostral	Rótulos na amostra	Dados amostrais
1	n_1	$a_1 = \{i_1, ..., i_{n_1}\}$	$y_{1i_1}, \cdots, y_{1i_{n_1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	n_h	$a_h = \{i_1, ..., i_{n_h}\}$	$y_{hi_1}, \cdots, y_{hi_{n_h}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
H	n_H	$a_H = \{i_1, ..., i_{n_H}\}$	$y_{Hi_1}, \cdots, y_{Hi_{n_H}}$

12.2.4 Estimação do total e da média populacional

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 descritos a seguir:

$\hat{Y}_h = \sum_{i \in a_h} w_{hi} y_{hi} = \frac{N_h}{n_h} \sum_{i \in a_h} y_{hi} = N_h \bar{y}_h$ é o estimador do total do estrato $h$da variável y.

Nota-se que 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.

$\bar{y}_h = \frac{1}{n_h} \sum_{i \in a_h} y_{hi}$ é o estimador da média do estrato h da variável y; e

$s_h^2(y) = \frac{1}{n_h - 1} \sum_{i \in a_h} (y_{hi} - \bar{y}_h)^2$ é o estimador da variância do estrato h da variável y.

Como temos AAS de n_h unidades dentro do estrato h, são válidas as seguintes propriedades:

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

$V_{AES}\left(\bar{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right)S_h^2(y)$ é a variância do estimador da média do estrato h da variável y;

$V_{AES}\left(\hat{Y}_h\right) = N_h^2 V_{AES}\left(\bar{y}_h\right)$ é a variância do estimador de total do estrato h da variável y;

$\hat{V}_{AES}\left(\bar{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right)s_h^2(y)$ é o estimador da variância do estimador da média do estrato h da variável y; e

$\hat{V}_{AES}\left(\hat{Y}_h\right) = N_h^2 \hat{V}_{AES}\left(\bar{y}_h\right)$ é o estimador da variância do estimador de total do estrato h da variável y.

Na estimação de parâmetros populacionais (globais), ou seja, para o conjunto da população temos:

O estimador do total Y é dado por:

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

e o estimador da média $\bar{Y}$ é dado por:

$$\bar{y}_{AES} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

Nota: Raramente é necessário estimar a variância global S_y^2. Se fosse necessário, como você faria isso?

O estimador de média é não viciado sob AES, isto é: $E_{AES}\left(\bar{y}_{AES}\right) = \bar{Y}$.

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

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

A variância do estimador $\bar{y}_{AES}$ pode ser obtida por:

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

Isto segue devido à independência da amostragem nos estratos, que implica em

$COV_{AES}(\bar{y}_h, \bar{y}_k) = 0,\ h \neq k.$

Então, $V_{AES}\left(\bar{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2(y)$ e $V_{AES}\left(\hat{Y}_{AES}\right) = N^2 \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2(y).$

Um estimador não viciado da variância de $\bar{y}_{AES}$ é dado por:

$$\hat{V}_{AES}\left(\bar{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2(y)$$

No caso da estimação de total $\hat{Y}_{AES}$, um estimador não viciado da variância é dado por:

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

A Tabela 12.3 apresenta um resumo da estimação de parâmetros média e total da variável y sob AES.

Tabela 12.3: Estimadores dos parâmetros média e total sob AES

Parâmetro	Estimador AES
$\bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_h$	$\bar{y}_{AES} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$
$Y = \sum_{h=1}^{H} Y_h = \sum_{h=1}^{H} N_h \bar{Y}_h$	$\hat{Y}_{AES} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} N_h \bar{y}_h$
$V_{AES}\left(\bar{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2(y)$	$\hat{V}_{AES}\left(\bar{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2(y)$
$V_{AES}\left(\hat{Y}_{AES}\right) = N^2 V_{AES}\left(\bar{y}_{AES}\right)$	$\hat{V}_{AES}\left(\hat{Y}_{AES}\right) = N^2 \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2(y)$

12.2.5 Intervalos de confiança

Se $n = \sum_{h=1}^{H} n_h$ 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 $1 - \alpha$ para $\bar{Y}$ é dado por:

$$IC_{AES}(\bar{y}_{AES}; 1 - \alpha) = \left[\bar{y}_{AES} \mp z_{\alpha/2} \sqrt{\hat{V}_{AES}\left(\bar{y}_{AES}\right)}\right]$$

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

Nesse caso:

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

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

$$IC_{AES}(\bar{y}_h; 1 - \alpha) = \left[\bar{y}_h \mp z_{\alpha/2}\sqrt{\hat{V}_{AES}\left(\bar{y}_h\right)}\right]$$

12.3 Estimadores de razão em amostragem estratificada simples

No capítulo 6 foi tratado o caso de utilização de estimador de razão para estimar o total populacional (Y) a partir de uma amostra aleatória simples sem reposição de tamanho n. No caso de uma amostra estratificada, há dois estimadores de razão usuais para estimar o total populacional (Y):

- estimador de razão combinada; e
- estimador de razão separada.

Considere então, o problema de estimar o total Y a partir de uma amostra aleatória estratificada selecionada de uma população com H estratos de tamanhos N_h $(h = 1, \ldots, L)$, tendo sido selecionadas n_h unidades e investigadas as variáveis x e y em cada unidade da amostra de cada estrato.

12.3.1 Estimador de razão combinada

Suponha que seja também conhecido, de alguma fonte externa a amostra, o total populacional (X) para a variável x.

O *estimador de razão combinada* $\hat{Y}_{RC}$ para estimar o total populacional Y é definido por:

$$\hat{Y}_{RC} = \frac{\hat{Y}_{AES}}{\hat{X}_{AES}}X = \frac{\bar{y}_{AES}}{\bar{x}_{AES}}X$$

É sabido que os estimadores de razão são viciados exceto se a população for de um tipo muito especial em termos da relação entre x e y. Apesar disso, tem-se afirmado que em muitos casos o estimador de razão é preferível ao estimador natural (simples) por que dá melhor precisão. Entretanto, esta afirmação só é verdadeira quando se consegue tornar desprezível o vício do estimador de razão,

De acordo com Cochran (1977), é usual considerar o vício desprezível quando

$$CV(\hat{X}_{AES}) = CV(\bar{x}_{AES}) \leq 0,10.$$

Há que notar a equivalência de fixar um coeficiente de variação de 10% para $\bar{x}_{AES}$ e de admitir um erro máximo de 20% na estimação de $\bar{X}$ com 95% de confiança.

Não se dispõe de uma expressão exata para a variância do estimador de razão combinada. Porém, se a amostra é de tamanho suficientemente grande para tornar o vício desprezível, pode-se obter uma expressão aproximada para a variância $\hat{Y}_{RC}$ dada por Cochran (1977), p. 166:

$$V_{AES}(\hat{Y}_{RC}) \doteq \sum_{h=1}^{H} \frac{N_h^2(1-f_h)}{n_h}\left(S_h^2(y) + R^2 S_h^2(x) - 2RS_h(x,y)\right)$$

onde:

$f_h = \frac{n_h}{N_h}$, $R = \frac{Y}{X}$ e

$S_h(x,y) = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(x_{hi} - \bar{X}_h)(y_{hi} - \bar{Y}_h)$ é a covariância populacional do estrato h das variáveis x e y.

Um estimador de $V(\hat{Y}_{RC})$ é dado por:

$$\hat{V}_{AES}(\hat{Y}_{RC}) = \sum_{h=1}^{H} \frac{N_h^2(1-f_h)}{n_h}\left(s_h^2(y) + \hat{R}_{AES}^2 s_h^2(x) - 2\hat{R}_{AES} s_h(x,y)\right)$$

onde: $\hat{R}_{AES} = \frac{\bar{y}_{AES}}{\bar{x}_{AES}}$

$s_h^2(y)$, $s_h^2(x)$ e $s_h(x,y)$ são estimadores não viciados de $S_h^2(y)$, $S_h^2(x)$ e $S_h(x,y)$, respectivamente.

$$s_h(x,y) = \frac{1}{n_h - 1} \sum_{i \in a_h} (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h).$$

O estimador de razão combinada para estimar a média $\bar{Y}$ é dado por:

$$\bar{y}_{RC} = \frac{\hat{Y}_{RC}}{N}$$

Neste caso, a variância $V(\bar{y}_{RC})$ é dada por:

$$V_{AES}(\bar{y}_{RC}) = \frac{1}{N^2} V_{AES}(\hat{Y}_{RC})$$

e um estimador de $V(\bar{y}_{RC})$ é dado por:

$$\hat{V}_{AES}(\bar{y}_{RC}) = \frac{1}{N^2} \hat{V}_{AES}(\hat{Y}_{RC})$$

12.3.2 Estimador de razão separada

Uma outra forma de utilizar estimadores e razão para conseguir maior precisão na amostragem estratificada é o chamado *estimador de razão separada*.

O *estimador de razão separada* $\hat{Y}_{RS}$ para estimar o total populacional Y é definido por:

$$\hat{Y}_{RS} = \sum_{h=1}^{H} \frac{\bar{y}_h}{\bar{x}_h} X_h = \sum_{h=1}^{H} \hat{R}_h X_h$$

sendo $\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h} = \frac{y_h}{x_h}$.

Note-se que é necessário conhecer os totais por estrato X_h da variável auxiliar x.

A principal diferença do estimador de razão separada para o estimador de razão combinada está no nível em que se faz o uso da estimação por razão: no estimador de razão separada são feitas razões em cada um dos estratos, enquanto no estimador de razão combinada uma única razão é feita para o estimador de total.

Quanto ao vício, este estimador precisa ser analisado com maior cuidado, porque depende de razões construídas em cada um dos estratos.

Se os n_h forem suficientemente grandes, pode-se admitir que o vício de $\hat{Y}_{RS}$ é desprezível. Caso isto não aconteça o uso deste estimador não é aconselhável.

De acordo com Cochran (1977), no uso do estimador de razão separada, há que verificar se:

$$\sqrt{H}\, CV_{AES}(\bar{x}_h) \leq 0,30 \quad \forall h$$

Supondo os n_h suficientemente grandes, a variância $\hat{Y}_{RS}$ é aproximada por:

$$V_{AES}(\hat{Y}_{RS}) \doteq \sum_{h=1}^{H} \frac{N_h^2(1-f_h)}{n_h}\left(S_h^2(y) + R_h^2 S_h^2(x) - 2R_h S_h(x,y)\right)$$

sendo $R_h = \dfrac{Y_h}{X_h}$.

Um estimador de $V(\hat{Y}_{RS})$ é dado por:

$$\hat{V}_{AES}(\hat{Y}_{RS}) = \sum_{h=1}^{H} \frac{N_h^2(1-f_h)}{n_h}\left(s_h^2(y) + \hat{R}_h^2 s_h^2(x) - 2\hat{R}_h s_h(x,y)\right)$$

O estimador de razão separada para estimar a média $\bar{Y}$ é dado por:

$$\bar{y}_{RS} = \frac{\hat{Y}_{RS}}{N}$$

Neste caso, a variância $V(\bar{y}_{RS})$ é dada por:

$$V_{AES}(\bar{y}_{RS}) = \frac{1}{N^2} V_{AES}(\hat{Y}_{RS})$$

e um estimador de $V(\bar{y}_{RS})$ é dado por:

$$\hat{V}_{AES}(\bar{y}_{RS}) = \frac{1}{N^2} \hat{V}_{AES}(\hat{Y}_{RS})$$

12.3.3 Comparação dos estimadores de razão combinada e separada

Em geral, para amostras de tamanhos idênticos, o estimador de razão combinada deve ter vício menor que o estimador de razão separada.

Em ambos os casos, os tamanhos de amostra que garantam um vício desprezível podem ser determinados.

Através da comparação das variâncias pode ser feita a avaliação da melhor precisão alcançada entre os estimadores de razão em amostragem estratificada:

$$V_{AES}(\hat{Y}_{RC}) - V_{AES}(\hat{Y}_{RS}) \doteq \sum_{h=1}^{H} \frac{N_h^2(1-f_h)}{n_h}\left[(R^2 - R_h^2)S_h^2(x) - 2(R - R_h)S_h(x,y)\right]$$

Os dois estimadores serão igualmente precisos se $R_h = R$ ou $Y_h/X_h = Y/X$ para todos os estratos.

A medida que os R_h sejam mais distantes de R, o estimador de razão separada tende a dar maior precisão, inclusive por se basear num conhecimento mais detalhado dos dados do universo da variável x.

12.4 Alocação da amostra pelos estratos

Uma importante consideração na amostragem estratificada é a forma na qual a amostra total é alocada em cada estrato, podendo ser feita de forma proporcional ou desproporcional:

- *proporcional*: quando a distribuição da amostra total é proporcional ao número de unidades em cada estrato. Neste caso, a fração amostral em cada estrato é constante e igual fração amostral da amostra inteira. Portanto, os estratos maiores ficam com amostras maiores;
- *igual*: quando as amostras têm o mesmo tamanho em todos os estratos;
- *de Neyman*: quando a distribuição da amostra total considera a variabilidade entre unidades dentro dos estratos, levando a tamanhos de amostras maiores em estratos com maior variabilidade entre unidades; e
- *ótima*: quando a distribuição da amostra total depende do tamanho, da variabilidade e do custo de coleta de cada estrato.

As situações que indicam a necessidade de alocação desproporcional (igual, de Neyman ou ótima) são quando as estimativas são requeridas por estrato levando a estratos menores precisarem ser amostrados com taxas maiores ou quando os estratos diferem em termos de variabilidade entre unidades e/ou custo por unidade.

12.4.1 Alocação Proporcional

Uma amostra 'representativa' deveria 'imitar' ou se parecer bastante com a população de onde foi extraída.

As unidades populacionais são distribuídas nos estratos segundo as proporções:

$W_h = N_h/N,\ \ h = 1, \ldots, H$, com $\sum_{h=1}^{H} W_h = 1.$

As proporções amostrais nos estratos são definidas como: $\lambda_h = n_h/n,\ \ h = 1, \ldots, H$, com $\sum_{h=1}^{H} \lambda_h = 1.$

Então o critério acima sugeriria tentar fazer $\lambda_h = W_h\ \ \forall h = 1, 2, \ldots, H.$

Isto implica fazer $\frac{n_h}{n} = \frac{N_h}{N}$ ou $n_h = n\frac{N_h}{N} = nW_h,\ \ \forall h = 1, 2, \ldots, H.$

Esta distribuição da amostra nos estratos é chamada *Alocação Proporcional*.

Um plano AES com $\frac{n_h}{n} = \frac{N_h}{N}$ é chamado de amostragem estratificada simples proporcional ou equiponderada.

Sob alocação proporcional, a *média amostral simples* é o estimador não viciado da média populacional, pois se $n_h = nW_h$, então:

$$\bar{y}_{AES} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} W_h \frac{1}{n_h} \sum_{i \in a_h} y_{hi} = \frac{1}{n} \sum_{h=1}^{H} \sum_{i \in a_h} y_{hi} = \bar{y}$$

A variância de $\bar{y}_{AES}$ sob alocação proporcional simplifica para:

$$V_{AES/Prop}\left(\bar{y}_{AES}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} W_h S_h^2$$

A expressão $\sum_{h=1}^{H} W_h S_h^2 = S_D^2$ mede a *variância dentro* dos estratos, dada por uma média ponderada dos S_h^2.

Então:

$$V_{AES/Prop}\left(\bar{y}_{AES}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S_D^2$$

Esta expressão tem a mesma forma que a correspondente ao caso de AAS, com S_y^2 substituído por S_D^2 .

Como a variância dentro é geralmente menor que a variância total $\left(S_D^2 < S_y^2\right)$, fica evidenciado que estratificação com alocação proporcional geralmente reduz a variância do estimador quando comparada com AAS de igual tamanho.

12.4.2 Alocação Ótima

A maioria das pesquisas sofre restrições orçamentárias.

Se o custo total da pesquisa é fixado em C unidades monetárias, então é necessário especificar uma *função custo* que descreva como esse custo varia para diferentes tamanhos amostrais e alternativas de alocação.

Suponha uma função custo linear dada por:

$$C = c_0 + \sum_{h=1}^{H} n_h c_h$$

O problema é minimizar a $V_{AES}\left(\bar{y}_{AES}\right)$ sujeito à restrição de não ultrapassar o orçamento previsto (custo total C).

Solução:

$$V_{AES}\left(\bar{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) = \sum_{h=1}^{H} W_h^2 S_h^2 / n_h - V_0$$

onde $V_0 = \sum_{h=1}^{H} W_h^2 S_h^2 / N_h$.

Como V_0 não depende de n_h, minimizando $V_{AES}\left(\bar{y}_{AES}\right)$ sujeito a $C = c_0 + \sum_{h=1}^{H} n_h c_h$ resulta em:

$$n_h \propto \left(W_h^2 S_h^2 / c_h\right)^{1/2} = W_h S_h / \sqrt{c}_h$$

Isto é:

$$n_h = n \times \frac{W_h S_h / \sqrt{c}_h}{\sum_{k=1}^{H} W_k S_k / \sqrt{c}_h}, \ \forall\, h = 1, \ldots, H$$

Esta alocação é chamada **Alocação Ótima**.

Cabe registrar que sob a *alocação ótima*, selecione uma amostra maior num estrato h sempre que:

a. O estrato tiver mais unidades (N_h grande);

b. A variabilidade no estrato for maior (S_h grande); e

c. O custo de amostragem no estrato for menor (c_h pequeno).

Se $S_h = S*$ e $c_h = c* \ \forall\, h = 1, 2, \ldots, H$, ambos constantes, então $n_h \propto N_h$, isto é, a alocação ótima coincide com a alocação proporcional.

Se $c_h = c* \ \forall\, h = 1, 2, \ldots, H$, isto é, os custos são constantes ao longo dos estratos, então $n_h \propto N_h S_h$, gerando a chamada **Alocação (Ótima) de Neyman**.

Esta alocação é muito usada em pesquisas de estabelecimentos quando os desvios padrões S_h crescem com o tamanho das unidades.

Para um custo fixado C, assumindo função linear de custos $C = c_0 + \sum_{h=1}^{H} n_h c_h$, o tamanho total da amostra n é:

$$n = \left(C - c_0\right) \times \frac{\sum_{h=1}^{H} N_h S_h / \sqrt{c}_h}{\sum_{h=1}^{H} N_h S_h \sqrt{c}_h}$$

Se a **Alocação de Neyman** é usada, então o valor da variância correspondente ao mínimo é dado por:

$$V_{AES/Ney}\left(\bar{y}_{AES}\right) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h S_h\right)^2 - \frac{1}{N}\left(\sum_{h=1}^{H} W_h S_h^2\right)$$

O segundo termo à direita corresponde à *correção de população finita*.

As soluções acima são 'aproximadas', pois ignoram restrições do tipo $n_h \leq N_h,\ n_h \leq 1,\ n_h$ inteiro $\forall\, h$. Brito (2005) oferece uma solução 'exata'.

12.4.3 Comparação de alternativas de alocação da amostra

Usando a partição da soma de quadrados total em parcelas devidas à variação dentro e entre estratos, e ignorando termos de ordem $1/N_h$, então sob *alocação de Neyman*, isto é, com $n_h \propto N_h S_h$ prova-se (Cochran (1977), p. 99) que:

$$V_{AES/Ney}\left(\bar{y}_{AES}\right) \leq V_{AES/Prop}\left(\bar{y}_{AES}\right) \leq V_{AAS}\left(\bar{y}\right)$$

ou seja, AES com alocação de Neyman é mais eficiente que AES com alocação proporcional, ambas superando AAS como plano amostral.

12.4.4 Alguns problemas com alocação ótima

Seguem algumas situações enfrentadas com a alocação ótima:

1. Se os valores de $S_h, h = 1, \ldots, H$, são desconhecidos.

As solucões possíveis são: usar informação de variável auxiliar x; usar S_{hx} para estimar S_{hy}; predizer y_{hi} usando x_{hi}, e então estimar S_{hy}; usar a soma ou a amplitude de x_{hi} no estrato h como proxy para S_{hy}; selecionar pequena amostra piloto (preliminar) e usar dados desta amostra para estimar S_{yh}.

2. Pode haver muitas variáveis de pesquisa.

Cada variável usualmente levaria a uma alocação ótima diferente. Qualquer método deve buscar um compromisso entre as diversas alternativas: tome a média das alocações alternativas; escolha uma ou duas variáveis principais; use alocação proporcional; construa um 'índice' das variáveis de pesquisa e use este índice para definir a alocação.

Brito et al. (2015) oferece abordagem promissora, para alocação ótima 'multivariada'. O pacote 'stratbr' do R está disponível para implementar; usa formulações de programação inteira binária; soluções disponíveis para minimizar o custo ou tamanho amostral para limites

de precisão especificados, ou para minimizar uma função de variância dado um orçamento ou tamanho amostral especificado; e permite fixar tamanho mínimo de amostra por estrato.

3. Se $n_h > N_h$ em alguns casos.

Ponha $n_h = N_h$ (estrato certo) para os estratos que $n_h > N_h$ e refaça a alocação ótima nos demais estratos.

4. Se $n_h = 1$ em alguns casos.

Se a estimação de variâncias for importante, então force $n_h \geq 2$.

Na prática, costuma-se fazer $n_h \geq 5$ devido à não resposta.

Caso contrário, use métodos aproximados somente para estimação de variâncias, tais como agregação de estratos ou similar (ver Cochran (1977), seção 5A.12).

5. Ganhos de eficiência podem ser modestos, particularmente para estimação de proporções.

Cochran (1977), p. 99, mostra que $V_{AES/Ney}\left(\bar{y}_{AES}\right) \leq V_{AES/Prop}\left(\bar{y}_{AES}\right) \leq V_{AAS}\left(\bar{y}\right)$.

Os ganhos possíveis de precisão dependem da relação entre a(s) variável(is) de estratificação e as variáveis de pesquisa.

Em geral, ganhos são pequenos para amostras de pessoas e variáveis ligadas a atitudes, opiniões, comportamentos, etc.

Para pesquisas amostrais de estabelecimentos ou instituições, os ganhos podem ser muito grandes.

12.5 Definição dos limites dos estratos

Se uma variável auxiliar x estiver disponível, seus valores podem ser usados para formar estratos.

Como devemos formar os estratos? Quais os limites que devemos usar para delimitar os estratos?

Conforme já dito anteriormente, a determinação dos limites dos estratos influencia na eficiência na amostragem estratificada.

Primeiro, escolha H, o número total de estratos.

Quanto maior for a correlação entre a variável de pesquisa y e a variável auxiliar x maior deve ser o número de estratos.

Evidências empíricas sugerem, entretanto, que $5 \leq H \leq 10$. Mais detalhes sobre esta escolha podem ser estudadas em diversas referências apresentadas na Figura 12.1.

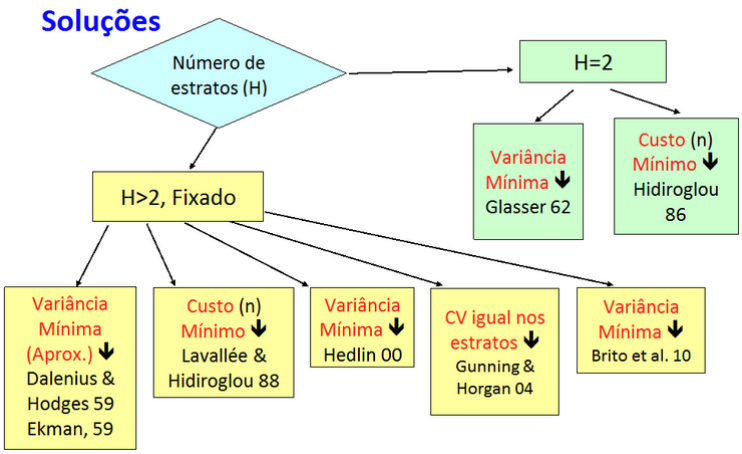

Figura 12.1: Soluções sobre escolha de número de estratos

12.6 Número de estratos na amostragem estratificada simples

Para estimação por domínios, recomenda-se utilizar tantos estratos quantos sejam os domínios de interesse.

Para estimação de total ou média global, Cochran (1977), seção 5A.8, recomenda usar até 6 (seis) estratos, se a variável de estratificação for bem correlacionada com as variáveis de interesse,conforme justificativa dada a seguir.

Sob as hipóteses: N grande, $f = n/N$ pequena; o modelo: $y_i = a + bx_i + \varepsilon_i$ para $i \in U$; estratificação "ótima" em x; alocação igual nos estratos ($n_h = n/H$), mostra-se que:

$EPA\left(\bar{y}_{AES}\right) = V_{AES}\left(\bar{y}_{AES}\right) / V_{AAS}\left(\bar{y}\right) = \rho^2/H^2 + (1 - \rho\ 2)$, onde ρ é correlação entre x e y.

A Tabela 12.4 apresenta valores de $EPA\left(\bar{y}_{AES}\right)$ para valores variados de H e de ρ.

Tabela 12.4: Valores de EPA para valores variados de H e de correlação

Valores de ρ	$H = 2$	$H = 3$	$H = 4$	$H = 5$	$H = 6$	$H = \infty$
$\rho = 0,85$	0,458	0,358	0,323	0,306	0,298	0,277
$\rho = 0,95$	0,323	0,198	0,154	0,134	0,123	0,098

A análise dos resultados acima evidencia que ganhos adicionais de eficiência com mais de seis estratos é modesto.

Na Figura 12.2 temos o gráfico dos ganhos de precisão versus número de estratos para $\rho = 0,85$ e $\rho = 0,95$.

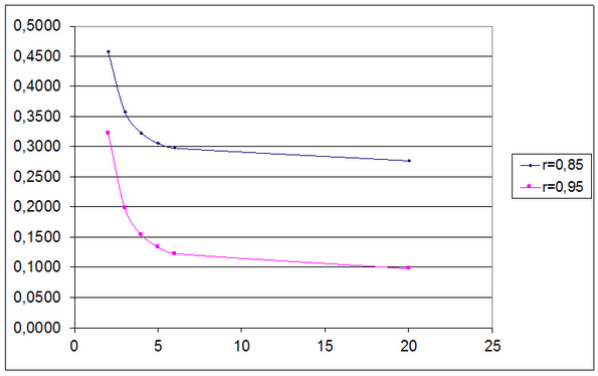

Figura 12.2: Ganhos de precisão versus número de estratos

12.7 Exercícios

Exercício 12.1 Mostre que a média amostral global $\bar{y} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i \in a_h} y_{hi}$ 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 $\frac{n_h}{n} = \frac{N_h}{N}$, $\forall h = 1, \ldots, H$ ou seja, a menos que se adote amostragem estratificada simples proporcional ou equiponderada.

Exercício 12.2 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 proporcional ou equiponderada?

Referências não incluídas no book.bib

Baillargeon, S. & Rivest, L. P. (2011). A General Algorithm for Univariate Stratification. Proceedings of the International Statistical Institute, Dublin.

Brito, J. A. M.; Maculan, N.; Lila, M. F. e Montenegro, F. T. (2010). An exact algorithm for the stratification problem with proportional allocation. Optimization Letters, v. 4, pp. 185 – 195.

Dalenius T. & Hodges Jr., Joseph L. (1959). Minimum Variance Stratification. Journal of the American Statistical Association, Vol. 54, No. 285, pp. 88-101.

Ekman, G. (1959). An Approximation Useful in Univariate Stratification. The Annals of Mathematical Statistics v. 30, p. 219–229.

Glasser, G.J. (1962) On the complete coverage of large units in a statistical study. International Statistical Review, 30, 28-32.

Gunning, P. & Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries. Survey Methodology, 30, No. 2, 159-166.

Hedlin, D. (2000). A procedure for stratification based on an extended Ekman rule. Journal of Official Statistics, 16, 15-29.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design. The American Statistician, 40, n. 1, 27-31.

Lavallée, P. & Hidiroglou, M. A. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.

Rivest, L. P. (2002). A generalization of the Lavallée-Hidiroglou algorithm for stratification in business surveys. Survey Methodology 28, 191-198.