SECRETARIA DE PLANEJAMENTO DA PRESIDENCIA DA REPÚBLICA
IBGE — DIRETORIA DE FORMAÇÃO E APERFEIÇOAMENTO DE PESSOAL — DF
SUPERINTENDENCIA DE ENSINO — SUDEN
ESCOLA NACIONAL DE CIENCIAS ESTATISTICAS — ENCE

PROFESSOR WALTER AUGUSTO DO NASCIMENTO

AMOSTRAGEM DE CONGLOMERADOS

1981

AMOSTRAGEM DE CONGLOMERADOS

A exposição que se segue, sobre Amostragem de Conglomerados, constitue um destaque do curso de Tecnologia da Amostragem do 6.º período da Escola Nacional de Ciências Estatísticas – ENCE. Desse modo, um melhor entendimento sobre o assunto pode ser obtido com a leitura das apostilas da parte restante do curso, principalmente no que se refere a Amostragem Aleatória Simples, Amostragem Estratificada e Estimador de Razão.

Os exemplos que ilustram as partes teóricas têm, apenas, objetivo didático, para mostrar a escolha e o manejo das fórmulas adequadas a cada situação.

CAPÍTULO 1

CAPÍTULO 2

PÁG.

CAPÍTULO 3

CAPÍTULO 4

CAPÍTULO 8

AMOSTRAGEM DE CONGLOMERADOS EM 3 ESTÁGIOS

CAPÍTULO 1 — Amostragem de conglomerados em 1 — estágio Acl — Tamanho desigual. Probabilidade igual de seleção.

1 — DEFINIÇÕES PRELIMINARES

1.1 — *População*

Chama-se População um conjunto finito de N elementos. Será representada por:

$$\pi_N = \{U_1, U_2, \ldots, U_N\}$$

Cada $U_i \in \pi_N$ é uma unidade da população.

1.1.1 — Exemplos

a) População de Domicílios de certa localidade.

b) População de Fazendas produtoras de café de certa região.

c) Alunos da rede escolar estadual.

1.2 — *População matriz*

Considere-se uma característica y de π_N. Por exemplo, na População de Domicílios, uma característica é a renda familiar; na População de Fazendas produtoras de café, uma característica é a quantidade produzida; na População de alunos, uma característica é o rendimento medido por um teste.

y associa a cada unidade $U_i \in \pi_N$ um número real Y_i de modo que se tem:

$$\begin{array}{rl} & U_1 \; U_2 \; \ldots \; U_N \\ y: & \downarrow \; \downarrow \qquad \downarrow \\ & Y_1 \; Y_2 \; \ldots \; Y_N \end{array}$$

O conjunto

$$\pi_N(y) = \{Y_1, Y_2, \ldots, Y_N\}$$

é a "População Matriz" gerada pela característica y.

1.2.1 — Observação

Na mesma população pode-se observar várias características y, x, z, ... gerando, cada uma, uma população matriz, $\pi_N(y)$, $\pi_N(x)$, $\pi_N(z)$, ...

1.3 — *Parâmetros da característica* y

Cada característica tem um conjunto de parâmetros cuja estimação constitue o objetivo principal das técnicas de amostragem.

Entre esses parâmetros serão particularmente considerados:

Total de y: $= Y = \sum_{i=1}^{N} Y_i$

Média de y: $\overline{Y} = \dfrac{Y}{N}$

Variância de y: $\sigma^2 = \dfrac{1}{N} \sum_{i=1}^{N} (Y_i - \overline{Y})^2$

Variância relativa de y: $\vartheta^2 = \dfrac{\sigma^2}{\overline{Y}^2}$

Coeficiente de variação de y: $\vartheta = \dfrac{\sigma}{\overline{Y}}$

Ainda é de interesse estimar a proporção das unidades de π_N que pertencem a um certo nível da característica como, por exemplo, a proporção de pessoas com salário abaixo do salário mínimo, a proporção de pessoas do sexo masculino, etc...

Observe-se que, nesse caso, não há uma característica gerando uma população matriz e sim, uma "contagem" das unidades de π_N que pertencem a determinado nível da característica.

1.4 — *Universo*

Considere-se a experiência aleatória de seleção de uma unidade de π_N, com determinada probabilidade de seleção. Seja u a unidade eventualmente selecionada. Pela característica y associa-se a u uma variável aleatória que se vai representar também por y.

y é o universo associado a π_N.

No que se segue, y está sempre relacionada com a seleção de uma amostra, que é uma experiência aleatória; portanto, y é sempre um universo. A utilidade do conceito decorre, também, de ser y uma variável aleatória e, em certas situações, ser suposta com determinada distribuição.

1.5 — *Configuração ou Desenho da amostra*

Com as técnicas de amostragem se pretende, em linhas gerais, obter um subconjunto ou amostra de π_N com o objetivo de estimar os parâmetros de y (ou de várias características). O modo como a amostra é selecionada constitue a configuração ou desenho da amostra.

A obtenção das unidades da amostra pode ser conseguida selecionando unidades de π_N ou grupos de unidades de π_N.

2 — AMOSTRAGEM DE CONGLOMERADOS

Na amostragem de conglomerados, as unidades de π_N são, inicialmente, grupadas em subconjuntos ou conglomerados. As unidades de amostra, em vez de serem unidades de π_N são os conglomerados.

No quadro que se segue, se mostra a definição de conglomerados para algumas populações.

População	Conglomerados
Alunos	Turmas
Domicílios	Quarteirões
Fichas de cadastro	Grupo de 10 fichas
Turmas	Escola

Observe-se que a turma pode ser unidade de população ou um conglomerado. Se a característica é o rendimento do aluno, a turma é um conglomerado; se a característica é o número de alunos por turma, a turma é uma unidade da população.

3 – AMOSTRAGEM DE CONGLOMERADOS EM 1 – ESTÁGIO

Suponham-se formados M conglomerados com as N unidades de π_N. Selecionem-se com igual probabilidade, m desses conglomerados. O conjunto das unidades de π_N que participam dos m conglomerados selecionados constitue a Amostra de Conglomerados em 1 – estágio de π_N. Observando a característica y das unidades de π_N que estão na amostra, obtém-se a Amostra de Conglomerados em 1 – estágio de y.

Esquematicamente, esse desenho de amostra é representado do modo que se segue.

Seja U_{ij} a j-ésima unidade de π_N no i-ésimo conglomerado C_i. Pela característica y associe-se a U_{ij} o número real Y_{ij}.

Tem-se para os M conglomerados:

C_1		C_2			C_M	
U_{11}	Y_{11}	U_{21}	Y_{21}		U_{M1}	Y_{M1}
U_{12}	Y_{12}	U_{21}	Y_{21}	...	U_{M2}	Y_{M2}
⋮		⋮			⋮	
U_{1N_1}	Y_{1N_1}	U_{2N_2}	Y_{2N_2}		U_{MN_M}	Y_{MN_M}

$N_1, N_2, \ldots N_M$ são, respectivamente, os tamanhos dos conglomerados $C_1, C_2, \ldots C_M$, sendo $N_1 + N_2 + \ldots + N_M = N$.

Selecionem-se m conglomerados, com probabilidade igual de seleção. Sejam $C'_1, C'_2, \ldots, C'_m$ esses conglomerados.

C'_1		C'_2		C'_m	
U'_{11}	Y'_{11}	U'_{21}	Y'_{21}	U'_{m1}	Y'_{m1}
U'_{12}	Y'_{12}	U'_{22}	Y'_{22}	U'_{m2}	Y'_{m2}
$\vdots$		$\vdots$		$\vdots$	
$U'_{1N'_1}$	$Y'_{1N'_1}$	$U'_{2N'_2}$	$Y'_{2N'_2}$	$U'_{mN'_m}$	$Y'_{mN'_m}$

Observe-se que C'_i $(i = 1, 2, \ldots m)$ é um acontecimento aleatório que decorre do processo de seleção. Eventualmente, C'_i pode ser C_1 ou C_2 ... ou C_M.

Conseqüentemente, os Y'_{ij} $(i = 1, 2, \ldots m;\ j = 1, 2, \ldots N'_i)$ e os N'_i $(i = 1, 2, \ldots m)$ são variáveis aleatórias.

A amostra de π_N é:

$$\{U'_{11}, \ldots, U'_{1N'_1}; \ldots\ldots; U'_{m1}, \ldots, U'_{mN'_m}\}$$

e a amostra de y é:

$$\{Y'_{11}, \ldots, Y'_{1N'_1}; \ldots\ldots; Y'_{m1}, \ldots, Y'_{mN'_m}\}$$

Observe-se que o tamanho da amostra é:

$$\sum_{i=1}^{m} N'_i$$

sendo uma variável aleatória, cujos valores dependem dos conglomerados selecionados. Em média, assume o valor:

$$\bar{n} = E\left(\sum_{i=1}^{m} N'_i\right) = m\ \frac{\sum_{i=1}^{M} N_i}{M} = \frac{m}{M}\ N = f_1 N \text{ sendo } f_1 = \frac{m}{M}$$

fração de amostragem de 1.º estágio.

Desse modo, a amostragem de conglomerados em 1 — estágio é caracterizada pelos seguintes fatos:

a) Pertencem à amostra *todas* as unidades dos conglomerados selecionados.

b) Só é necessário listar as unidades de π_N nos m conglomerados selecionados para a amostra. É evidente a economia de tempo e de custo quando se compara com a amostragem aleatória simples ou estratificada, nas quais são listadas todas as unidades de π_N.

c) O tamanho da amostra não pode ser exatamente prefixado tendo em vista que depende dos conglomerados selecionados.

d) Cada unidade de π_N tem a mesma probabilidade de participar da amostra, probabilidade esta representada pela fração de amostragem: $f_1 = \frac{m}{M}$.

e) Conforme se verá em próximos Capítulos, são muitas as ocasiões em que a precisão da amostragem de conglomerados é inferior à precisão da amostragem aleatória simples. No entanto, a vantagem que resulta da economia de tempo e de custo pode, em certos casos, compensar a desvantagem da menor precisão.

Observação — No que se segue, é freqüente o uso da abreviação Ac1 para representar a Amostragem de Conglomerados em 1 — estágio e da abreviação Als para representar a Amostragem Aleatória Simples sem Reposição.

A seguir, especificam-se, em detalhes, os parâmetros já anunciados em 1.3, adaptando à situação decorrente da formação de conglomerados.

3.1 — *Parâmetros de y*

Total de y em C_i: $Y_i = \sum_{j=1}^{N_i} Y_{ij}$

Média de y em C_i: $\overline{Y}_i = \frac{Y_i}{N_i}$

Variância de y em C_i: $S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \overline{Y}_i)^2$

Total de y: $Y = \sum_{i=1}^{M} Y_i$

Média de y: por unidade de π_N $\overline{\overline{Y}} = \frac{Y}{N}$

Média por conglomerado: $\overline{Y} = \dfrac{Y}{M}$

Variância de y: $S^2 = \dfrac{1}{N-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} (Y_{ij} - \overline{\overline{Y}})^2$

Variância relativa: $\gamma^2 = \dfrac{S^2}{\overline{\overline{Y}}^2}$

Como resultado da seleção dos conglomerados, seguem-se as seguintes estatísticas.

3.2 — *Estatísticas*

Total de y em C_i': $Y_i' = \sum_{j=1}^{N_i'} Y_{ij}'$

Média de y em C_i': $\overline{Y}_i' = \dfrac{Y_i'}{N_i'}$

Variância de y em C_i': $S_i'^2 = \dfrac{1}{N_i' - 1} \sum_{j=1}^{N_i'} (Y_{ij}' - \overline{Y}_i')^2$

4 — TEOREMA

Um estimador não tendencioso de Y, total de y é:

$$y_{Ac1}^{*} = \frac{M}{m} \sum_{i=1}^{m} Y_i'$$

Prova

$$E(y_{Ac1}^{*}) = \frac{M}{m} \sum_{i=1}^{m} E(Y_i') = \frac{M}{m} \sum_{i=1}^{m} \left(\frac{1}{M} \sum_{i=1}^{M} Y_i \right) =$$

$$= \sum_{i=1}^{M} Y_i = Y$$

4.1 — *Corolário*

Um estimador não tendencioso de $\overline{\overline{Y}}$, média por unidade de π_N é:

$$\overline{\overline{y}}_{Acl} = \frac{y^*_{Acl}}{N} = \frac{M}{mN} \sum_{i=1}^{m} Y'_i = \frac{1}{m\overline{N}} \sum_{i=1}^{m} Y'_i$$

onde $\overline{N} = \frac{N}{M}$, tamanho médio por conglomerado.

Prova

$$E(\overline{\overline{y}}_{Acl}) = \frac{E(y^*_{Acl})}{N} = \frac{Y}{N} = Y$$

4.2 — *Corolário*

Um estimador não tendencioso de $\overline{Y}$, média por conglomerado é:

$$\overline{y}_{Acl} = \frac{y^*_{Acl}}{M} = \frac{1}{m} \sum_{i=1}^{m} Y'_i$$

Prova

Fazer como exercício.

5 — VARIÂNCIA DE y^*_{Acl}

Por definição:

$$V(y^*_{Acl}) = E(y^*_{Acl} - Y)^2$$

ou,

$$V(y^*_{Acl}) = E\left(\frac{M}{m} \sum_{i=1}^{m} Y'_i - Y\right)^2 =$$

$$= \frac{M^2}{m^2} E\left[\sum_{i=1}^{m} Y'_i - m\overline{Y}\right]^2 =$$

$$= \frac{M^2}{m^2} E\left[\sum_{i=1}^{m} (Y'_i - \overline{Y})\right]^2 =$$

$$= \frac{M^2}{m^2} E\left[\sum_{i=1}^{m} (Y'_i - \overline{Y})^2 + \sum_{\substack{i=1 \\ (i \neq j)}}^{m} \sum_{j=1}^{m} (Y'_i - \overline{Y})(Y'_j - \overline{Y})\right] =$$

$$= \frac{M^2}{m^2}\left[\frac{m}{M} \sum_{i=1}^{M} (Y_i - \overline{Y})^2 + \frac{m(m-1)}{M(M-1)} \sum_{\substack{i=1 \\ (i \neq j)}}^{M} \sum_{j=1}^{M} (Y_i - \overline{Y})(Y_j - \overline{Y})\right] =$$

$$= \frac{M}{m}\left[\sum_{i=1}^{M} (Y_i - \overline{Y})^2 - \frac{m-1}{M-1} \sum_{i=1}^{M} (Y_i - \overline{Y})^2\right]$$

Pondo

$$S_e^2 = \frac{1}{M-1} \sum_{i=1}^{M} (Y_i - \overline{Y})^2$$

vem,

$$V(\overline{y}^*_{Ac1}) = \frac{M}{m}\left[(M-1) S_e^2 - (m-1) S_e^2\right]$$

donde.

$$V(\overline{y}^*_{Ac1}) = M^2 \frac{M-m}{M} \cdot \frac{S_e^2}{m}$$

5.1 — *Exercícios*

a) Mostrar que

$$V(\overline{y}_{Ac1}) = \frac{M-m}{M} \frac{S_e^2}{m} \quad \text{com} \quad S_e^2 = \frac{1}{M-1} \sum_{i=1}^{M} (Y_i - \overline{Y})^2$$

b) Mostrar que

$$V(\overline{\overline{y}}_{Ac1}) = \frac{M-m}{M} \frac{\overline{S}_e^2}{m} \quad \text{com} \quad \overline{S}_e^2 = \frac{1}{M-1} \sum_{i=1}^{M} \left(\frac{Y_i}{\overline{N}} - \overline{\overline{Y}}\right)^2$$

c) Supor uma situação em que caiba a formação de conglomerados. Definir uma característica e fazer uma breve descrição do processo de estimação da média por unidade da população.

6 – TEOREMA

Um estimador não tendencioso de $V(y^*_{Ac1})$ é:

$$\hat{V}(y^*_{Ac1}) = M^2 \frac{M-m}{M} \frac{s_e^2}{M}$$

onde

$$s_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y'_i - \bar{y}_{Ac1})^2$$

Prova

Pode-se escrever:

$$s_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y'_i - \bar{y}_{Ac1})^2 =$$

$$= \frac{1}{m-1} \sum_{i=1}^{m} [(Y'_i - \bar{Y}) - (\bar{y}_{Ac1} - \bar{Y})]^2 =$$

$$= \frac{1}{m-1} \left[\sum_{i=1}^{m} (Y'_i - \bar{Y})^2 - m (\bar{y}_{Ac1} - \bar{Y})^2 \right]$$

donde,

$$E(s_e^2) = \frac{1}{m-1} \left[\frac{m}{M} \sum_{i=1}^{M} (Y_i - \bar{Y})^2 - mV(\bar{y}_{Ac1}) \right] =$$

$$= \frac{1}{m-1} \left[\frac{m(M-1)}{M} S_e^2 - m \frac{M-m}{M} \frac{S_e^2}{m} \right] = S_e^2$$

Portanto,

$$E[\hat{V}(y^*_{Ac1})] = \frac{M-m}{M} \frac{E(s_e^2)}{m} = \frac{M-m}{M} \frac{S_e^2}{m} = V(y^*_{Ac1})$$

6.1 – *Exercício*

a) Mostrar que

$$\hat{V}(\bar{y}_{Ac1}) = \frac{M-m}{M} \frac{s_e^2}{m} \text{ com } s_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y'_i - \bar{y}_{Ac1})^2$$

b) Mostrar que

$$\hat{V}(\bar{\bar{y}}_{Ac1}) = \frac{M-m}{M} \frac{\bar{s}_e^2}{m} \text{ com } \bar{s}_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(\frac{Y'_i}{\bar{N}} - \bar{\bar{y}}_{Ac1} \right)^2$$

7 — EXEMPLO

Trata-se de avaliar o rendimento dos alunos da 1.ª série — 1.º grau, na rede de ensino público de certa localidade.

A partir da relação das 3 500 turmas existentes, foram preparados conglomerados, juntando turmas de diferentes Escolas, com o objetivo de grupar alunos o mais possível diferentes no que se refere ao rendimento (a necessidade dos conglomerados serem heterogêneos pode ser vista no Cap. 2).

Os conglomerados foram formados com 5 turmas e, aproximadamente, 150 alunos, supondo uma base de 30 alunos por turma.

Há possibilidade de tempo e de recursos financeiros para observar uma amostra de 1 500 alunos.

Considerando que $\overline{n} = m\overline{N}$ tem-se:

$$1\,500 = m\ 150$$

donde $m = 10$ conglomerados.

Conglomerados da amostra C_i'	N.º de alunos N_i'	Soma dos escores Y_i'
1	162	1004,4
2	170	952,0
3	145	1015,0
4	151	830,5
5	160	960,0
6	162	793,8
7	145	855,5
8	148	947,2
9	171	1214,1
10	178	1032,4
Soma	1592	9604,9

Tem-se:

$M = 700$

$m = 10$

$\overline{N} = 150$

Estimativa do escore médio por aluno:

$$\bar{\bar{y}}_{Acl} = \frac{1}{m\overline{N}} \sum_{i=1}^{m} Y'_i = \frac{9604{,}9}{10\,(150)} = 6{,}4$$

Estimativa da variância de $\bar{\bar{y}}_{Acl}$*:*

$$s_e^2 = \frac{1}{m-1}\left[\sum_{i=1}^{m} Y'^2_i - \frac{\left(\sum_{i=1}^{m} Y'_i\right)^2}{m}\right] = 14482{,}59$$

$$\hat{V}(\bar{\bar{y}}_{Acl}) = \frac{1}{\overline{N}^2}\,\frac{M-m}{M}\cdot\frac{s_e^2}{m} = 0{,}0634$$

Estimativa do coeficiente de variação de $\bar{\bar{y}}_{Acl}$*:*

$$\hat{\gamma}(\bar{\bar{y}}_{Acl}) = \frac{\sqrt{\hat{v}(\bar{\bar{y}}_{Acl})}}{\bar{\bar{y}}_{Acl}} = 0{,}039$$

8 – ESTIMAÇÃO DE PROPORÇÃO

Considere-se a população dividida em duas classes: A e $\bar{A}$ (não A), de acordo com algum atributo associado às unidades de π_N.

Por exemplo, se a população é de domicílios,

– A pode ser a classe dos domicílios próprios. $\bar{A}$ a classe dos não próprios.

– A pode ser a classe dos domicílios em que há TV a cores. $\bar{A}$ a classe dos domicílios sem TV a cores.

– A pode ser a classe dos domicílios em que há pelo menos um carro. $\bar{A}$ a classe dos domicílios sem carro.

Em conseqüência, se a população é grupada em M conglomerados, cada conglomerado é dividido também nas classes A e $\bar{A}$. Seja

A_i e $\bar{A}_i$ o número de unidades de π_N em A e $\bar{A}$, respectivamente, no conglomerado i $(i = 1, 2, \ldots M)$.

A_i pode assumir os valores $0, 1, 2, \ldots N_i$ e se tem:

$$A_i + \bar{A}_i = N_i$$

Donde $Y_i = A_i$ $(i = 1, 2, \ldots M)$ obtem-se:

$$\overline{Y}_i = \frac{A_i}{N_i} = P_i \text{ proporção da classe } A \text{ em } C_i$$

$$\overline{\overline{Y}} = \frac{\sum_{i=1}^{M} Y_i}{N} = \frac{\sum_{i=1}^{M} A_i}{N} = P \text{ proporção da classe } A \text{ em } \pi_N.$$

Com essas substituições, torna-se fácil achar o estimador de P usando as expressões dos estimadores estudados nas secções anteriores.

8.1 — *Estimador não tendencioso de P*

Das expressões de $\overline{\overline{y}}_{Ac1}$, $V(\overline{\overline{y}}_{Ac1})$ $\hat{V}(\overline{\overline{y}}_{Ac1})$ do estimador não tendencioso de $\overline{\overline{Y}}$ obtem-se:

$$p_{Ac1} = \frac{1}{m\overline{\overline{N}}} \sum_{i=1}^{m} A_i'$$

$$V(p_{Ac1}) = \frac{M - m}{M} \cdot \frac{\overline{S}_e^2}{m} \text{ com } \overline{S}_e^2 = \frac{1}{M - 1} \sum_{i=1}^{M} \left(\frac{A_i}{\overline{\overline{N}}} - P \right)^2$$

$$\hat{V}(p_{Ac1}) = \frac{M - m}{M} \cdot \frac{\bar{s}_e^2}{m} \text{ com } \bar{s}_e^2 = \frac{1}{m - 1} \sum_{i=1}^{m} \left(\frac{A_i'}{\overline{\overline{N}}} - p_{Ac1} \right)^2$$

8.2 — *Exemplo*

No Exemplo 7 observou-se, também, o número de alunos que fumam, obtendo-se:

Conglomerados da amostra	N.º de alunos (N_i')	N.º de alunos que fumam (A_i')
1	162	50
2	170	63
3	145	47
4	151	48
5	160	68
6	162	59
7	145	36
8	148	45
9	171	71
10	178	75
Soma	1592	562

Estimativas da proporção de alunos que fumam

$$p_{Ac1} = \frac{1}{m\overline{N}} \sum_{i=1}^{m} A_i' = \frac{562}{10\,(150)} = 0{,}375 \text{ ou } 37{,}5\%$$

Estimativa da variância de p_{Ac1}

$$s_e^2 = \frac{\sum_{i=1}^{m} A_i'^2 - \frac{\left(\sum_{i=1}^{m} A_i'\right)^2}{m}}{m-1} = \frac{33074 - \frac{(562)^2}{10}}{9} = 165{,}51$$

donde,

$$\hat{V}(p_{Ac1}) = \frac{1}{\overline{N}^2} \frac{M-m}{M} \frac{s_e^2}{m} = \frac{1}{(150)^2} \frac{700-10}{700} \frac{165{,}51}{10} = 0{,}000725$$

$$\sqrt{\hat{V}(p_{Ac1})} = 0{,}0269 \text{ ou } 2{,}69\%$$

9 – DIMENSIONAMENTO DA AMOSTRA

O erro de amostra do estimador $\overline{y}_{Ac1}$ é definido por:

$$|\overline{y}_{Ac1} - \overline{Y}|$$

Em princípio, esse erro não é conhecido, posto que se supõe o desconhecimento de $\overline{Y}$. No entanto, tem-se, com probabilidade θ, que esse erro seja inferior a

$$d = z_\theta \sqrt{V(\overline{y}_{Ac1})}$$

onde z_θ é o valor da normal $(0; 1)$ correspondendo a uma área central θ.

O erro relativo da amostra é definido por $\dfrac{|\overline{y}_{Ac1} - \overline{Y}|}{\overline{Y}}$ sendo inferior a $d_r = z_\theta \ \gamma(\overline{y}_{Ac1})$ com probabilidade θ. d_r é a precisão relativa de $\overline{y}_{Ac1}$ e $\gamma(\overline{y}_{Ac1})$ o coeficiente de variação de $\overline{y}_{Ac1}$.

Fixado d_r, o tamanho m da amostra de conglomerados que torna

$$\frac{|\overline{y}_{Ac1} - \overline{Y}|}{\overline{Y}} < d_r$$

com probabilidade θ é obtido da igualdade:

$$d_r = z_\theta \ \gamma(\overline{y}_{Ac1})$$

ou,

$$d_r = z_\theta \sqrt{\frac{M - m}{M} \cdot \frac{\gamma_e^2}{m}}$$

onde γ_e é o coeficiente de variação entre os totais dos conglomerados e dado por $\dfrac{S_e}{\overline{Y}}$.

Deste modo, obtem-se para m:

$$m = \frac{M\ z_\theta\ \gamma_e^2}{M\ d_r^2 + 4\ \gamma_e^2}$$

É usual fixar $\theta = 0{,}95$ (95%) donde $z_{0,95} = 1{,}96 \doteq 2$

donde,

$$m = \frac{M\ 4\ \gamma_e^2}{M\ d_r^2 + 4\ \gamma_e^2}$$

O coeficiente de variação dos totais, γ_e, pode ser obtido de pesquisa anterior ou de dados do Censo. Tem-se, ainda, a alternativa de substituir γ_e por um estimador $\hat{\gamma}_e = \frac{s_e}{\bar{y}_{Ac1}}$ obtido de uma amostra preliminar.

9.1 — *Exemplo*

Considere-se o Exemplo 7.

A precisão relativa obtida com a amostra de 10 conglomerados foi:

$$d_r = 2\ \hat{\gamma}\,(\bar{y}_{Ac1}) = 2(0{,}039) = 0{,}078 \text{ ou } 7{,}8\%$$

Para achar o tamanho da amostra com precisão relativa de 5%, calcula-se:

$$\hat{\gamma}_e^2 = \frac{\bar{y}_{Ac1}}{s_e^2} = \frac{14482{,}59}{(960{,}49)^2} = 0{,}0157$$

$$m \doteq \frac{4\ M\ \hat{\gamma}_e^2}{M\,(0{,}05)^2 + 4\ \hat{\gamma}_e^2} = \frac{4(700)\ (0{,}0157)}{700\,(0{,}0025) + 4\,(0{,}0157)} =$$

$= 24$ conglomerados.

9.2 — *Exercício*

Achar a expressão de m para estimar uma proporção.

9.3 — *Observação*

A validade da informação sobre d_r e sobre m, conforme Exemplo 9.1 depende de haver boa aproximação entre s_e^2 e S_e^2.

10 — *EFICIÊNCIA DA Acl EM RELAÇÃO A Als*

Representando por *Ef* esta eficiência, define-se:

$$Ef = \frac{V(\bar{y})}{V(\bar{\bar{y}}_{Acl})}$$

onde $V(y)$ é a variância do estimador da média por unidade de π_N na *Als* e $V(\bar{\bar{y}}_{Acl})$ a variância da mesma média na *Acl*. Observe-se que a eficiência cresce a medida que $V(\bar{\bar{y}}_{Acl}) < V(\bar{y})$.

Sabe-se que $V(\bar{\bar{y}}_{Acl}) = \frac{M-m}{M} \cdot \frac{S_e^2}{m\bar{N}^2}$ e que $V(\bar{y}) = \frac{N-n}{N} \cdot \frac{S^2}{n}$ onde S^2 é a variância de y em π_N, sem formação de conglomerados. Nessa última expressão, $N = M\bar{N}$ mas $n = \sum_{i=1}^{m} N_i'$ é uma variável aleatória cujo valor depende dos conglomerados que forem selecionados. A expectância de n é:

$$E(n) = E\left(\sum_{i=1}^{m} N_i'\right) = \frac{m}{M} \sum_{i=1}^{M} N_i = m\bar{N}$$

Substituindo N por $M\bar{N}$ e n por $m\bar{N}$ em $V(\bar{y})$ obtém-se:

$$V(\bar{y}) \doteq \frac{M\bar{N} - m\bar{N}}{M\bar{N}} \frac{S^2}{m\bar{N}} = \frac{M-m}{M} \cdot \frac{S^2}{m\bar{N}}$$

Fazendo as substituições em *Ef* vem:

$$Ef = \frac{\frac{M-m}{M} \cdot \frac{S^2}{m\bar{N}}}{\frac{M-m}{M} \cdot \frac{S_e^2}{m\bar{N}^2}} = \frac{\bar{N}S^2}{S_e^2}$$

11 – TAMANHO DO CONGLOMERADO

O tamanho do conglomerado não é arbitrário. Depende do número de observações independentes. Por exemplo, suponha-se que o objetivo de determinada pesquisa seja estimar a proporção de moradores da raça negra em certa localidade em que há segregação racial. Se o conglomerado é o quarteirão e se o entrevistador vai a um domicílio e é atendido por um morador da raça negra, é de esperar que haja no quarteirão muitos outros da raça negra. Desse modo, há poucas observações independentes e o quarteirão não é um bom conglomerado.

O modo de se avaliar a maior ou menor independência nas observações dentro dos conglomerados é com o emprego do coeficiente de correlação intraclasse.

12 – COEFICIENTE DE CORRELAÇÃO INTRACLASSE

Foi dito na Seção 11, que o valor de $\overline{N}$ não é arbitrário, dependendo do número de observações independentes e que o modo de medir esse número é com o emprego do coeficiente de correlação intraclasse.

Representando por δ o coeficiente de correlação intraclasse, define-se:

$$\delta = \frac{E(Y'_{ij} - \overline{\overline{Y}})\ (Y'_{ik} - \overline{\overline{Y}})}{E(Y'_{ij} - \overline{\overline{Y}})^2} \quad (i \neq j) \tag{12.I}$$

A média se extende aos valores de i, de j e de k (sendo $j \neq k$). Desse modo, δ mede a correlação entre pares de valores de y, em cada conglomerado.

No que se segue, admite-se $N_i \doteq N$ $(i=1,2, \ldots M)$ permitindo uma maior simplificação das expressões.

O numerador pode ser escrito do seguinte modo:

$$E(Y'_{ij} - \overline{Y})\,(Y'_{ik} - \overline{\overline{Y}}) = E(Y'_{ij} - \overline{Y}'_i + \overline{Y}'_i - \overline{\overline{Y}})\,(Y'_{ik} - \overline{Y}'_i + \overline{Y}'_i - \overline{\overline{Y}}) =$$
$$= E(Y'_{ij} - \overline{Y}'_i)(Y'_{ik} - \overline{Y}'_i) + E(Y'_{ij} - \overline{Y}'_i)(\overline{Y}'_i - \overline{\overline{Y}}) + E(\overline{Y}'_i - \overline{\overline{Y}})(Y'_{ik} - \overline{Y}'_i) +$$
$$+ E(\overline{Y}'_i - \overline{\overline{Y}})^2$$

Para a primeira parcela do segundo membro, tem-se:

$$E(Y'_{ij}-\overline{Y}'_i)(Y'_{ik}-\overline{Y}'_i)=\underset{C'_i}{E}\left\{E[(Y'_{ij}-\overline{Y}'_i)(Y'_{ik}-\overline{Y}'_i)\mid C'_i \text{ fixado}]\right\}=$$

$$=\underset{C'_i}{E}\left[\frac{1}{\overline{N}(\overline{N}-1)}\sum_{\substack{j=1\\(j\neq k)}}^{\overline{N}}\sum_{k=1}^{\overline{N}}(Y'_{ij}-\overline{Y}'_i)(Y'_{ik}-\overline{Y}'_i)\right]=$$

$$=\underset{C'_i}{E}\left[-\frac{1}{\overline{N}(\overline{N}-1)}\sum_{j=1}^{\overline{N}}(Y'_{ij}-\overline{Y}'_i)^2\right]=$$

$$=\underset{C'_i}{E}\left[-\frac{S'^2_i}{\overline{N}}\right]=-\frac{1}{M\overline{N}}\sum_{i=1}^{M}S_i^2=-\frac{S_d^2}{\overline{N}}$$

Para a segunda parcela do segundo membro, tem-se:

$$E(Y'_{ij}-\overline{Y}'_i)(\overline{Y}'_i-\overline{\overline{Y}})=\underset{C'_i}{E}\left\{E[(Y'_{ij}-\overline{Y}'_i)(\overline{Y}'_i-\overline{\overline{Y}})\mid C'_i \text{ fixado}]\right\}=$$

$$=\underset{C'_i}{E}\left[(\overline{Y}'_i-\overline{\overline{Y}})\frac{1}{N}\sum_{j=1}^{\overline{N}}(Y'_{ij}-\overline{Y}'_i)\right]=0$$

$$\text{posto que}\quad\sum_{j=1}^{\overline{N}}(Y'_{ij}-\overline{Y}'_i)=0$$

Para a terceira parcela, obtém-se, de modo semelhante, o valor 0.

Para a quarta parcela, tem-se:

$$E(\overline{Y}'_i-\overline{\overline{Y}})^2=\frac{1}{M}\sum_{i=1}^{M}(\overline{Y}_i-\overline{\overline{Y}})^2=\frac{M-1}{M}\,\overline{S}_e^2$$

Desse modo, o numerador de δ é:

$$E[(Y'_{ij}-\overline{\overline{Y}})(Y'_{ik}-\overline{\overline{Y}})]=-\frac{S_d^2}{\overline{N}}+\frac{M-1}{M}\,\overline{S}_e^2 \qquad (12.\text{II})$$

Para o denominador de δ tem-se:

$$E(Y'_{ij} - \overline{\overline{Y}})^2 = \underset{C'_i}{E}\left\{E[(Y'_{ij} - \overline{\overline{Y}})^2 \mid C'_i \text{ fixado}]\right\} =$$

$$= \underset{C'_i}{E}\left[\frac{1}{\overline{N}} \sum_{j=1}^{\overline{N}} (Y'_{ij} - \overline{\overline{Y}})^2\right] =$$

$$= \frac{1}{M\overline{N}} \sum_{i=1}^{M} \sum_{j=1}^{\overline{N}} (Y_{ij} - \overline{\overline{Y}})^2 = \frac{M\overline{N} - 1}{M\overline{N}} S^2 \quad (12.III)$$

Substituindo (12.II) e (12.III) em (12.I), obtém-se:

$$\delta = \frac{\dfrac{M-1}{M} \overline{S}_e^2 - \dfrac{1}{\overline{N}} S_d^2}{\dfrac{M\overline{N} - 1}{M\overline{N}} S^2} \quad (12.IV)$$

Em particular, para M grande:

$$\delta \doteq \frac{\overline{S}_e^2 - \dfrac{S_d^2}{\overline{N}}}{S^2} \quad (12.V)$$

12.1 — *Variação de* δ

Para o melhor entendimento da participação de δ na medida da heterogeneidade do conglomerado, considere-se a seguinte igualdade:

$$\sum_{i=1}^{M} \sum_{j=1}^{\overline{N}} (Y_{ij} - \overline{\overline{Y}})^2 = \sum_{i=1}^{M} \sum_{j=1}^{\overline{N}} [(Y_{ij} - \overline{Y}_i) + (Y_i - \overline{\overline{Y}})]^2$$

Obtém-se:

$$\sum_{i=1}^{M} \sum_{j=1}^{\overline{N}} (Y_{ij} - \overline{\overline{Y}})^2 = \sum_{i=1}^{M} \sum_{j=1}^{\overline{N}} (Y_{ij} - \overline{Y}_i)^2 + \sum_{i=1}^{M} \overline{N} (\overline{Y}_i - \overline{\overline{Y}})^2 \quad (12.VI)$$

Sabe-se que:

$$S^2 = \frac{1}{M\overline{N} - 1} \sum_{i=1}^{M} \sum_{j=1}^{\overline{N}} (Y_{ij} - \overline{\overline{Y}})^2$$

(variância total)

$$\overline{S}_e^2 = \frac{1}{M - 1} \sum_{i=1}^{M} (\overline{Y}_i - \overline{\overline{Y}})^2$$

(variância entre conglomerados)

Ponha-se ainda:

$$S_d^2 = \frac{1}{M} \sum_{i=1}^{M} S_i^2$$

(variância dentro dos conglomerados)

Então (12.VI) pode ser escrita:

$$(M\overline{N} - 1)\, S^2 = M\,(\overline{N} - 1) \cdot S_d^2 + (M - 1)\,\overline{N} \cdot \overline{S}_e^2 \qquad (12.VII)$$

Em primeiro lugar, recorde-se que $V(\overline{\bar{y}}_{Ac1}) = \frac{M - m}{M} \frac{\overline{S}_e^2}{m}$ de modo que a variância do estimador decresce a medida que S_e^2 decresce. De acordo com (12.VII), isso implica em S_d^2 crescer, ou seja, os conglomerados se tornaram heterogêneos. Está, assim, justificada uma propriedade básica dos conglomerados: a heterogeneidade.

Quanto a influência da variação de δ na maior ou menor heterogeneidade do conglomerado, tem-se:

Se os conglomerados são homogêneos para uma determinada característica,

$$S_d^2 = 0$$

Nesse caso, de acordo com (12.VII):

$$(M\overline{N} - 1)\, S^2 = (M - 1)\,\overline{N}\ \overline{S}_e^2$$

donde,

$$S^2 = \frac{(M-1)\,\overline{N}}{M\overline{N}-1}\,\overline{S}_e^2 \qquad \text{(12.VIII)}$$

Por outro lado, de acordo com (12.IV) obtém-se:

$$\delta = \frac{\dfrac{M-1}{M}\,\overline{S}_e^2}{\dfrac{M\overline{N}-1}{M\overline{N}}\,S^2} \qquad \text{(12.IX)}$$

Substituindo (12.VIII) em (12.IX) vem:

$$\delta = \frac{\dfrac{M-1}{M}\,\overline{S}_e^2}{\dfrac{(M-1)\,\overline{N}}{M\overline{N}}\,\overline{S}_e^2} = 1$$

Portanto, se os conglomerados são homogêneos, o coeficiente de correlação intraclasse é igual a 1.

A medida que os conglomerados se tornam heterogêneos na característica, $\overline{S}_d^2$ cresce e $\overline{S}_e^2$ decresce.

Quando $\overline{S}_e^2 = 0$ ou seja, quando:

$\overline{Y}_1 = \overline{Y}_2 = \ldots = \overline{Y}_M$, obtém-se de (12.VII):

$$(M\overline{N}-1)\,S^2 = M\,(\overline{N}-1)\,S_d^2$$

donde,

$$S^2 = \frac{M\,(N-1)}{M\overline{N}-1}\,S_d^2$$

Substituindo em (12.VI) obtém-se:

$$\delta_{\cdot} = \frac{-\dfrac{S_d^2}{\overline{N}}}{\dfrac{M\overline{N}-1}{M\overline{N}}\,S_d^2} = -\frac{1}{\overline{N}-1}$$

Em resumo:

a) $$\delta \in \left[-\frac{1}{\overline{N} - 1} ; 1 \right]$$

b) Valores negativos de δ são raros. Na situação usual δ é positivo e os valores mais elevados correspondem a características econômicas. Como exemplo, a renda familiar e o fato do domicílio ser próprio ou não. Como muitas características sociais são influenciadas pelas econômicas, conclue-se que são muitas as características com δ positivo.

c) Se δ é positivo, significa que há homogeneidade dentro de vários conglomerados, resultando serem poucas as observações independentes. Desse modo, um número grande de conglomerados — $\overline{N}$ pequeno — torna-se mais conveniente.

d) Se δ é negativo, significa que há heterogeneidade em vários conglomerados resultando serem muitas as observações independentes. Nesse caso poucos conglomerados — $\overline{N}$ grande — são necessários para se obter a informação desejada.

13 — EFICIÊNCIA DA *Ac1* EM RELAÇÃO A *Als*

A eficiência da *Ac1* em relação a *Als* em vez de ser estudadas pela relação:

$$Ef = \frac{\overline{N} S^2}{S_e^2}$$

pode ser estudada a partir do tamanho $\overline{N}$ e do coeficiente de correlação intraclasse.

De (12.IV) obtém-se:

$$\frac{S_d^2}{\overline{N}} = \frac{M - 1}{M} \overline{S}_e^2 - \frac{M\overline{N} - 1}{M\overline{N}} S^2 \delta$$

ou,

$$MS_d^2 = (M - 1) \overline{N} \overline{S}_e^2 - (M\overline{N} - 1) S^2 \delta$$

Substituindo em (12.VII) vem,

$$(M\overline{N} - 1)\ S^2 = (M - 1)\overline{N}\ (\overline{N} - 1)\ \overline{S}_e^2 - (M\overline{N} - 1)\ (\overline{N} - 1)\ S^2\ \delta + \\ + (M - 1)\ \overline{N}\ \overline{S}_g^2$$

donde,

$$(M\overline{N} - 1)\ S^2 = (M - 1)\ \overline{N}^2\ \overline{S}_e^2 - (M\overline{N} - 1)\ (\overline{N} - 1)\ S^2\ \delta$$

ou,

$$\frac{\overline{S}_e^2}{m} \doteq \frac{S^2}{m\overline{N}}\ [1 + (\overline{N} - 1)\ \delta] \qquad (13.I)$$

para M grande, de modo a se supor

$$M\overline{N} - 1 \doteq M\overline{N} \quad \text{e} \quad M - 1 \doteq M$$

Observe-se que:

$$\frac{\overline{S}_e^2}{m} \doteq V(\overline{\overline{y}}_{Ac1})$$

e que,

$$\frac{S^2}{m\overline{N}} \doteq V(\overline{y})$$

variância do estimador de $\overline{\overline{Y}}$ na *Als*.

donde,

$$V(\overline{\overline{y}}_{Ac1}) \doteq V(\overline{y})\ [1 + (\overline{N} - 1)\ \delta]$$

isto é, a variância do estimador na *AC-1* é igual a variância do estimador na *ALS* multiplicada pelo fator $[1 + (\overline{N} - 1)\ \delta]$

Portanto, esse fator mede a influência da formação de conglomerados na variância do estimador. Daí, a denominação:

efeito da conglomeração. A medida que $\overline{N}$ cresce, esse efeito cresce se $\delta > 0$, posto que, embora δ tenda a decrescer, a taxa de crescimento de $\overline{N}$ é maior que a taxa de decréscimo de δ.

Da igualdade (13.I) conclue-se que $V(\bar{\bar{y}}_{Acl})$ equivale a $V(\bar{y})$ quando:

$$\frac{\bar{S}_e^2}{m[1+(\bar{N}-1)\,\delta]} = \frac{S^2}{m\bar{N}}$$

ou seja, quando são selecionadas $m\ [1+(N-1)\,\delta]$ conglomerados. Nesse caso, o número de unidades de π_N na amostra e $m\ [1+(\bar{N}-1)\,\delta]\ \bar{N} = m\bar{N} + m\bar{N}\,(\bar{N}-1)\,\delta$ ou seja, há um acréscimo de $m\bar{N}\,(\bar{N}-1)\,\delta$ unidades em relação a amostra aleatória simples sem reposição.

14 – ESTIMAÇÃO DO COEFICIENTE DE CORRELAÇÃO INTRACLASSE

Considere-se a expressão:

$$\delta = \frac{\bar{S}_e^2 - \dfrac{S_d^2}{\bar{N}}}{S^2}$$

Sabe-se (Seção 6) que um estimador não tendencioso de $\bar{S}_e^2$ é:

$$\bar{s}_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{Y}'_i - \bar{\bar{y}}_{Acl})^2$$

Um estimador não tendencioso de $S_d^2 = \dfrac{1}{M} \sum_{i=1}^{M} S_i^2$ é:

$$s_d^2 = \frac{1}{m} \sum_{i=1}^{m} S_i'^2$$

Um estimador não tendencioso de S^2 pode ser obtido da expressão (12.VII) substituindo S_d^2 e $\bar{S}_e^2$ pelos respectivos estimadores:

$$s^2 = \frac{M(\bar{N}-1)\,s_d^2 + (M-1)\,\bar{N}\,\bar{s}_e^2}{M\bar{N}-1}$$

Substituindo esses três estimadores na expressão de δ vem:

$$\hat{\delta} = \frac{\overline{s}_e^2 - \frac{s_d^2}{\overline{N}}}{s^2}$$

estimador consistente de δ.

15 – EXEMPLO

a) Tem-se um fichário de 20.000 segurados de uma Companhia de seguros, em um plano A. O objetivo é calcular, por amostragem, a reserva técnica do plano A, com uma amostragem de conglomerados, probabilidade igual de seleção.

As 20.000 fichas estão dispostas em 400 gavetas, com 50 fichas cada.

Considerando as gavetas como conglomerados, tem-se:

$$M=400$$

$$\overline{N}=50$$

Selecionou-se uma amostra de 10 gavetas, correspondendo a 500 fichas. Nas gavetas selecionadas foram calculadas as reservas técnicas de todas as fichas, obtendo-se:

Gavetas da amostra	Reserva total (Y_i)	Variância das reservas (S_i^2)
1	321	25
2	170	17
3	610	30
4	405	32
5	350	35
6	155	20
7	254	40
8	328	18
8	328	18
9	652	25
10	269	35
Soma	3514	277

Estimativa de S_d^2

$$s_d^2 = \frac{1}{m} \sum_{i=1}^{m} S_i'^2 = \frac{277}{10} = 27{,}7$$

Estimativa de $\overline{S}_e^2$

Do quadro, obtém-se: $\sum_{i=1}^{m} Y_i'^2 = 1484156$

Para aproveitar esse valor, pode-se escrever $\overline{s}_e^2$ do seguinte modo:

$$\overline{s}_e^2 = \frac{1}{(m-1)\,\overline{N}^2} \sum_{i=1}^{m} (Y_i' - \overline{y}_{Aç1})^2$$

$$= \frac{1}{(m-1)\,\overline{N}^2} \left[\sum_{i=1}^{m} Y_i'^2 - \frac{\left(\sum_{i=1}^{m} Y_i' \right)^2}{m} \right] =$$

$$= \frac{1}{9\,(50)^2} \left[1484156 - \frac{(3514)^2}{10} \right] = 11{,}082$$

Estimativa de S^2

$$s^2 = \frac{M(\overline{N}-1)\; s_d^2 + (M-1)\,\overline{N}\; \overline{s}_e^2}{M\overline{N} - 1}$$

$$= \frac{400\,(49)\;\;(27{,}7) + 399\,(50)\;\;(11{,}082)}{20000 - 1} = 38{,}20$$

Estimativa de $\overline{\overline{Y}}$

$$\overline{\overline{y}}_{Ac1} = \frac{1}{m\overline{N}} \sum_{i=1}^{m} Y_i' = \frac{3514}{10\,(50)} = 7{,}028$$

Estimativa do coeficiente de correlação intraclasse

$$\hat{\delta} \doteq \frac{\overline{s}_e^2 - \dfrac{s_d^2}{\overline{N}}}{S^2}$$

$$= \frac{11{,}032 - 0{,}554}{38{,}20} = 0{,}276$$

Efeito da conglomeração

$$1 + (\overline{N} - 1)\ \delta = 1 + 49\,(0{,}276) = 14{,}524$$

Tamanho da amostra para dar a mesma precisão de uma Als

$$m\ [1 + (\overline{N} - 1)\ \delta] = 10\ (14{,}524) \doteq 145 \text{ conglomerados}$$

O elevado valor do efeito da conglomeração, mostra que a gaveta com 50 fichas não constitue um bom conglomerado. Seria necessário fazer uma subamostragem, para reduzir esse efeito (Capítulo 4).

16 – EXERCÍCIO

Achar a expressão de δ e do respectivo estimador, para o caso de estimação de proporção.

CAPÍTULO 2 — Amostragem de conglomerados em 1 — estágio — Controle da variação de tamanho: Estimador de razão.

1 — INTRODUÇÃO

Observe-se que $V(y^*_{Ac1}) = M^2 \frac{M-m}{M} \frac{S_e^2}{m}$ aumenta e $Ef = \frac{\overline{N}S^2}{S_e^2}$ diminue quando S_e^2 aumenta. Mas, de acordo com a expressão:

$$S_e^2 = \frac{1}{M-1} \sum_{i=1}^{M} (Y_i - \overline{Y})^2$$

o aumento de S_e^2 é tanto maior quanto mais diferentes forem os totais dos conglomerados. É evidente que um processo ideal para reduzir S_e^2 seria tornar próximos os totais dos conglomerados. Mas esse processo é inviável, posto que os totais não são conhecidos.

No entanto, os totais tendem a ser correlacionados com os tamanhos de modo que, quando os tamanhos crescem, os totais tendem a crescer. Então, basta controlar a variação de tamanho para reduzir S_e^2.

Os processos usuais de controle do tamanho são:

a) Usar um estimador de razão, com característica auxiliar definida pelo tamanho do conglomerado.

b) Selecionar os conglomerados da amostra com probabilidade proporcional ao tamanho.

c) Estratificar os conglomerados, de modo que a característica de estratificação seja o tamanho.

2 – ESTIMADOR DE RAZÃO DE $\overline{\overline{Y}}$

Considere-se a população de conglomerados e duas características associadas: o total e o tamanho.

A média por unidade de π_N é:

$$\overline{\overline{Y}} = \frac{\sum_{i=1}^{M} Y_i}{\sum_{i=1}^{M} N_i} = \frac{\overline{Y}}{\overline{N}}$$

(Dividindo numerador e denominador por M)

Observe-se que $\overline{\overline{Y}}$ pode ser entendida como a razão das médias por conglomerado das características. Sabe-se, pelas propriedades do Estimador de Razão, que um estimador consistente $\hat{R}$ de uma razão R é obtido substituindo o numerador e o denominador pelos respectivos estimadores não tendenciosos.

Desse modo, um estimador consistente de $\overline{\overline{Y}}$ é:

$$\frac{\frac{M}{m}\sum_{i=1}^{m} Y'_i}{\frac{M}{m}\sum_{i=1}^{m} N'_i} = \frac{\sum_{i=1}^{m} Y'_i}{\sum_{i=1}^{m} N'_i}$$

que será representado por $\overline{\overline{y}}^{R}_{Ac1}$.

Observe-se que esse estimador só depende dos tamanhos dos conglomerados da amostra e dos totais correspondentes, enquanto que o estimador não tendencioso também depende do conhecimento de N, tamanho da população.

3 – VARIÂNCIA DE $\overline{\overline{y}}^{R}_{Ac1}$

Recorde-se que, em uma população de M unidades da qual se seleciona uma *Als* de m unidades, a variância do estimador de razão é dada por:

$$V(\hat{R}) = \frac{M-m}{M\overline{X}^2}\,\frac{S^2_{eR}}{m} \quad \text{com } S^2_{eR} = \frac{1}{M-1}\sum_{i=1}^{M}(Y_i - RX_i)^2$$

supondo m suficientemente grande para tornar desprezível a tendenciosidade de $\hat{R}$. Adaptando para $\overline{\overline{y}}^{R}_{Ac1}$ obtém-se:

$$V(\overline{\overline{y}}^{R}_{Ac1}) = \frac{M-m}{M\overline{N}^2}\,\frac{S^2_{eR}}{m}$$

$$\text{com } S_{eR}^2 = \frac{1}{M-1} \sum_{i=1}^{M} (Y_i - \overline{\overline{Y}} N_i)^2 =$$

$$= \frac{1}{M-1} \sum_{i=1}^{M} N_i^2 (\overline{Y}_i - \overline{\overline{Y}})^2$$

4 – ESTIMADOR CONSISTENTE DE $V(\overline{\overline{y}}_{Acl}^{R})$

Sabe-se, também, que um estimador consistente de $V(\hat{R})$ é dado por:

$$\hat{V}(\hat{R}) = \frac{M-m}{M\overline{X}^2} \frac{s_{eR}^2}{m} \quad \text{onde } s_{eR}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (y_i - \hat{R}x_i)^2$$

onde x_i e y_i são elementos da amostra. Adaptando a $V(\overline{\overline{y}}_{Acl}^{R})$ obtém-se:

$$\hat{V}(\overline{\overline{y}}_{Acl}^{R}) = \frac{M-m}{M\overline{N}^2} \frac{s_{eR}^2}{m}$$

$$\text{com } s_{eR}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i' - \overline{\overline{y}}_{R}^{Acl} N_i')^2 =$$

$$= \frac{1}{m-1} \sum_{i=1}^{m} N_i' (\overline{Y}_i' - \overline{\overline{y}}_{Acl}^{R})^2$$

Se $\overline{N}$ não é conhecido, substitue-se pelo estimador não tendencioso:

$$\overline{N}' = \frac{1}{m} \sum_{i=1}^{m} N_i'$$

5 – ESTIMADOR DE RAZÃO DE Y

Considerando que $Y = N\overline{Y} = M\overline{N}\overline{Y}$ tem-se que um estimador consistente de Y é:

$$\overset{*R}{y}_{Acl} = M\overline{N}\ \overline{\overline{y}}_{Acl}^{R} = M\overline{N}\ \frac{\sum_{i=1}^{m} Y_i'}{\sum_{c=1}^{m} N_i'}$$

5.1 — *Variância de* y_{Ac1}^{*R}

Para a variância do estimador do total, tem-se:

$$V(y_{Ac1}^{*R}) = (M\overline{N})^2 V(y_{Ac1}^{R}) = M^2 \frac{M - m}{M} \cdot \frac{S_{eR}^2}{m}$$

5.2 — *Estimador consistente de* $V(y_{Ac1}^{*R})$

É imediata a expressão:

$$V(y_{Ac1}^{*R}) = M^2 \frac{M - m}{M} \cdot \frac{s_{eR}^2}{m}$$

6 — EXEMPLO

O objetivo da pesquisa é estimar o consumo média semanal por domicílio (em unidades do produto) de determinado produto para alimentação.

Dispõe-se, apenas, de um mapa da localidade onde podem ser detectados 400 quarteirões, que serão considerados como conglomerados. Sabe-se que existem na localidade cerca de 26.000 domicílios, dando uma média de 65 domicílios por quarteirão.

Tenciona-se selecionar uma amostra de 650 domicílios, correspondendo a 10 quarteirões e usando um estimador de razão com o tamanho do quarteirão (número de domicílios) como característica auxiliar.

Quanto ao fato de serem os quarteirões considerados conglomerados, convém levar em conta as considerações feitas na Seção 12 — Capítulo 1.

Observe-se, ainda, a correlação entre tamanho e total, importante para o estimador.

Quarteirões da amostra	N.º de unidades consumidas (Y_i')	N.º de domicílios (N_i')	$\overline{Y}_i'$
1	230	70	3,286
2	152	52	2,923
3	250	84	2,976
4	150	54	2,778
5	192	60	3,200
6	184	60	3,067
7	198	66	3,000
8	270	76	3,553
9	225	74	3,041
10	145	50	2,900
Soma	1996	646	30,724

$$\overline{N}' = \frac{646}{10} = 64{,}6$$

Estimativa do número de unidades do produto, por domicílio:

$$\bar{y}_{Ac1} = \frac{1996}{646} = 3{,}090$$

Estimativa da variância de $\bar{y}_{Ac1}^{R}$

$$s_{eR}^2 = \frac{1}{m-1} \sum_{i=1}^{m} N_i'^2 (\overline{Y}_i' - y_{Ac1}^R)^2 = \frac{2061{,}54}{9} = 229{,}06$$

$$\text{donde } V(y_{Ac1}^R) = \frac{M-m}{M\overline{N}^2} \frac{s_{eR}^2}{m} = \frac{400-10}{400(65)^2} \cdot \frac{229{,}06}{10} = 0{,}0053$$

Estimativa do coeficiente de variação de $\bar{y}_{Ac1}^{R}$*:*

$$\hat{\gamma}(\bar{y}_{Ac1}^R) = \frac{\sqrt{0{,}0053}}{3{,}09} = 0{,}0236 \text{ ou } 2{,}36\%$$

É interessante comparar as variâncias do estimador não tendencioso do total e do estimador de razão do total, para se observar a redução decorrente de se considerar a variabilidade das médias em vez da variabilidade dos totais.

Para o estimador não tendencioso, tem-se:

$$\hat{V}(y^*_{Ac1}) = M^2 \frac{M - m}{M} \frac{s_e^2}{m} \quad \text{com } s_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y'_i - \bar{y}_{Ac1})^2$$

$$\text{e } \bar{y}_{Ac1} = \frac{1}{m} \sum_{i=1}^{m} Y_i$$

donde,

$$\hat{V}(y^*_{Ac1}) = 400(390) \frac{1925{,}37}{10} = 30035772$$

Para o estimador de razão, tem-se:

$$\hat{V}(y^{*R}_{Ac1}) = M^2 \frac{M - m}{M} \frac{s_{eR}^2}{m} = 400(390) \frac{229{,}06}{10} = 3573336$$

Portanto, $V(y^*_{Ac1}) = V(y^{*R}_{Ac1})$ 8,4

7 – ESTIMADOR DE 1 AZÃO DE PROPORÇÃO

Das expressões de y^R_{Ac1}, $V(y^R_{Ac1})$ e $\overline{V}(y^R_{Ac1})$ do estimador de razão de $\overline{\overline{Y}}$ obtém-se:

$$p^R_{Ac1} = \frac{\sum_{i=1}^{m} A'_i}{\sum_{i=1}^{m} N'_i}$$

$$V(p^R_{Ac1}) = \frac{M - m}{Mm(M-1)} \sum_{i=1}^{M} \left(\frac{N_i}{\overline{\overline{N}}} \right)^2 (P_i - P)^2$$

$$\hat{V}(p^R_{Ac1}) = \frac{M - m}{Mm(M-1)} \sum_{i=1}^{m} \left(\frac{N'_i}{\overline{\overline{N}}} \right)^2 (P'_i - p^R_{Ac1})^2$$

7.1 — *Exemplo*

Considere-se o Exemplo 2.1.3. Suponha-se que, além de investigar o número de unidades consumidas do produto em cada domicílio levantou-se, também, o fato do domicílio ser próprio ou não, obtendo-se o quadro:

Quarteirões	N.º de domicílios (N_i')	N.º de domicílios próprios (A_i')	$P_i' = \frac{A_i'}{N_i'}$
1	70	25	0,357
2	52	10	0,192
3	84	36	0,429
4	54	15	0,278
5	60	15	0,250
6	60	18	0,300
7	66	22	0,333
8	76	36	0,474
9	74	30	0,405
10	50	21	0,420
Soma	646	228	

$$p_{Acl}^{R} = \frac{228}{646} = 0{,}353 \text{ ou } 35{,}3\%$$

$$\overline{N}' = \frac{646}{10} = 64{,}6$$

$$\hat{V}(p_{Acl}^{R}) = \frac{400 - 10}{400\,(10)\,(9)}\,0{,}0689081 = 0{,}000764$$

$$\sqrt{\hat{V}(p_{Acl}^{R})} = 0{,}0276 \text{ ou } 276\%$$

8 — ESTIMADOR DE RAZÃO EM RELAÇÃO A UMA CARACTERÍSTICA QUE NÃO SEJA O TAMANHO

No estimador de razão em relação ao tamanho, usa-se o tamanho como característica auxiliar x, o que permite estimar a média de y

por unidade da população. Por exemplo, estimar o número de mulheres por domicílio, a produção por estabelecimento da localidade, etc.

No estimador de razão em relação a uma característica auxiliar x que não seja o tamanho, estima-se a média de y por unidade da característica auxiliar. Por exemplo, estimar a taxa de mulheres analfabetas, a produção por hectare, etc.

8.1 — *Estimador consistente de R*

A relação desejada entre as duas características é:

$$R = \frac{\bar{\bar{Y}}}{\bar{\bar{X}}} = \frac{Y}{X}$$

Sabe-se, conforme já foi relembrado na Seção 2.1, que um estimador consistente de R é obtido substituindo Y e X por estimadores não tendenciosos. Desse modo, tem-se para o estimador consistente:

$$\hat{R} = \frac{\frac{M}{m} \sum_{i=1}^{m} Y_i'}{\frac{M}{m} \sum_{i=1}^{m} X_i'} = \frac{\sum_{i=1}^{m} Y_i'}{\sum_{i=1}^{m} X_i'}$$

8.2 — *Variância de* $\hat{R}$

Repete-se a expressão dada em 3.

$$V(\hat{R}) = \frac{M - m}{M\bar{X}^2} \frac{S_{eR}^2}{m} \quad \text{com} \quad S_{eR}^2 = \frac{1}{M - 1} \sum_{i=1}^{M} (Y_i - RX_i)^2$$

8.3 — *Estimador consistente de* $V(\hat{R})$

$$\hat{V}(\hat{R}) = \frac{M - m}{M\bar{X}^2} \frac{s_{eR}^2}{m} \text{ com } s_{eR}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i' - \hat{R}X_i')^2$$

8.4 — *Estimador consistente de* Y

Conforme se sabe pelo estudo do estimador de razão na *Als*, é:

$$y_{Ac1}^{*R} = \hat{R}\ X$$

onde X é o total da característica auxiliar, suposto conhecido.

8.4.1 — Variância de y_{Ac1}^{*R}

$$V(y_{Ac1}^{*R}) = V(\hat{R})\ X^2 = M^2 \frac{M - m}{M} \cdot \frac{S_{eR}^2}{m}$$

8.4.2 — Estimador consistente de $V(y_{Ac1}^{*R})$

$$V(y_{Ac1}^{*R}) = M^2 \frac{M - m}{M} \cdot \frac{s_{eR}^2}{m}$$

9 — EXEMPLO

Considere-se o Exemplo 6. Suponha-se que se deseja estimar o número de unidades consumidas por morador. Nesse caso, a característica auxiliar é o número de moradores dos quarteirões, em vez do número de domicílios. Tem-se o seguinte quadro da amostra:

QUARTEIRÕES DA AMOSTRA	N.º DE UNIDADES CONSUMIDAS (Y_i')	N.º DE MORADORES (X_i')
1	230	301
2	152	176
3	250	428
4	150	140
5	192	160
6	184	171
7	198	180
8	270	320
9	225	290
10	145	130
Soma	1 996	2 296

Estimativa do número de unidades consumidas, por morador:

$$\hat{R} = \frac{1996}{2296} = 0{,}87$$

O número médio de moradores por quarteirão é, aproximadamente, $\overline{X} = 230$

$$s^2_{eR} = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \hat{R}X_i)^2 = \frac{24360{,}57}{9} = 2706{,}73$$

donde $\hat{V}(\hat{R}) = \frac{400-10}{400(230)^2} \cdot \frac{2706{,}73}{10} = 0{,}00499$ e $\sqrt{\hat{V}(\hat{R})} = 0{,}0706$

CAPÍTULO 3 — Amostragem de conglomerados em 1 — estágio — Controle da variação de tamanho: Probabilidade desigual de seleção, com reposição.

1 — INTRODUÇÃO

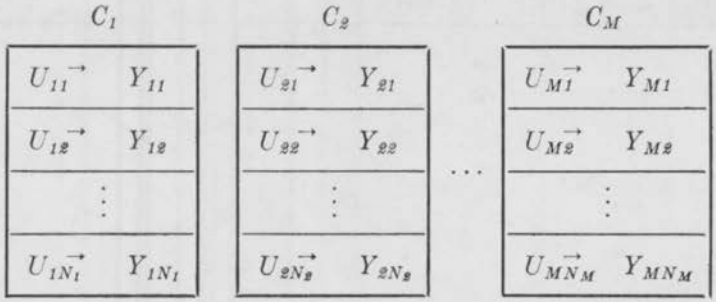
A formação de conglomerados com tamanho igual controla a variação de tamanho da amostra e a influência da variação de tamanho na variância do estimador.

No entanto, a ocorrência de conglomerados com tamanho igual não é um fato comum. Geralmente a igualdade de tamanho é obtida artificialmente, retirando partes dos conglomerados maiores ou reunindo conglomerados menores. Por exemplo, em pesquisas domiciliares em que os conglomerados sejam quarteirões, os quarteirões com muitos domicílios podem ser desmembrados em 2 ou mais conglomerados, de modo a se obter um número pelo menos aproximadamente igual de domicílios em cada conglomerado.

Neste Capítulo, estuda-se outra forma de controle, mantendo o tamanho desigual mas variando a probabilidade de seleção. Inicialmente apresenta-se a teoria com probabilidade desigual de seleção e, posteriormente, como essas probabilidades podem ser estabelecidas. A seleção é feita com reposição, com o objetivo de manter a probabilidade de seleção constante, simplificando, em conseqüência, as expressões dos estimadores.

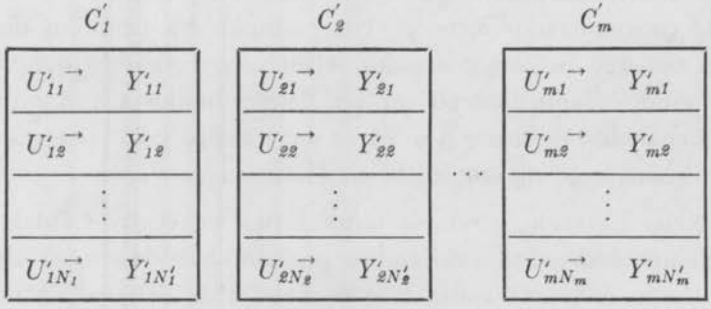
2 – CONFIGURAÇÃO DA AMOSTRA

Como das situações anteriores, as unidades de π_N são grupadas em M conglomerados que podem ter tamanhos desiguais:



Seja P_i a probabilidade de seleção de C_i $(i = 1, 2, \ldots, M)$ com $\sum_{i=1}^{M} P_i = 1$

Selecione-se uma amostra de m conglomerados, de acordo com as probabilidades de seleção P_i, com reposição.



A reposição influe, mantendo as probabilidades de cada conglomerado e, em decorrência, simplificando as expressões dos estimadores.

2.1 — *Parâmetros de y*

Total em C_i: $Y_i = \sum_{j=1}^{N_i} Y_{ij}$

Média em C_i: $\overline{Y}_i = \frac{Y_i}{N_i}$

Variância em C_i: $S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \overline{Y}_i)^2$

Total de y: $Y = \sum_{i=1}^{M} \sum_{j=1}^{N_i} Y_{ij} = \sum_{i=1}^{M} Y_i$

Média de y

(por unidade de π_N): $\overline{\overline{Y}} = \frac{Y}{N}$; $N = \sum_{i=1}^{M} N_i$

Média por conglomerados: $\overline{Y} = \frac{Y}{M}$

2.2 — *Estatísticas*

Total em C_i: $Y'_i = \sum_{j=1}^{N'_i} Y'_{ij}$

Média em C_i: $\overline{Y}'_i = \frac{Y'_i}{N'_i}$

Variância em C_i: $S_i'^2 = \frac{1}{N'_i - 1} \sum_{j=1}^{N'_i} (Y'_{ij} - \overline{Y}'_i)^2$

3 — TEOREMA

Um estimador não tendencioso de Y é:

$$y_{Act}^{*P} = \frac{1}{m} \sum_{i=1}^{m} \frac{Y'_i}{P'_i}$$

Prova

$$E(y_{Ac1}^{*P}) = \frac{1}{m} \sum_{i=1}^{m} E\left(\frac{Y_i'}{P_i'}\right) = \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{i=1}^{M} \frac{Y_i}{P_i} P_i\right) =$$

$$= \sum_{i=1}^{M} Y_i = Y$$

3.1 — *Corolário*

Um estimador não tendencioso de $\overline{\overline{Y}}$ é :

$$\overline{\overline{y}}_{Ac1}^{P} = \frac{1}{mN} \sum_{i=1}^{m} \frac{Y_i'}{P_i'}$$

Prova

Fazer como exercício

4 — VARIÂNCIA DE y_{Ac1}^{*P}

$$V(y_{Ac1}^{*P}) = E(y_{Ac1}^{*P} - Y)^2 = E(y_{Ac1}^{*P}) - Y^2 =$$

$$= E\frac{1}{m}\left(\sum_{i=1}^{m} \frac{Y_i'}{P_i'}\right)^2 - Y^2 =$$

$$= \frac{1}{m^2} E\left(\sum_{i=1}^{m} \frac{Y_i'^2}{P_i'^2} + \sum_{\substack{i=1 \\ (\neq j))}}^{m} \sum_{j=1}^{m} \frac{Y_i'}{P_i'} \frac{Y_j'}{P_j'}\right) - Y^2 =$$

$$= \frac{1}{m^2}\left[\sum_{i=1}^{m} E\left(\frac{Y_i'^2}{P_i'^2}\right) + \sum_{i=1}^{m} \sum_{j=1}^{m} E\left(\frac{Y_i'}{P_i'} \cdot \frac{Y_j'}{P_j'}\right)\right] - Y^2 =$$

Considerando que há reposição dos conglomerados, $\frac{Y_i'}{P_i'}$ e $\frac{Y_j'}{P_j'}$ são independentes. Por conseguinte,

$$E\left(\frac{Y_i'}{P_i'} \cdot \frac{Y_j'}{P_j'}\right) = E\left(\frac{Y_i'}{P_i'}\right) E\left(\frac{Y_j'}{P_j'}\right) = \left(\sum_{i=1}^{M} \frac{Y_j}{P_i} P_i\right)^2 = Y^2$$

donde,

$$V(y^{*P}_{Ac1}) = \frac{1}{m^2}\left[m \sum_{i=1}^{M} \frac{Y_i^2}{P_i^2} P_i + m(m-1)\ Y^2\right] - Y^2 =$$

$$= \frac{1}{m}\left[\sum_{i=1}^{M} \frac{Y_i^2}{P_i^2} P_i - Y^2\right] =$$

$$= \frac{1}{m} \sum_{i=1}^{M} \left(\frac{Y_i^2}{P_i^2} - Y^2\right) P_i = \frac{1}{m} \sum_{i=1}^{M} \left(\frac{Y_i}{P_i} - Y\right)^2 P_i$$

Pondo

$$S^2_{eP} = \sum_{i=1}^{M} \left(\frac{Y_i}{P_i} - Y\right)^2 P_i$$

pode-se escrever:

$$V(y^{*P}_{Ac1}) = \frac{S^2_{eP}}{m}$$

5 – TEOREMA

Um estimador não tendencioso de $V(y^{*P}_{Ac1})$ é:

$$\hat{V}(y^{*P}_{Ac1}) = = \frac{s^2_{eP}}{m} \text{ onde } s^2_{eP} = \frac{1}{m-1} \sum_{i=1}^{m} \left(\frac{Y'_i}{P'_i} - y^{*P}_{Ac1}\right)^2$$

Prova

Escreve-se: $\hat{V}(y^{*P}_{Ac1}) = \frac{1}{m(m-1} \left[\sum_{i=1}^{m} \frac{Y'^2_i}{P'^2_i} - m\ y^{*P}_{Ac1}\right]$

donde,

$$E[V(y^{*P}_{Ac1})] = \frac{1}{m(m-1)}\left[m \sum_{i=1}^{M} \frac{Y_i^2}{P_i^2} P_i - m\ E(y^{*P^2}_{Ac1})\right] =$$

$$= \frac{1}{m-1}\left[\sum_{i=1}^{M} \frac{Y_i^2}{P_i^2} P_i - V(y^{*P}_{Ac1} - y^2\right] =$$

$$= \frac{1}{m-1}\left[\sum_{i=1}^{M} \left(\frac{Y_i}{P_i} - Y\right)^2 - V(y^{*P}_{Ac1})\right] =$$

$$= \frac{1}{m-1}\left[mV(y^{*P}_{Ac1}) - V(y^{*P}_{Ac1})\right] = V(y^{*P}_{Ac1})$$

6 – PROBABILIDADE DE SELEÇÃO PROPORCIONAL AO TAMANHO DO CONGLOMERADO

Considere-se a variância de y_{Acl}^{*P} estimador não tendencioso de Y:

$$V(y_{Acl}^{*P}) = \frac{1}{m} \sum_{i=1}^{M} \left(\frac{Y_i}{P_i} - Y \right)^2 P_i$$

Se $P_i = \frac{Y_i}{Y}$ proporção do total de C_i $(i = 1, 2, \ldots M)$ obtem-se $V(y_{Acl}^{*}) = 0$. Portanto, a seleção com probabilidade proporcional ao total controla a variação de tamanho dos conglomerados.

O problema com a definição de P_i proporcionalmente ao total, está em que não são conhecidos os totais Y_i para todos os conglomerados. Desse modo, é necessário dar uma nova definição a P_i que traga um valor próximo ao obtido com a definição anterior.

Admitindo que os totais sejam proporcionais aos tamanhos, hipótese essa que pode ser considerada com razoável aproximação, define-se:

$$P_i = \frac{N_i}{N} \qquad (i = 1, 2 \ldots, M)$$

Embora seja usual fazer $P_i = \frac{N_i}{N}$, nem sempre se conhecem os valores N_i para todos os conglomerados.

Nesse caso, a solução é considerar uma característica x conhecida para todas as unidade de π_N e que assuma valores proporcionais aos valores de y e definir:

$$P_i = \frac{X_i}{X} \text{ com } X = \sum_{i=1}^{M} X_i$$

Essa característica x é chamada "medida de tamanho".

Em particular, uma boa medida de tamanho é a própria característica y, observada em outra ocasião. Por exemplo, na estimação da população de certa localidade, pode-se usar como medida de tamanho, a população observada no último Censo.

Na estimação da produção de café, a área das fazendas é uma medida de tamanho. Outra medida pode ser a área plantada.

Se as *UP* são hospitais, pode-se usar como medida de tamanho o número de leitos ou o número de leitos ocupados.

Na estimação do faturamento de estabelecimentos comerciais, uma medida de tamanho é o número de empregados.

7 – MODO DE SELECIONAR A AMOSTRA COM PROBABILIDADE DE SELEÇÃO PROPORCIONAL AO TAMANHO (OU A UMA MEDIDA DE TAMANHO)

7.1 – *Seleção com uma tabela de números aleatórios*

Para explicar o processo, suponha-se que certa localidade tem 8 640 domicílios em 270 conglomerados. Trata-se de selecionar uma amostra de 10 conglomerados.

a) Parte-se da listagem dos 270 conglomerados com os respectivos tamanhos; número de domicílios (ou uma medida de tamanho).

CONGLOMERADOS (1)	N.º DE DOMICÍLIOS (2)	ACUMULADO (3)
1	48	48
2	29	77
3	35	112
4	28	140
5	32	172
6	33	205
7	21	226
.	.	.
.	.	.
.	.	.
269	42	8 598
270	30	8 540

b) Com uma tabela de números aleatórios, selecionam-se 10 números de 0 001 a 8 640. Suponha-se que o primeiro número selecionado seja 0 138. A posição deste número é marcada na coluna (3) do quadro com 1. Se o segundo número selecionado é 0 218, marca-se na coluna (3) com 2. Se o terceiro é 0 225, marca-se com 3 na mesma posição do anterior e assim por diante, até marcar 10 números.

c) Cada marca na coluna (3) indica o conglomerado que vai para a amostra. Observe-se que pode ocorrer mais de uma marca para o mesmo conglomerado, conforme aconteceu com o conglomerado 7. Nesse caso, o conglomerado 7 tem os valores de y repetidos duas vezes na amostra.

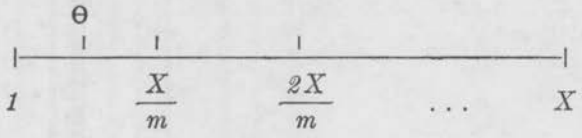
7.2 — *Seleção sistemática*

A seleção dos conglomerados com probabilidade proporcional ao tamanho, é feita com reposição. No entanto, se M é grande e $M \ldots m$, a probabilidade de um conglomerado ser selecionado mais de uma vez é muito pequena e, como aproximação, pode-se usar a seleção sistemática.

Se a seleção é proporcional a uma medida de tamanho, a probabilidade de seleção do conglomerado i é:

$$m \frac{X_i}{X} = \frac{X_i}{X/m}$$

Divide-se X em partes, sendo $\frac{X}{m}$ o intervalo de amostra para fins de seleção sistemática,



Seleciona-se um ponto de partida θ no intervalo $\left[1; \frac{X}{m}\right]$ ponto esse que vai determinar o primeiro conglomerado da amostra. A θ soma-se $\frac{X}{m}$ determinando o segundo conglomerado da amostra e assim por diante, conforme se esclarece no exemplo que se segue.

Se $X_i > \frac{X}{m}$ o conglomerado é auto-representado, isto é, o seu total Y_i é somado ao total estimado dos demais conglomerados não auto-representados.

8 – EXEMPLO

Para estimar a população de certa localidade, considerou-se os Setores Censitários como conglomerados dos domicílios da localidade. Existem 420 setores na localidade, dos quais 10 serão selecionados para a amostra. A seleção é feita sistematicamente, usando-se como medida de tamanho a população registrada no último Censo, conforme o quadro:

SETORES	POPULAÇÃO X_i (Censo)	ACUMULADO
1	570	570
2	180	750
3	270	1 020
4	400	1 420
5	480	1 900
6	377	2 277
6	377	2 277
.	.	.
.	.	.
.	.	.
420	221	121 212
Soma	121212	—

Tem-se:

$$X = 121212$$

$$\frac{X}{m} = 12121$$

Selecione-se um número inteiro no intervalo $[1;12121]$. Seja $\theta = 00920$ esse número.

. Na coluna "acumulado" localiza-se esse número e marca-se *, correspondendo ao conglomerado 3 que, assim, vai para a amostra.

O número seguinte é $\theta + 12121 = 920 + 12121 = 13040$ que também é marcado na coluna "acumulado" e vai determinar o segundo conglomerado da amostra. E assim por diante, até serem selecionados 10 conglomerados.

Os dados da amostra foram os seguintes:

SETORES	N.º DE MORADORES CENSO — (X'_i)	PROBABILIDADE DE SELEÇÃO (P'_i)	N.º DE MORADORES AMOSTRA (Y'_i)
1	270	0,0022	320
2	386	0,0032	300
3	401	0,0033	540
4	424	0,0035	471
5	247	0,0020	240
6	377	0,0031	401
7	473	0,0039	502
8	206	0,0017	320
9	367	0,0030	450
10	275	0,0023	370

Observação — As probabilidades de seleção dos setores da amostra foram obtidas do primeiro quadro.

Estimativa do número de moradores:

$$y^{*P}_{Ac1} = \frac{1}{m} \sum_{i=1}^{m} \frac{Y'_i}{P'_i} = \frac{1414589,98}{10} = 141459 \text{ habitantes}$$

Estimativa da variância de y^{*}_{Ac1}

$$s^2_{eP} = \frac{1}{(m-1)} \sum_{i=1}^{m} \left(\frac{Y'_i}{P'_i} - y^{*P}_{Ac1} \right)^2 = \frac{6238460878}{9}$$

$$= 693162319,8$$

$$V(y^{*P}_{Ac1}) = 69316231,98$$

$$\sqrt{\hat{V}(y^{*P}_{Ac1})} = 8325,64$$

$$\hat{\gamma}(y^{*P}_{Ac1}) = 0,059 \text{ ou } 5,9\%$$

9 – EXERCÍCIO

Achar as expressões dos estimadores para a estimação de proporção, com probabilidade desigual de seleção.

10 – EXEMPLO

Suponha-se, no Exemplo 8, que também se levantou o número de pessoas alfabetizadas nos Setores da amostra, obtendo-se:

SETORES	N.º DE MORADORES	N.º DE MORADORES ALFABETIZADOS
1	320	200
2	300	130
3	540	340
4	471	290
5	240	90
6	401	240
7	502	310
8	320	160
9	450	340
10	370	160

Estimar a proporção de moradores alfabetizados e a respectiva variância.

11 – COEFICIENTE DE CORRELAÇÃO INTRACLASSE

As expressões que se seguem, supõem $P_i = \frac{N_i}{N}$, ou seja, probabilidade de seleção dos conglomerados proporcional ao tamanho. As expressões resultantes serão usadas como aproximação, mesmo no caso em que as probabilidades de seleção sejam proporcionais a uma medida de tamanho.

No Capítulo 1 mostrou-se que o coeficiente de correlação intraclasse, podia ser escrito na forma:

$$\delta = \frac{E(Y'_{ij} - \overline{Y}'_i)\,(Y'_{ik} - \overline{Y}'_i) + E(\overline{Y}'_i - \overline{\overline{Y}})}{E(Y'_{ij} - \overline{\overline{Y}})} \quad (j \neq k) \quad (11.1)$$

Nessa expressão tem-se, agora:

a) $E(Y'_{ij}-\overline{Y}'_i)(Y'_{ik}-\overline{Y}'_i)=\underset{C'_j}{E}\{E[(Y'_{ij}-\overline{Y}'_i)(Y'_{ik}-\overline{Y}'_i)\,|\,C'_i \text{ fix.}]\}=$

$$=\underset{C'_i}{E}\left[\frac{1}{N'_i(N'_i-1)}\sum_{\substack{j=1\\(j\neq k)}}^{N'_i}\sum_{k=1}^{N'_i}(Y'_{ij}-\overline{Y}'_i)(Y'_{ik}-\overline{Y}'_i)\right]=$$

$$=\underset{C'_i}{E}\left[-\frac{1}{N'_i(N'_i-1)}\sum_{j=1}^{N'_i}(Y'_{ij}-\overline{Y}'_i)^2\right]=$$

$$=\underset{C'_i}{E}\left[-\frac{S'^2_i}{N'_i}\right]=-\sum_{i=1}^{M}\frac{S_i^2}{N_i}P_i=$$

$$=-\sum_{i=1}^{M}\frac{S_i^2}{N_i}=-\sum_{i=1}^{M}\frac{S_i^2}{N}$$

Sendo $\overline{N}=\frac{N}{M}$ tamanho médio dos conglomerados, tem-se $N=M\,\overline{N}$. Pondo, ainda,

$$\frac{1}{M}\sum_{i=1}^{M}S_i^2=S_d^2$$

obtém-se, finalmente:

$$E(Y'_{ij}-\overline{Y}'_i)(Y'_{ik}-\overline{Y}'_i)=-\frac{S_d^2}{\overline{N}}$$

b) $E(\overline{Y}'_i-\overline{\overline{Y}})^2=\sum_{i=1}^{M}(\overline{Y}_i-\overline{\overline{Y}})^2P_i=\sum_{i=1}^{M}\frac{N_i}{N}(\overline{Y}_i-\overline{\overline{Y}})^2=\overline{S}_{eP}^2$

c) $E(Y'_{ij}-\overline{\overline{Y}})^2=\underset{C'_i}{E}\left\{E[(Y'_{ij}-\overline{\overline{Y}})^2\mid C'_i \text{ fix.}]\right\}=$

$$=\underset{C'_i}{E}\left[\sum_{j=1}^{N'_i}\frac{(Y'_{ij}-\overline{\overline{Y}})^2}{N'_i}\right]=\sum_{j=1}^{M}\sum_{j=1}^{N_i}\frac{(Y_{ij}-\overline{\overline{Y}})^2}{N_i}\frac{N_i}{N}=$$

$$=\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(Y_{ij}-\overline{\overline{Y}})^2=\frac{N-1}{N}S^2$$

Substituindo a), b) e c) em (11.I) obtém-se uma expressão aproximada para δ:

$$\delta = \frac{\overline{S}^2_{eP}}{S^2} - \frac{S^2_d}{\overline{N}} \qquad \text{com } \frac{N-1}{N} = 1$$

12 – RELACIONAMENTO ENTRE S^2, S^2_{eP} e S^2_d

Ponha-se,

$$\sum_{i=1}^{M}\sum_{j=1}^{N_i}(Y_{ij}-\overline{\overline{Y}})^2 = \sum_{i=1}^{M}\sum_{j=1}^{N_i}[(Y_{ij}-\overline{Y}_i)+(\overline{Y}_i-\overline{\overline{Y}})]^2 =$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}(Y_{ij}-\overline{Y}_i)^2 + \sum_{i=1}^{M} N_i N(\overline{Y}_i-\overline{\overline{Y}})^2$$

ou,

$$(N-1)S^2 = \sum_{i=1}^{M}(N_i-1)S_i^2 + \sum_{i=1}^{M} N_i(\overline{Y}_i-\overline{\overline{Y}})^2$$

$$\frac{N-1}{N}S^2 = \sum_{i=1}^{M}\frac{N_i-1}{N}S_i^2 + \sum_{i=1}^{M}\frac{N_i}{N}(\overline{Y}_i-\overline{\overline{Y}})^2$$

Com as aproximações,

$$\frac{N-1}{N} \doteq 1 \text{ e } \frac{N_i-1}{N} \doteq \frac{\overline{N}-1}{M\overline{N}} \doteq \frac{1}{M}$$

obtem-se:

$$S^2 \doteq S^2_d + \overline{S}^2_{eP} \qquad (12.I)$$

13 – ESTIMADOR DE δ

Estimam-se, a seguir, as variâncias que participam da expressão de δ.

a') *Estimador não tendencioso de* S^2_d

Seja o estimador,

$$s^2_d = \frac{\overline{N}}{m}\sum_{i=1}^{m}\frac{S_i'^2}{N_i'}$$

Tem-se:

$$E(s_d^2) = \frac{\overline{N}}{m} \sum_{i=1}^{m} \left[\sum_{i=1}^{M} \frac{S_i^2}{N_i} \frac{N_i}{N} \right] = \frac{\overline{N}}{M\overline{N}} \sum_{i=1}^{M} S_i^2 = S_d^2$$

b') *Estimador não tendencioso de* $\overline{S}_{eP}^2$

Recorde-se a expressão: $S_{eP}^2 = \sum_{i=1}^{M} \left(\frac{Y_i'}{P_i} - Y \right)^2 P^i$

Com $P_i = \frac{N_i}{N}$ obtém-se:

$$S_{eP}^2 = \sum_{i=1}^{M} N^2 (\overline{Y}_i - \overline{\overline{Y}})^2 \frac{N_i}{N} = N^2 \overline{S}_{eP}^2$$

donde,

$$\overline{S}_{eP}^2 = \frac{S_{eP}^2}{N^2}$$

Considerando que um estimador não tendencioso de S_{eP}^2 é:

$$s_{eP}^2 = \frac{N^2}{m-1} \sum_{i=1}^{m} (\overline{\overline{Y}}_i' - \overline{\overline{y}}_{Ac1})^2$$

conclui-se que um estimador não tendencioso de $\overline{S}_{eP}^2$ é o mesmo de $\frac{S_{eP}^2}{N^2}$, ou seja,

$$\overline{s}_{eP}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\overline{Y}' - \overline{\overline{y}}_{Ac1}^P)^2$$

c') Substituindo os estimadores definidos em a') e b') em (12.I), obtém-se o estimador de S^2:

$$s \doteq s_d^2 - \overline{s}_{eP}^2$$

Finalmente, substituindo os resultados de a'), b') e c') na expressão de δ, obtém-se o estimador:

$$\hat{\delta} = \frac{\overline{s}_{eP}^2 - \frac{s_d^2}{\overline{N}}}{s^2}$$

14 – EXEMPLO

Considere-se o Exemplo 9. Complete-se o quadro da amostra com as informações referentes ao "N.º de domicílios" e a "Variância entre os domicílios da amostra (N.º de moradores)".

SETORES	N.º DE MORADORES (X'_i)	PROB. DE SELEÇÃO (P'_i)	N.º DE DOMICÍLIOS (N'_i)	TOTAL DE MORADORES (Y'_i)	VARIANCIA DO N.º DE MORADORES (S'^2_i)
1	270	0,0022	71	320	16,09
2	386	0,0032	86	300	18,00
3	401	0,0033	129	540	15,25
4	424	0,0035	84	471	21,05
5	247	0,0020	52	240	13,80
6	377	0,0031	93	401	17,70
7	473	0,0039	116	502	21,30
8	206	0,0017	94	320	26,40
9	367	0,0030	98	450	18,40
10	275	0,0023	67	370	17,00

Supondo que as probabilidades de seleção sejam proporcionais ao tamanho, como uma aproximadação, tem-se:

$$\bar{s}^2_{eP} = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{Y}'_i - \bar{\bar{y}}^P_{Ac1})^2 \text{ com } \bar{\bar{y}}^P_{Ac1} = \frac{1}{m} \sum_{i=1}^{m} \bar{Y}'_i = \frac{44,561}{10} = 4,456$$

donde $\bar{s}^2_{eP} = \frac{4,6613}{9} = 0,5179$

$$s^2_d = \frac{\bar{N}}{m} \sum_{i=1}^{m} \frac{S'^2_i}{N'_i} = \frac{90}{10}\ 2,1664 = 19,4976$$

$$s^2 = \bar{s}^2_{eP} + s^2_d = 20,0155$$

Então, o coeficiente de correlação intraclasse é estimado por:

$$\hat{\delta} = \frac{0,5179 - \frac{19,4976}{90}}{20,0155} = 0,015$$

O efeito da conglomeração é estimado em:

$$1 + (\overline{N} - 1)\ \hat{\delta} = 2{,}35$$

Para reduzir esse efeito da conglomeração, pode-se fazer subamostragem, conforme será estudado no Capítulo 4.

15 – EXERCÍCIO

Dar as expressões dos estimadores, variância, coeficiente de correlação intraclasse, relativas a estimação de proporção.

16 – ESTRATIFICAÇÃO DE CONGLOMERADOS

16.1 – *Introdução*

Estratificando os conglomerados com a característica de estratificação representada pelo tamanho, as variâncias dentro de cada estrato dependem de totais próximos, de modo que se obtém um controle da variação de tamanho.

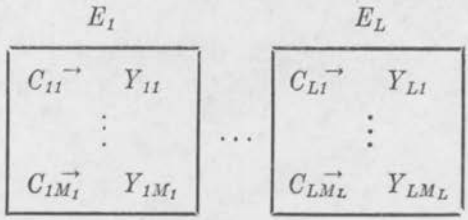
Por exemplo, se a população é de domicílios e os conglomerados são Setores Censitários, esses setores podem ser estratificados pelo número de domicílios. Se os tamanhos não são conhecidos, estratifica-se por uma "medida de tamanho", isto é, por uma característica correlacionada com o tamanho. Assim, no exemplo acima, os Setores Censitários podem ser estratificados pela população registrada no último Censo.

A estratificação pode ser simples, com probabilidade igual de seleção dos conglomerados e sem reposição, dentro de cada estrato, ou pode ser feita com um estimador de razão ou com probabilidade desigual de seleção.

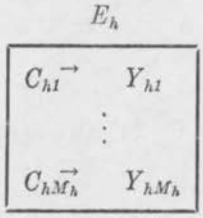
No que se segue, estuda-se a primeira situação. A adaptação às outras duas modalidades, pode ser feita sem dificuldade.

16.2 — *Configuração da amostra*

Suponham-se os M conglomerados grupados em L estratos $E_1, E_2, \ldots E_L$ e, associado a cada conglomerado, o respectivo total:



Seja um conglomerado genérico E_h $(h = 1, 2, \ldots L)$



Tamanho do Estrato: M_h

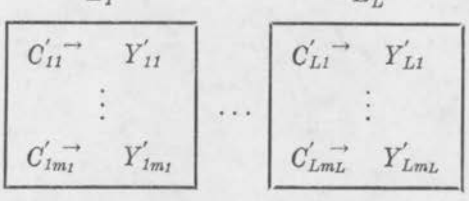
Total do Estrato: $Y_h = \sum_{i=1}^{M_h} Y_{hi}$

Média por conglomerado no Estrato: $\overline{Y}_h = \dfrac{Y_h}{M_h}$

Variância entre os totais dos conglomerados do Estrato:

$$S_{he}^2 = \frac{1}{M_h - 1} \sum_{i=1}^{M_h} (Y_{hi} - \overline{Y})^2$$

Selecione-se em cada estrato, uma amostra aleatória simples sem reposição, de $m_1, m_2, \ldots m_L$ conglomerados, respectivamente.



A amostra estratificada é:

$$\{Y'_{11},\ Y'_{12},\ \ldots\ Y'_{1m_1};\ \ldots,\ Y'_{L1},\ Y'_{L2},\ \ldots\ Y'_{Lm_L}\}$$

16.3 — *Estimadores em* E_h

Com a amostra de m_h conglomerados em E_h, aplicam-se as fórmulas já conhecidas do estimador não tendencioso, acrescentando-se o índice h.

16.3.1 — Estimador não tendencioso do total Y_h

$$y^*_{h.Ac1} = \frac{M_h}{m_h} \sum_{i=1}^{m_h} Y'_{hi}$$

16.3.2 — Variância de $y^*_{h.Ac1}$

$$V(y^*_{h.Ac1}) = M_h^2 \frac{M_h - m_h}{M_h} \cdot \frac{S^2_{he}}{m_h}$$

16.3.3 — Estimador não tendencioso de $V(y^*_{h.Ac1})$

$$V(y^*_{h.Ac1}) = M_h^2 \frac{M_h - m_h}{M_h} \cdot \frac{s^2_{he}}{m_h}$$

$$\text{onde } s^2_{he} = \frac{1}{m_h - 1} \sum_{i=1}^{m_h} (Y'_{hi} - \bar{y}_{h.Ac1})^2$$

$$\text{e } \bar{y}_{h.Ac1} = \frac{1}{m_h} \sum_{i=1}^{m_h} Y'_{hi}$$

16.4 — *Estimador e variância do total geral* Y

16.4.1 — Estimador não tendencioso de Y

$$y^{*Est}_{Ac1} = \sum_{h=1}^{L} y^*_{h.Ac1} = \sum_{h=1}^{L} \frac{M_h}{m_h} \sum_{i=1}^{m_h} Y'_{hi}$$

16.4.2 — Variância de y^{*Est}_{Ac1}

$$V(y^{*Est}_{Ac1}) = \sum_{h=1}^{L} V(y^*_{h.Ac1}) = \sum_{h=1}^{L} M_h^2 \frac{M_h - m_h}{M_h} \cdot \frac{S^2_{he}}{m_h}$$

16.4.3 — Estimador não tendencioso de $V(y_{Acl}^{*Est})$

$$V(y_{Acl}^{*Est}) = \sum_{h=1}^{L} M_h^2 \frac{M_h - m_h}{M_h} \cdot \frac{s_{he}^2}{m_h}$$

16.5 — *Amostra autoponderada*

Se a fração de amostragem $\frac{m_h}{M_h}$, em E_h é constante $(h = 1, 2, \ldots L)$ e é igual à fração geral de amostragem $f = \frac{m}{M}$ obtém-se:

$$y_{Acl}^{*Est} = \frac{1}{f} \sum_{h=1}^{E} \sum_{i=1}^{m_h} Y'_{hi}$$

$$V(y_{Acl}^{*Est}) = \frac{1-f}{f} \sum_{h=1}^{L} M_h S_{he}^2$$

$$\hat{V}(y_{Acl}^{*Est}) = \frac{1-f}{f} \sum_{h=1}^{L} M_h s_{he}^2$$

16.6 — *Exemplo*

Em certa localidade, existem 1.200 Setores Censitários que vão ser considerados como conglomerados de domicílios. Foram formados 6 estratos, de acordo com a população do último Censo.

Quadro 1

ESTRATOS	N.º DE SETORES (M_h)
	90
1	100
2	140
3	250
4	295
5	325
6	
Soma	1 200

A população total da localidade, de acordo com o Censo, foi de 1.960.800 habitantes, o que corresponde a uma média de 1.634 habitantes por Setor ou 380 domicílios por Setor (na base de 4,3 pessoas por domicílio, de acordo com pesquisa anterior). Considerando as disponibilidades de tempo e custo, foi fixada uma amostra de 24 Setores ou, aproximadamente, 9.120 domicílios, o que corresponde à fração de amostragem de $\frac{24}{1200} = \frac{1}{50}$

Aplicando essa fração no Quadro 1, obtém-se a amostra de Setores. Nesses Setores, levantou-se o número de habitantes, com o objetivo de estimar a população atual da localidade, obtendo-se:

Quadro 2

ESTRATOS	N.º DE SETORES (M_h)	N.º DE SETORES DA AMOSTRA (M_h)	N.º DE HABITANTES NOS SETORES DA AMOSTRA (Y'_{hi})
1	90	2	3450;3120
2	100	2	2980;3060
3	140	3	2320;2850;2010
4	250	5	1910;1990;1300;1400;1520
5	295	6	1040;1090;1200;990;1460;1310
6	325	6	980;1010;870;1100;900;930

Estimativa do número de habitantes da localidade

$$y_{Acl}^{*Est} = \frac{M}{m} \sum_{h=1}^{L} \sum_{i=1}^{m_h} Y'_{hi} = 50(40730) = 2036500 \text{ habitantes.}$$

Estimativa da variância de y_{Acl}^{*Est}

Em cada estrato calcula-se $\bar{y}_{h.Acl} = \frac{1}{m_h} \sum_{i=1}^{m_h} Y'_{hi}$ média da amostra por Setor, no Estrato h, e

$s_{he}^2 = \frac{1}{m_h - 1} \sum_{i=1}^{m_h} (Y'_{hi} - \bar{y}_{h.Acl})^2$ variância da amostra entre os Setores.

Quadro 3

ESTRATOS	MÉDIA DA AMOSTRA POR SETOR($\bar{y}_{h\,.\,Acl}$)	VARIÂNCIA ENTRE OS SETORES (s^2_{he})
1	3 285	54 450
2	3 020	3 200
3	2 393	360 867
4	1 624	381 720
5	1 172	129 084
6	965	34 950

$$\hat{V}(y_{Acl}^{*Est}) = \frac{1-f}{f} \sum_{h=1}^{L} M_h s_{he}^2 = 49(64226395) = 3147093351$$

$$\sqrt{\hat{V}(y_{Acl}^{*Est})} = 56098{,}96 \qquad \hat{\gamma}(y_{Acl}^{*Est}) = \frac{56098{,}96}{2036500} = 0{,}0275$$

CAPÍTULO 4 — Amostragem de conglomerados em 2 — estágios — Ac2 — Tamanho desigual das Unidades Primárias. Probabilidade igual de seleção.

1 - INTRODUÇÃO

Mostrou-se que o efeito da conglomeração $1 + (\overline{N} - 1)\,\delta$ pode determinar uma redução na eficiência da Amostragem de Conglomerados em relação a Amostragem Aleatória Simples, desde que o coeficiente de correlação intraclasse seja positivo, conforme aliás, é comum acontecer. E essa redução é tanto maior, quanto maior for o tamanho dos conglomerados.

Para reduzir a influência do tamanho na eficiência, em vez de se considerar todo o conglomerado, conforme se faz com *1* — estágio, considera-se apenas uma amostra das unidades que compõem o conglomerado, ou seja, faz-se uma amostra de conglomerados com subamostragem.

Por exemplo, se os conglomerados de domicílios são representados por quarteirões, seleciona-se uma amostra de quarteirões e, nesses quarteirões da amostra, seleciona-se uma amostra de domicílios.

Esse desenho, constituído de uma amostra de conglomerados com subamostragem é chamado: Amostragem de Conglomerados em *2 — estágios.*

No que se segue, os conglomerados passam a ser chamados "unidades primárias — UP" e as unidades dentro dos conglomerados "unidades secundárias — US". No 1.º estágio há uma seleção de unidades primárias e no 2.º estágio uma seleção de unidades secundárias.

2 – CONFIGURAÇÃO DA AMOSTRA

As N unidades de π_N são grupadas em M unidades primárias, observando-se uma característica y.

UP_1		UP_M	
$US_{11}^{\rightarrow}$	Y_{11}	$US_{M1}^{\rightarrow}$	Y_{M1}
$US_{12}^{\rightarrow}$	Y_{12}	$US_{M2}^{\rightarrow}$	Y_{M2}
⋮		⋮	
$US_{1N_1}^{\rightarrow}$	Y_{1N_1}	$US_{MN_M}^{\rightarrow}$	Y_{MN_M}

Desse modo, na Unidade Primária i (UP_i) há N_i unidades secundárias (US_{i1}, US_{i2}, ... US_{iN_i})

Selecione-se uma ALS de m Unidades Primárias:

	UP'_1		UP'_m	
Amostra de 1.º estágio	$US'^{\rightarrow}_{11}$	Y'_{11}	$US'^{\rightarrow}_{m1}$	Y'_{m1}
	$US'^{\rightarrow}_{12}$	Y'_{12}	$US'^{\rightarrow}_{m2}$	Y'_{m2}
	⋮		⋮	
	$US'^{\rightarrow}_{1N'_1}$	$Y'_{1N'_1}$	$US'^{\rightarrow}_{mN'_m}$	$Y'_{mN'_m}$

Observe-se que UP'_i é um acontecimento aleatório, que, dependendo da unidade selecionada, poderá ser UP_1, UP_2, ..., UP_M.

Em cada UP' selecione-se uma ALS de Unidades Secundárias, obtendo-se:

	UP'_1		UP'_m	
Amostra de 2.º estágio	$US''^{\rightarrow}_{11}$	y_{11}	$US''^{\rightarrow}_{m1}$	y_{m1}
	$US''^{\rightarrow}_{12}$	y_{12}	$US''^{\rightarrow}_{m2}$	y_{m2}
	⋮		⋮	
	$US''^{\rightarrow}_{1n'_1}$	$y_{1n'_1}$	$US^{\rightarrow}_{mn'_m}$	$y_{mn'_m}$

A amostra de y é:

$$\{y_{11}, y_{12}, \ldots y_{1n'_1}; \ldots\ldots\ldots; y_{m1}, y_{m2}, \ldots y_{mn'_m}\}$$

Agora, em vez de se ter integralmente os conglomerados na amostra, com N'_1 N'_2, ... N'_m unidades, tem-se as subamostras, de tamanhos n'_1, n''_{2m}, ... n_m, respectivamente.

A Amostragem de Conglomerados em 2 — estágios é caracterizada pelo seguintes fatos:

a) A amostra total é constituída por um conjunto de amostras obtidas em conglomerados selecionados, em vez de ser constituída por todas as unidades dos conglomerados selecionados.

b) Há duas frações de amostragem: uma no 1.º estágio, correspondendo à seleção equiprovável das UP e representada por $f_1 = \frac{m}{M}$ e outra no 2.º estágio, representado por $f_{2i} = \frac{n_i}{N_i}$ $(i = 1, 2, \ldots M)$. Na situação usual, a probabilidade de 2.º estágio é mantida constante nas UP e representada por f_2.

c) O tamanho da amostra é:

$$\sum_{i=1}^{m} n'_i$$

e é uma variável aleatória, cujos valores dependem das UP selecionadas no 1.º estágio.

Em média, é igual a:

$$\bar{n} = E\left(\sum_{i=1}^{m} n'_i\right) = E\left(\sum_{i=1}^{m} f_2 N'_i\right) =$$

$$= f_2 m \frac{\sum_{i=1}^{M} N_i}{M} = f_1 f_2 N$$

d) No caso da fração de amostragem constante no 2.º estágio, a probabilidade de qualquer unidade de π_N pertencer à amostra é $f_1 f_2$.

3 – PARÂMETROS

Total de y em UP_i: $Y_i = \sum_{j=1}^{N_i} Y_{ij}$

Média de " " : $\overline{Y}_i = \frac{1}{N_i} Y_i$

Variância " " : $S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \overline{Y}_i)^2$

Total de y : $Y = \sum_{i=1}^{M} Y_i$

Média por US; $\overline{\overline{Y}} = \frac{Y}{N}$

Média por UP; $\overline{Y} = \frac{Y}{M}$

4 – ESTATÍSTICAS

Total de UP'_i: $Y'_i = \sum_{j=1}^{N'_i} Y'_{ij}$

Média " : $\overline{Y}'_i = \frac{Y'_i}{N'_i}$

Variância " : $S_i'^2 = \frac{1}{N'_i - 1} \sum_{j=1}^{N'_i} (Y'_{ij} - \overline{Y}'_i)^2$

Total da subamostra em UP'_i: $y_i = \sum_{j=1}^{n'_i} y_{ij}$

Média da subamostra em UP'_i: $\overline{y}_i = \frac{y_i}{n'_i}$

Variância da subamostra em UP'_i: $s_i^2 = \frac{1}{n'_i - 1} \sum_{j=1}^{n'_i} (y_{ij} - \overline{y}_i)^2$

5 – TEOREMA

Um estimador não tendencioso de Y é:

$$\overset{*}{y}_{Ac2} = \frac{M}{m} \sum_{i=1}^{m} N'_i \, \overline{y}_i$$

Prova

Observe-se que o valor de $\overline{y}_i$ depende de qual unidade primária tenha sido selecionada. Desse modo, é necessário achar a média condicional – primeiro fixando a unidade primária e depois fazendo a variação das unidades primárias.

$$E(\overset{*}{y}_{Ac2}) = \underset{UP'_i}{E}\left[E\left(\frac{M}{m} \sum_{i=1}^{m} N'_i \, \overline{y}_i \mid UP'_i \text{ fixada} \right) \right] =$$

$$= \underset{UP'_i}{E}\left[\frac{M}{m} \sum_{i=1}^{m} N'_i \, \overline{Y}'_i \right] =$$

$$= \frac{M}{m} \sum_{i=1}^{m} \left(\frac{1}{M} \sum_{i=1}^{M} N_i \, \overline{Y}_i \right) =$$

$$= \sum_{i=1}^{M} Y_i = Y$$

5.1 – *Corolário*

Um estimador não tendencioso de $\overline{\overline{Y}}$ é:

$$\overline{\overline{y}}_{Ac2} = \frac{1}{m\overline{N}} \sum_{i=1}^{m} N'_i \, \overline{y}_i$$

Prova

$$\overline{\overline{y}}_{Ac2} = \frac{\overset{*}{y}_{Ac2}}{N}$$

$$\text{donde} \quad E(\overline{\overline{y}}_{Ac2}) = \frac{Y}{N} = \overline{\overline{Y}}$$

5.2 – *Corolário*

Um estimador não tendencioso de $\overline{Y}$ é:

$$\overline{y}_{Ac2} = \frac{1}{m} \sum_{i=1}^{m} N'_i \, \overline{y}_i$$

Prova

Imediata

6 – VARIÂNCIA DE y^*_{Ac2}

Ainda em decorrência do fato de que o valor de $\bar{y}_I$ depende da unidade primária selecionada, é necessário achar a variância condicional. A fórmula é constituída de duas partes. Na primeira parte, toma-se a média fixando a unidade primária e depois toma-se a variância deixando a unidade primária variar. Na segunda parte, troca-se esta ordem de operação.

$$V(y^*_{Ac2}) = \underset{UP'_i}{V}\left[E\left(\frac{M}{m}\sum_{i=1}^{m} N'_i\,\bar{y}_i \mid UP'_i \text{ fixado}\right)\right] + \\ + \underset{UP'_i}{E}\left[V\left(\frac{M}{m}\sum_{i=1}^{m} N'_i\,\bar{y}_i \mid UP'_i \text{ fixado}\right)\right]$$

Tem-se, para as parcelas do segundo membro:

a) $$\underset{UP'_i}{v}\left[E\left(\frac{M}{m}\sum_{i=1}^{m} N'_i\,\bar{y}_i \mid UP'_i \text{ fixado}\right)\right] =$$

$$\underset{UP'_i}{V}\left[\frac{M}{m}\sum_{i=1}^{m} N'_i = \bar{Y}'_i\right] = \underset{UP'_i}{V}\left[\frac{M}{m}\sum_{i=1}^{m} Y'_i\right]$$

Observe-se que esta variância corresponde à variância de y^*_{Ac1} e que é:

$$M^2 \cdot \frac{M-m}{M}\,\frac{S_e^2}{m} \quad \text{com} \quad S_e^2 = \frac{1}{M-1}\sum_{i=1}^{M}(Y_i - \bar{Y})^2$$

b) $$\underset{UP'_i}{E}\left[V\left(\frac{M}{m}\sum_{i=1}^{m} N'_i\,\bar{y}_i \mid UP'_i \text{ fixado}\right)\right] =$$

$$= \underset{UP'_i}{E}\left[\frac{M^2}{m^2}\sum_{i=1}^{m} N'^2 \cdot V(\bar{y}_i \mid UP'_i \text{ fixado}).\right] =$$

$$= \underset{UP'_i}{E}\left[\frac{M^2}{m^2}\sum_{i=1}^{m} N'^2_i \cdot \frac{N'_i - n'_i}{N'_i}\,\frac{S'^2_i}{n'_i}\right] =$$

$$= \frac{M}{m}\sum_{i=1}^{M} N_i^2 \cdot \frac{N_i - n_i}{N_i}\,\frac{S_i^2}{n_i}$$

Substituindo os resultados de a) e b) em $V(y^*_{Ac2})$ vem:

$$V(y^*_{Ac2}) = M^2 \frac{M-m}{M} \frac{S_e^2}{m} + \frac{M}{m} \sum_{i=1}^{M} N_i^2 \cdot \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i}$$

As duas parcelas do 2.º membro são chamadas "componentes da variância."

Se $m = M$ (1.ª componente nula)

$$V(y^*_{Ac2}) = \sum_{i=1}^{M} N_i^2 \cdot \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i}$$ que corresponde a $V(y^*_{Est})$

variância do estimador de Y na amostragem estratificada.

Se $n_i = N_i$ $(i = 1, 2, \ldots, M)$ (2.ª componente nula)

$V(y^*_{Ac2}) = M^2 \frac{M-m}{M} \frac{S_e^2}{m}$, variância do estimador de Y *com*

1 — *estágio.*

6.1 —.*Exercício*

Mostrar que:

a) $V(\overline{y}_{Ac2}) = \frac{M-m}{M} \cdot \frac{S_e^2}{m} + \frac{1}{Mm} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$

b) $V(\overline{\overline{y}}_{Ac2}) = \frac{M-m}{M} \cdot \frac{\overline{S}_e^2}{m} + \frac{1}{Mm} \sum_{i=1}^{M} \left(\frac{N_i}{\overline{N}}\right)^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$

7 — TEOREMA

Um estimador não tendencioso de $V(y^*_{Ac2})$ é:

$$\hat{V}(y^*_{Ac2}) = M^2 \frac{M}{M-m} \frac{s_e^2}{m} + \frac{M}{m} \sum_{i=1}^{m} N_i'^2 \frac{N_i'}{N_i' - n_i'} \cdot \frac{s_i^2}{n_i'}$$

$$\text{onde } s_e^2 = \frac{1}{m-1} \sum_{i=1}^{m} (N_i' \overline{y}_i - \overline{y}_{Ac2})^2$$

Prova

Trata-se de mostrar que $E[\hat{V}(\bar{y}^*_{Ac2})] = V(\bar{y}^*_{Ac2})$

a) $$E(s_e^2) = \frac{1}{m-1} E\left[\sum_{i=1}^{m} (N'_i\,\bar{y}_i - \bar{y}_{Ac2})^2\right] =$$

$$= \frac{1}{m-1} E\left[\sum_{i=1}^{m} (N'_i\,\bar{y}_i)^2\right] - \frac{m}{m-1} E[\bar{y}^2_{Ac2}]$$

a.1) Para a primeira expectância no segundo membro, tem-se:

$$E\left[\sum_{i=1}^{m} (N'_i\,\bar{y}_i)^2\right] = \underset{UP'_i}{E}\left\{E\left[\sum_{i=1}^{m} (N'_i\,\bar{y}_i)^2 \mid UP'_i \text{ fix.}\right]\right\} =$$

$$= \underset{UP'_i}{E}\left[\sum_{i=1}^{m} V(N'_i\,\bar{y}_i \mid UP'_i \text{ fix.})\right] + \underset{UP'_i}{E}\left\{\sum_{i=1}^{m} [(N'_i\,\bar{y}_i \mid UP'_i \text{ fix.})]\right\} =$$

$$= \underset{UP'_i}{E}\left[\sum_{i=1}^{m} N'^2_i \cdot \frac{N'_i - n'_i}{N'_i} \frac{S'^2_i}{n_i}\right] + \underset{UP'_i}{E}\left[\sum_{i=1}^{m} (N'_i\,\bar{Y}'_i)^2\right] =$$

$$= \frac{m}{M} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} + \frac{m}{M} \sum_{i=1}^{M} Y_i^2$$

a.2) Para a segunda expectância no segundo membro, tem-se:

$$E(\bar{y}_{Ac2}) = V(\bar{y}_{Ac2}) + [E(\bar{y}_{Ac2})]^2 =$$

$$= \frac{M-m}{M} \frac{S_e^2}{m} + \frac{1}{Mm} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} + \bar{Y}^2$$

Substituindo em a) os resultados obtidos em a.1) e a.2) vem:

$$E(s_e^2) = \frac{1}{m-1}\left[\frac{m}{M} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} + \frac{m}{M} \sum_{i=1}^{M} Y_i^2 -\right.$$

$$\left. - \frac{M-m}{M} S_e^2 - \frac{1}{M} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} - m\,\bar{Y}^2\right] =$$

$$= \frac{1}{m-1}\left[\left(\frac{M}{m} - \frac{1}{M}\right) \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} +\right.$$

$$\left. + \frac{m}{M} \sum_{i=1}^{M} Y_i^2 - \frac{M-m}{M} S_e^2 - m\bar{Y}^2\right] =$$

$$= \frac{1}{m-1}\left[\frac{m-1}{M}\sum_{i=1}^{M} N_i^2 \frac{M_i - n_i}{N_i}\frac{S_i^2}{n_i} + \right.$$

$$\left. + \frac{m}{M}\sum_{i=1}^{M}(Y_i - \overline{Y})^2 - \frac{M-m}{M}S^2\right] =$$

$$= \frac{1}{m-1}\left[\frac{m-1}{M}\sum_{i=1}^{M} N_i^2 \cdot \frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i} + \frac{m(M-1)}{M}S_e^2 - \right.$$

$$\left. - \frac{M-m}{M}S_e^2\right] =$$

$$= \frac{1}{m-1}\left[\frac{m-1}{M}\sum_{i=1}^{M} N_i^2 \cdot \frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i} + (m-1)S_e^2\right]$$

donde,

$$E(s_e^2) = S_e^2 + \frac{1}{M}\sum_{i=1}^{M} N_i^2 \cdot \frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i} \qquad (7.I)$$

Esta expressão (7.1) mostra que s_e^2 é estimador tendencioso de S_e^2.

b) $$E\left[\sum_{i=1}^{m} N_i'^2 \frac{N_i - n_i}{N_i'}\frac{s_i^2}{n_i'}\right] =$$

$$= \underset{UP_i'}{E}\left\{E\left[\sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'}\frac{s_i^2}{n_i'} \mid UP_i' \text{ fixado}\right]\right\} =$$

$$= \underset{UP_i'}{E}\left[\sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'}\frac{S_i'^2}{n_i'}\right] =$$

$$= \frac{M}{m}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i}$$

Então, substituindo os resultados de a) e b) em:

$$E[\hat{V}(\bar{y}_{Ac2}^*)] = M^2 \frac{M-m}{M}\frac{E(s_e^2)}{m} + \frac{M}{m}E\left[\sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'}\frac{s_i^2}{n_i'}\right]$$

vem,

$$E[V(y^*_{Ac2})] = M^2 \frac{M-m}{M} \frac{1}{m}\left[S_e^2 + \frac{1}{M}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i}\right] +$$

$$+ \frac{M}{m}\left[\frac{m}{M}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i}\right] =$$

$$= M^2 \cdot \frac{M-m}{M} \frac{S_e^2}{m} + \left(M\frac{M-m}{Mm} + 1\right)\sum_{i=1}^{M} N_i^2 \cdot \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} =$$

$$= M^2 \cdot \frac{M-m}{M} \frac{S_e^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} = V(y^*_{Ac2})$$

7.1 — *Estimador de* $V(y^*_{Ac2})$ *sem desmembramento nas componentes da variância.*

Mostrou-se que:

$$E(s_e^2) = S_e^2 + \frac{1}{M}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

Multiplicando por $\frac{M^2}{m}$ vem:

$$E\left(\frac{M^2 s_e^2}{m}\right) = \frac{M^2 S_e^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

Por outro lado, para $M >> m$, $\left(\frac{m}{M} < 1\%\right)$, tem-se:

$$V(y^*_{Ac2}) \doteq M^2 \cdot \frac{S_e^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

Comparando as duas expressões acima, tem-se:

$$V(y^*_{Ac2}) \doteq E\left(\frac{M^2 s_e^2}{m}\right)$$

donde se obtém o estimador:

$$\hat{V}(y^*_{Ac2}) \doteq \frac{M^2 s_e^2}{m}$$

7.1.1 – Exercício

Achar as expressões aproximadas de $V(\bar{\bar{y}}_{Ac2})$ e de $\hat{V}(\bar{y}_{Ac2})$

8 – COMPONENTES DA VARIÂNCIA

As duas parcelas que compõem $V(y^*_{Ac2})$ são chamadas "componentes da variância":

$$M^2 \frac{M-m}{M} \cdot \frac{S_e^2}{m}$$

e

$$\frac{M}{m} \sum_{i=1}^{m} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

No que se segue, estimam-se essas duas componentes.

a) *Estimador da componente "entre"*

Sabe-se que:

$$E(s_e^2) = S_e^2 + \frac{1}{M} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

Um estimador não tendencioso da segunda parcela do segundo membro é:

$$\frac{1}{m} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \cdot \frac{S_i'^2}{n_i'}$$

donde se pode escrever,

$$E(s_e^2) = S_e^2 + E\left[\frac{1}{m} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \frac{s_i^2}{n_i'}\right]$$

donde,

$$s_e^2 = \hat{S}_e^2 + \frac{1}{m} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \cdot \frac{s_i^2}{n_i'}$$

representando por $\hat{S}_e^2$ o estimador não tendencioso de S_e^2.

Então,

$$\hat{S}_e^2 = s_e^2 - \frac{1}{m} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \cdot \frac{s_i^2}{n_i'}$$

Substituindo S_e^2 por $\hat{S}_e^2$ na componente entre, vem:

Estimador não tendencioso de $M^2 \frac{M-m}{M} \cdot \frac{S_e^2}{m} = M^2 \frac{M-m}{M} \cdot \frac{\hat{S}_e}{m}$

b) *Estimador da componente "dentro"*

É imediata a expressão:

Estimador não tendencioso de $\frac{M}{m} \sum_{i=1}^{m} N_i^2 \frac{N_i - n_i}{n_i} \frac{S_i^2}{n_i} =$

$$= \frac{M^2}{m^2} \sum_{i=1}^{m} N_i' \frac{N_i' - n_i'}{N_i'} \frac{s_i^2}{n_i'}$$

9 – AMOSTRA AUTOPONDERADA

A probabilidade de uma *US* pertencer à amostra é:

$$\frac{m}{M} \frac{n_i}{N_i}$$

isto é, é o produto da probabilidade de seleção da *UP* pela probabilidade de seleção da *US* na subamostra da *UP*.

Fixada uma fração de amostragem geral $\frac{n}{N}$, diz-se que amostra é autoponderada se:

$$\frac{m}{M} \cdot \frac{n_i}{N_i} = \frac{n}{N} \quad (i = 1, 2, \ldots M) \tag{9.I}$$

donde,

$\frac{n_i}{N_i} = \frac{Mn}{mN} = \frac{\bar{n}}{\bar{N}}$ com $\bar{n} = \frac{n}{M}$, tamanho médio da subamostra e $\bar{N} = \frac{N}{M}$ tamanho médio do conglomerado. Desse modo, a fração de amostragem de 2.º estágio se tornou constante.

Sendo $f_1 = \frac{m}{M}$ a fração de amostragem de 1.º estágio, $f_2 = \frac{\bar{n}}{\bar{N}}$ a fração de amostragem de 2.º estágio e $f = \frac{n}{N}$ a fração geral de amostragem, pode-se escrever (9.I) na forma:

$$f_1 f_2 = f$$

9.1 — *Adequação da expressão de* y^*_{ac2}

Da expressão geral,

$$y^*_{Ac2} = \frac{M}{m} \sum_{i=1}^{m} \frac{N'_i}{n'_i} \sum_{j=1}^{n'_i} y_{ij}$$

obtém-se:

$$y^*_{Ac2} = \frac{M \cdot \overline{N}}{m \cdot \overline{n}} \sum_{i=1}^{m} \sum_{j=1}^{n'_i} y_{ij} = \frac{N}{n} \sum_{i=1}^{m} \sum_{j=1}^{n'_i} y_{ij}$$

ou ainda,

$$y^*_{Ac2} = \frac{1}{f} \sum_{i=1}^{m} \sum_{j=1}^{n'_i} y_{ij}$$

9.2 — *Adequação da expressão de* $V(y^*_{Ac2})$

Da expressão,

$$V(y^*_{Ac2}) = M^2 \frac{M-m}{M} \frac{S_e^2}{m} + \frac{M}{m} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

obtém-se:

$$V(y^*_{Ac2}) = M^2 \frac{M-m}{M} \frac{S_e^2}{m} + \frac{M\overline{N}}{m\overline{n}} \frac{\overline{N} - \overline{n}}{\overline{N}} \sum_{i=1}^{M} N_i S_i^2$$

$$\text{pondo} \quad S_d^2 = \frac{1}{M\overline{N}} \sum_{i=1}^{M} N_i S_i^2$$

vem,

$$V(y^*_{Ac2}) = M^2 \frac{M-m}{M} \frac{S_e^2}{m} + (M\overline{N})^2 \frac{\overline{N} - \overline{n}}{\overline{N}} \cdot \frac{S_d^2}{m\overline{n}}$$

Em termos das frações de amostragem, pode-se escrever:

$$V(y^*_{Ac2}) = M \frac{1 - f_1}{f_1} S_e^2 + \frac{1 - f_2}{f_1 f_2} S_d^2$$

9.3 — *Adequação da expressão de* $\hat{V}(y^*_{Ac2})$

Pondo $s_d^2 = \frac{1}{m\overline{N}} \sum_{i=1}^{m} N'_i \, s_i^2$, estimador não tendencioso de S_d^2, obtém-se:

$$\hat{V}(y^*_{Ac2}) = M^2 \frac{M-m}{M} \frac{s_2^2}{m} + (M\overline{N})^2 \frac{\overline{N}-\overline{n}}{\overline{N}} \frac{s_d^2}{m\overline{n}}$$

ou em termos das frações de amostragem,

$$\hat{V}(y^*_{Ac2}) = M \cdot \frac{1-f_1}{f_1} s_e^2 + \frac{1-f_2}{f_1 f_2} s_d^2$$

Para o estimador aproximado,

$$\hat{V}(y^*_{Ac2}) = \frac{M^2 s_e^2}{m}$$

tem-se, com amostra autoponderada.

$$\hat{V}(y^*_{Ac2}) \doteq m \left(\frac{N}{n}\right)^2 \frac{1}{m-1} \sum_{i=1}^{m} \left(\sum_{j=1}^{n'_i} y_{ij} - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n'_i} y_{ij} \right)^2$$

que só depende das amostras finais de cada conglomerado.

9.4 — *Exercício*

Com amostra autoponderada, achar as expressões de $\overline{y}_{Ac2}$, $\overline{\overline{y}}_{Ac2}$ e respectivas variâncias e estimadores das variâncias.

9.5 — *Exemplos*

Em determinada Região, de acordo com o último Censo, há 150 Setores Censitários e, aproximadamente, 36.400 domicílios. Há condições financeiras e de tempo para selecionar uma amostra de 364 domicílios, com a finalidade de estimar o número de habitantes da região, o que corresponde a uma fração geral de amostragem,

$$f = \frac{364}{36\,400} = \frac{1}{100}$$

Há, em média, $\frac{36.400}{150} \doteq 243$ domicílios por Setor Censitário.

Serão selecionados 10 setores, com probabilidade igual de seleção, o que corresponde a uma fração de amostragem de 1.º estágio

$$f_1 = \frac{10}{150} = \frac{1}{15}$$

Da relação: $f = f_1 f_2$ (amostra autoponderada), obtém-se a fração de amostragem de 2.º estágio,

$$f_2 = f \div f_1 = \frac{15}{100}$$

Dados da amostra

SETORES CENSITÁRIOS	N.º DE DOMICÍLIOS DO SETOR (N_i')	N.º DE DOMICÍLIOS NA SUBAMOSTRA $n_i' = f_2 N_i'$	N.º DE MORADORES NA SUBAMOSTRA $\sum_{j=1}^{n_i'} y_{ij}$	s_i^2
1	320	48	168	4,018
2	210	32	138	5,224
3	180	27	130	5,905
4	400	60	222	1,044
5	250	38	201	2,840
6	221	33	149	4,345
7	120	18	97	6,000
8	500	75	300	2,012
9	262	39	199	3,484
10	238	36	108	3,000
	2 701	406	1 712	—

Estimativa do número de habitantes:

$$\overset{*}{y}_{Ac2} = \frac{1}{f} \sum_{i=1}^{m} \sum_{j=1}^{n_i'} y_{ij} = 100(1712) = 171200 \text{ habitantes}$$

Estimativa da variância aproximada de y^*_{Ac2}

$$\hat{V}(y^*_{Ac2}) = m\left(\frac{N}{n}\right)^2 \frac{1}{m-1}\left(\sum_{j=1}^{n'_i} y_{ij} - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n'_i} y_{ij}\right)^2 =$$

$$= 10(100)^2 \frac{1}{9}(33633{,}6) = 373332960$$

$$\sqrt{\hat{V}(y^*_{Ac2})} = 19321{,}825$$

$$\hat{\gamma}(y^*_{Ac2}) = 0{,}113 \quad \text{ou} \quad 11{,}3\%$$

10 — DIMENSIONAMENTO DA AMOSTRA

Na amostragem de conglomerados em 1 — estágio, dimensionou-se a amostra de conglomerados fixando a variância do estimador, isto é, fixou-se d e, conseqüentemente, fixou-se $V(\bar{y}_{Ac1})$ em $\frac{d^2}{z^2_\theta}$.

Na *Ac2* há que se achar dois valores: o número de unidades primárias da amostra e o número de unidades secundárias da subamostra. O problema não pode ser resolvido apenas fixando a variância, que depende desses dois valores.

Considera-se, então, uma função auxiliar, referente ao custo da execução do desenho e que também depende do número de *UP* e do número de *US*.

No que se segue, para possibilitar uma solução simples para o problema acima abordado, considera-se um tamanho médio $\overline{N}$ das unidades primárias e um tamanho médio $\overline{n}$ da subamostra.

Dois critérios para o cálculo de m — número de *UP* da amostra — e de $\overline{n}$ — número de *US* da subamostra, podem ser estabelecidos:

- Minimizar a variância com um custo fixado
- Minimizar o custo com variância fixada.

10.1 — *Função custo*

Para ambos os critérios propostos, necessita-se de uma função custo. Uma definição da função custo é a seguinte:

$$C_t = C_f + C_1 m + C_2 m\overline{n}$$

onde:

C_f: custo fixo — Despesas que não variam com o processo de seleção e com o tamanho da amostra.

Incluem-se nessas despesas:

a) planejamento e orientação, incluindo salários do pessoal técnico e despesas de administração.

b) preparação de mapas e outras informações que não dependem do tamanho da amostra.

c) impressão de tabelas e treinamento de pessoal de campo que não dependa do tamanho da amostra.

C_1m: — Despesas que variam com o número de unidades primárias da amostra. C_1 é o custo por unidade primária e é formado das seguintes parcelas:

a) despesas de seleção da amostra de unidades primárias.

b) preparação de roteiros de viagem para as unidades primárias.

c) impressão de material para a amostra de unidades primárias.

d) tempo de treinamento para investigação das unidades primárias.

e) gastos de transporte para as unidades primárias e entre as mesmas.

$C_2m\overline{n}$: — Despesas que variam com o número de unidades secundárias. C_2 é o custo por unidade e é formado das seguintes parcelas:

a) custo de entrevista de cada unidade.

b) impressão de material referente as unidades da amostra.

c) despesas de transporte dentro das unidades primárias.

10.2 — *Minimizar a variância com custo fixado*

Para obter o m_o e $\overline{n}_o$, valores de m e $\overline{n}$ respectivamente, que minimizam a variância $V(\overline{\overline{y}}_{Ac2})$ com custo fixado, minimiza-se a função:

$$F = V(\overline{\overline{y}}_{Ac2}) + \lambda[C_1 m + C_2 m\overline{n} - C]$$

$$\text{onde} \quad C = C_t - C_f$$

e λ é um multiplicador de Lagrange.

Pode-se escrever:

$$F = \frac{M - m}{M} \frac{\overline{S}_e^2}{m} + \frac{\overline{N} - \overline{n}}{\overline{N}} \frac{S_d^2}{m\overline{n}} + \lambda[C_1 m + C_2 m\overline{n} - C]$$

Tomando as derivadas parciais em relação a m e a n e igualando a zero, vem:

$$\frac{\partial F}{\partial \overline{n}} = -\frac{S_d^2}{m\overline{n}^2} + \lambda C_2 m = 0 \qquad (10.I)$$

$$\frac{\partial F}{\partial m} = -\frac{\overline{S}_e^2}{m^2} - \frac{\overline{N} - \overline{n}}{\overline{N}} \frac{S_d^2}{m^2 \overline{n}} + \lambda[C_1 + C_2 \overline{n}] = 0 \qquad (10.II)$$

De (10.I) obtém-se:

$$\lambda C^2 m^2 \overline{n}^2 = S_d^2$$

De (10.II) obtém-se:

$$\lambda[C_1 + C_2 \overline{n}] m^2 \overline{N}\overline{n} = \overline{S}_e^2 \overline{N} \overline{n} + (\overline{N} - \overline{n}) S_d^2$$

Dividindo membro a membro as duas últimas igualdades, vem:

$$\frac{(C_1 + C_2 \overline{n}) \overline{N}}{C_2 \overline{n}} = \frac{\overline{S}_e^2}{S_d^2} \cdot \overline{N} \overline{n} + (\overline{N} - \overline{n})$$

ou,

$$(C_1 + C_2 \overline{n}) \overline{N} S_d^2 = \overline{S}_e^2 \cdot C_2 \overline{N} \overline{n}^2 + (\overline{N} - \overline{n}) C_2 \overline{n} S_d^2$$

ou,

$$(C_1 + C_2 \overline{n}) \overline{N} S_d^2 - (\overline{N} - \overline{n}) C_2 \overline{n} S_d^2 = \overline{S}_e^2 \cdot C \overline{N} \overline{n}^2$$

donde,

$$C_1 \overline{N} S_d^2 = (\overline{S}_e^2 \cdot C_2 \overline{N} - C_2 S_d^2) \overline{n}^2$$

A solução para n é o ótimo $\overline{n}_o$:

$$\overline{n}_0 = \sqrt{\frac{C_1}{C_2} \cdot \frac{S_d^2}{\overline{S}_e^2 - \frac{S_d^2}{\overline{N}}}} \qquad (10.III)$$

Substituindo $\overline{n}_o$ no lugar de $\overline{n}$ em $m = \dfrac{C}{C_1 + C_2 \overline{n}}$ obtém-se o ótimo m_o.

Observe-se que:

a) $\overline{n}_o$ cresce se C_1 cresce em relação a C_2, ou seja, se cresce a parte do custo referente as unidades primárias, cabe aumentar $\overline{n}_o$ e diminuir m_o.

b) para achar $\overline{n}_o$ basta se ter uma informação sobre a razão $\dfrac{C_1}{C_2}$. Pequenas variações sobre esse valor, repercutem pouco sobre $\overline{n}_o$ posto que se toma a raiz quadrada. Em levantamentos de âmbito nacional nos Estados Unidos, $\dfrac{C_1}{C_2}$ varia de 25 a 50. Em levantamentos de âmbito regional C_1 baixa e $\dfrac{C_1}{C_2}$ varia de 15 a 20. Em pesquisas contínuas, com rotação de amostras, C_1 pode ser repartido ao longo do tempo, baixando $\dfrac{C_1}{C_2}$ para 4.

$\overline{n}_o$ pode ser estimado a partir da expressão (10.III) observando-se que:

$$\overline{S}_e^2 - \frac{S_d^2}{\overline{N}} = E\left(\hat{\overline{S}}_e^2 - \frac{s_d^2}{\overline{N}}\right) =$$

$$= E\left(\overline{s}_e^2 - \frac{\overline{N} - \overline{n}}{\overline{N}} \frac{s_d^2}{\overline{n}} - \frac{s_d^2}{\overline{N}}\right) =$$

$$= E\left(\overline{s}_e^2 - \frac{s_d^2}{\overline{n}}\right)$$

Desse modo, $\overline{n}_o$ pode ser estimado por:

$$\hat{\overline{n}}_0 = \sqrt{\frac{C_1}{C_2} \cdot \frac{s_d^2}{\overline{s}_e^2 - \frac{s_d^2}{\overline{n}}}} \qquad (10.IV)$$

Nesta expressão, supõe-se $\left[\overline{s}_e^2 - \frac{s_d^2}{\overline{n}}\right] > 0$

Se isso não acontece, $\overline{n}_o$ é obtido de outro modo.

Considere-se a função custo:

$$C = m\,(C_1 + C_2\,\overline{n})$$

Se $C > C_1 + C_2\,\overline{N}$ então

$$\overline{n}_0 = \text{máximo de } \overline{n} = \overline{N}$$

$$\text{donde, } \quad m_0 = \frac{C}{C_1 + C_2\,\overline{N}}$$

Se $C \leqslant C_1 + C_2\overline{N}$ então, $\overline{n}_o$ é a solução para $\overline{n}$ na equação $C = C_1 + C_2\overline{n}$ donde $\overline{n}_o = \dfrac{C - C_1}{C}$.

$$\text{e} \quad m_0 = 1$$

10.3 — *Minimizar o custo com variância fixada*

Agora, a função a minimizar é:

$$G = C + \mu\ V(\overline{\overline{y}}_{Ac2})$$

onde μ é um multiplicador de Lagrande.

Pode-se escrever:

$$G = (C_1\,m + C_2\,m\,\overline{n}) + \mu\left(\frac{M - m}{M}\ \frac{\overline{S}_e^2}{m} + \frac{\overline{N} - \overline{n}}{\overline{N}}\ \frac{S_d^2}{m\overline{n}}\right)$$

donde,

$$\frac{\partial\,G}{\partial\,\overline{n}} = C_2\,m - \mu\,\frac{S_d^2}{m\overline{n}^2} = 0$$

$$\frac{\partial\,G}{\partial\,\cdot\,m} = C_1 + C_2\,n - \mu\left(\frac{\overline{S}_e^2}{m^2} + \frac{\overline{N} - \overline{n}}{\overline{N}}\ \frac{S_d^2}{\overline{n}}\right) = 0$$

Observe-se que essas duas equações correspondem às equações (10.I) e (10.II) com $\lambda = \frac{1}{\mu}$.

Conseqüentemente, tem-se a mesma solução para o ótimo de $\overline{n}$, seja minimizando a variância com custo fixado, seja minimizando o custo, com variância fixada.

Quanto a m_o, é obtido fixando $V(\overline{\overline{y}}_{Acz})$ com $\overline{n}_o$ no lugar de $\overline{n}$ e tirando o valor de m.

10.4 — *Expressão de $\overline{n}_o$ em função do coeficiente de correlação intraclasse*

Considerem-se as seguintes expressões, já encontradas no caso de 1 — estágio (Capítulo 1):

$$\delta = \frac{\frac{M-1}{M}\,\overline{S}_e^2 - \frac{S_d^2}{\overline{N}}}{\frac{M\overline{N}-1}{M\overline{N}}\,S^2} \qquad (10.V)$$

$$(M\overline{N}-1)\,S^2 = M\,(\overline{N}-1)\,S_d^2 + (M-1)\,\overline{N}\,\overline{S}_e^2 \qquad (10.VI)$$

Substituindo $(MN\text{-}1)\,S^2$ tirado de (10.VI) em (10.V), obtém-se:

$$\delta = \frac{\frac{M-1}{M}\,\overline{S}_e^2 - \frac{S_d^2}{\overline{N}}}{\frac{\overline{N}-1}{\overline{N}}\,S_d^2 + \frac{M-1}{M}\,\overline{S}_e^2}$$

donde,

$$I - \delta = \frac{S_d^2}{\frac{\overline{N}-1}{\overline{N}}\,S_d^2 + \frac{M-1}{M}\,\overline{S}_e^2}$$

$$\frac{1-\delta}{\delta} = \frac{S_d^2}{\frac{M-1}{M}\,\overline{S}_e^2 - \frac{S_d^2}{\overline{N}}} \doteq \frac{S_d^2}{\overline{S}_e^2 - \frac{S_d^2}{\overline{N}}}$$

Então, substituindo em (10.III) obtém-se:

$$\bar{n}_0 = \sqrt{\frac{C_1}{C_2}\,\frac{1-\delta}{\delta}}$$

A necessidade do conhecimento do coeficiente de correlação já foi enfatizada no estudo da *AC-1*. Ratifica-se, agora, esta necessidade, para o cálculo de $\bar{n}_0$.

11 – EXEMPLO

Em certa localidade existem 740 setores censitários. Trata-se de estimar a produção total dos estabelecimentos agrícolas, produtores de café, com uma amostra de conglomerados em 2 – estágios, sendo os Setores as unidades primárias e os estabelecimentos as unidades secundárias.

De pesquisa anterior, sabe-se que, para a característica e os setores em questão:

$$\delta = 0{,}201$$

$$\frac{C_1}{C_2} = 10$$

Conseqüentemente, o tamanho médio da subamostra é:

$$\bar{n}_0 = \sqrt{10\,\frac{1-0{,}201}{0{,}201}} \doteq 6$$

O custo de investigação de um estabelecimento foi estimado em 300, de modo que a função custo é da forma:

$$C = 3000\,m + 300\,m\bar{n}$$

A quantia disponível para a pesquisa é 300.000, donde:

$$m = \frac{300000}{3000 + 3000\,(6)} = 62 \text{ setores.}$$

correspondendo a $6\,(62) = 372$ estabelecimentos.

A fração de amostragem de 1.º estágio é:

$$f_1 = \frac{m}{M} = \frac{62}{740} \doteq \frac{1}{12}$$

Os setores têm, em média, $\overline{N} = 30$ estabelecimentos, de modo que a fração de amostragem de 2.º estágio é:

$$f_2 = \frac{\overline{n}}{\overline{N}} = \frac{6}{30} = \frac{1}{5}$$

Portanto, a fração geral de amostragem é:

$$f = f_1 \cdot f_2 = \frac{1}{60}$$

Feita a seleção das 62 unidades primárias (Setores) aplicou-se em cada uma a fração de amostragem de 2.º estágio, obtendo-se o seguinte quadro:

SETORES DA AMOSTRA	N.º DE ESTABELECIMENTOS (N_i')	TAMANHO DA SUBAMOSTRA (n_i')	PRODUÇÃO TOTAL DA SUBAMOSTRA $\sum_{j=1}^{n'} y_{ij}$
1	58	12	1 300
2	35	7	980
.	.	.	.
.	.	.	.
.	.	.	.
62	47	9	640
Soma	1 920	412	57 320

Estimativa da produção total

$$y^*_{Ac2} = \frac{1}{f} \sum_{i=1}^{m} \sum_{j=1}^{n_i'} y_{ij} = 60(57320) = 3439200$$

Estimativa da variância de $\bar{y}^*_{Ac2}$

$$\hat{V}(\bar{y}^*_{Ac2}) = \frac{m}{m-1}\left(\frac{1}{f}\right)^2 \sum_{i=1}^{m}\left(\sum_{j=1}^{n'_i} y_{ij} - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n'_i} y_{ij}\right)^2 =$$

$$= \frac{62}{61}\,(60)^2\,(4650307{,}73)$$

$$\sqrt{\hat{V}(\bar{y}^*_{Ac2})} = 130443{,}67$$

$$\hat{\gamma}(\bar{y}^*_{Ac2}) = 0{,}038$$

12 – EFEITO DA CONGLOMERAÇÃO

Recorde-se, ainda, da Amostragem de Conglomerados em 1 – estágio (Capítulo 1), as seguintes expressões:

a) $(M\overline{N} - 1)\,S^2 = M\,(\overline{N} - 1)\,S_d^2 + (M - 1)\,\overline{N}\,\overline{S}_e^2$

b) $\overline{S}_e^2 = \dfrac{M\overline{N} - 1}{(M-1)\,\overline{N}}\,\dfrac{S^2}{\overline{N}}\,[1 + (\overline{N} - 1)\,\delta]$

Substituindo b) em a) obtém-se:

$$(M\overline{N}-1)\,S^2 = M\,(\overline{N}-1)\,S_d^2 + (M-1)\,\overline{N}\left\{\frac{M\overline{N}-1}{(M-1)\,\overline{N}}\,\frac{S^2}{\overline{N}}\,[1+(\overline{N}-1)\delta]\right\}$$

donde,

$$(M\overline{N} - 1)\,S^2 - (M\overline{N} - 1)\,\frac{S^2}{\overline{N}}\,[1 + (\overline{N} - 1)\,\delta] = M\,(\overline{N} - 1)\,S_d^2$$

ou,

$$(M\overline{N} - 1)\,(\overline{N} - 1)\frac{S^2}{\overline{N}}\,(1 - \delta) = M\,(\overline{N} - 1)\,S_d^2$$

donde,

$$S_d^2 = \frac{M\overline{N} - 1}{M\overline{N}} S^2 (1 - \delta)$$

Considerando as seguintes aproximações:

$$\frac{M - 1}{M} \doteq 1 \quad \text{e} \quad \frac{M\overline{N} - 1}{M\overline{N}} \doteq 1$$

obtém-se,

$$\overline{S}_e^2 \doteq \frac{S^2}{\overline{N}} [1 + (\overline{N} - 1)\, \delta]$$

e,

$$S_d^2 \doteq S^2 (1 - \delta)$$

que, substituídas na expressão aproximada de $V(\overline{\overline{y}}_{Ac2})$,

$$V(\overline{\overline{y}}_{Ac2}) \doteq \frac{\overline{S}_e^2}{m} + \frac{S_d^2}{m\overline{n}} \quad (\text{com } M >> m \quad \text{e} \quad \overline{N} >> \overline{n})$$

dão,

$$V(\overline{\overline{y}}_{Ac2}) \doteq \frac{S^2}{m\overline{N}} [1 + (\overline{N} - 1)\, \delta] + \frac{S^2 (1 - \delta)}{m\overline{n}} =$$

$$= \frac{S^2}{m} \left[\frac{1}{N} + \delta \right] + \frac{S^2 (1 - \delta)}{m\overline{n}} =$$

$$= \frac{S^2\, \delta}{m} + \frac{S (1 - \delta)}{m\overline{n}} = \frac{S^2}{m\overline{n}} [1 + (\overline{n} - 1)\, \delta]$$

Considerando que $\frac{S^2}{m\overline{n}}$ é a expressão de $V(\overline{y})$ na *Als*, escreve-se:

$$V(\overline{\overline{y}}_{Ac2}) \doteq V(\overline{y}) [1 + (\overline{n} - 1)\, \delta]$$

Agora, o efeito da conglomeração é $[1 + (\bar{n} - 1)\ \delta]$.

Se $\delta > 0$, é interessante manter $\bar{n}$ pequeno, o que implica em m grande. Isto é, se trabalha com mais unidades primárias e menores subamostras.

Se $\delta < 0$, o interessante é aumentar $\bar{n}$. É claro que a melhor solução é fazer $\bar{n} = \overline{N}$ ou seja, fazer uma *Ac1*.

No Exemplo 11, o efeito da conglomeração é:

$$1 + (\bar{n} - 1)\ \delta = 1 + (6 - 1),\ 201 \doteq 0$$

Para baixar esse efeito, poder-se-ia reduzir a relação de custos $\frac{C_1}{C_2}$ ou partir para uma nova unidade primária, com menor δ.

13 – ESTIMAÇÃO DE PROPORÇÃO

Suponha-se, como se tem feito em Capítulos anteriores, que as unidades secundárias estejam separadas em duas classes disjuntas: A e $\bar{A}$ (não A).

Seja a_i' o número de *US* da subamostra na UP_i', que pertencem à classe A.

Com amostra autoponderada, o estimador da proporção da classe A é:

$$p_{Ac2} = \frac{1}{n} \sum_{i=1}^{m} a_i'$$

e com aproximação para o estimador da variância,

$$V(p_{Ac2}) = \frac{m}{m-1} \left(\frac{1}{n}\right)^2 \sum_{i=1}^{m} \left(a_i' - \frac{1}{m} \sum_{i=1}^{m} a_i'\right)^2$$

13.1 — *Exemplo*

No Exemplo 11, deseja-se, também, estimar a proporção de estabelecimentos com equipamento mecanizado. Para isso, levantou-se na subamostra o número de estabelecimentos com equipamento mecanizado, obtendo-se o quadro:

SETORES DA SUBA-MOSTRA	N.º DE ESTABELE-CIMENTOS (N_i)	TAMANUO DA SUBA-MOSTRA (n'_i)	PRODUÇÃO TOTAL DA SUBA-MOSTRA $\sum_{j=1}^{n'_i} y_{ij}$	N.º DE ESTABELE-CIMENTOS C/EQUI-PAMENTO (a'_i)
1	58	12	1 300	8
2	35	7	980	4
62	47	9	640	7
Soma	1 920	412	57 320	393

Estimativa da proporção de estabelecimentos com equipamento mecanizado.

$$p_{Ac2} = \frac{393}{412} = 0{,}954 \quad \text{ou} \quad 95{,}4\%$$

Estimativa da variância aproximada de P_{Ac2}

$$\hat{V}(p_{Ac2}) = \frac{62}{61}\left(\frac{1}{412}\right)^2 (179{,}139) = 0{,}00107$$

$$\sqrt{\hat{V}(p_{Ac2})} = 0{,}0327$$

14 – CONTROLE DE VARIAÇÃO DE TAMANHO DAS UNIDADES PRIMÁRIAS

Se o coeficiente de correlação intraclasse é positivo, a subamostragem melhora a eficiência, posto que se substitue $\overline{N}$ por $\overline{n}$ no efeito da conglomeração. Esse fato pode ser observado no Capítulo 5, quando se trata de unidades primárias com tamanho igual.

No entanto, a influência da variação de tamanho ainda persiste na estimação do total, posto que a variância do estimador,

$$V(\hat{Y}^{*}_{Ac2}) = M^2 \frac{M-m}{M} \frac{S_e^2}{m} + \frac{M}{m} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{m_i}$$

ainda depende da variabilidade das *UP*.

Desse modo, as diversas formas de controle da variação de tamanho enunciadas na *Ac1*, podem ser repetidas na *Ac2*.

CAPÍTULO 5 — Amostragem de Conglomerados em 2 — estágios. Controle da variação de tamanho: Estimador de razão.

1 — INTRODUÇÃO

Estuda-se, neste Capítulo, o Estimador de Razão, tendo como característica auxiliar o tamanho das unidades primárias. O caso em que a característica auxiliar é uma medida de tamanho, é proposto como exercício.

2 — ESTIMADOR DE RAZÃO DE $\overline{\overline{Y}}$

Sabe-se que a média por *US* é:

$$\overline{\overline{Y}} = \frac{\sum_{i=1}^{M} Y_i}{\sum_{i=1}^{M} N_i} = \frac{\overline{Y}}{\overline{\overline{N}}}$$

o que mostra que $\overline{\overline{Y}}$ pode ser entendida como uma razão de duas médias.

Um estimador consistente de $\overline{\overline{Y}}$ é obtido substituindo o numerador e o denominador por estimadores não tendenciosos.

Desse modo, representando por $\bar{\bar{y}}^{R}_{Ac2}$ esse estimador consistente. tem-se:

$$\bar{\bar{y}}^{R}_{Ac2} = \frac{\frac{1}{m}\sum_{i=1}^{m} N'_i \bar{y}_i}{\frac{1}{m}\sum_{i=1}^{m} N'_i} = \frac{\sum_{i=1}^{m} N'_i \bar{y}_i}{\sum_{i=1}^{m} N'_i}$$

3 – VARIÂNCIA DE $\bar{\bar{y}}^{R}_{Ac2}$

Sabe-se que se $R = \frac{\bar{Y}}{\bar{X}}$, a variância de $\hat{R} = \frac{\bar{y}}{\bar{x}}$ estimador consistente de R, é:

$$V(\hat{R}) = \frac{1}{\bar{X}^2}\left[V(\bar{y}) + R^2 V(\bar{x}) - 2RC(\bar{y},\bar{x})\right] \qquad (3.I)$$

onde $\bar{y}$ é estimador não tendencioso de $\bar{Y}$ e $\bar{x}$ é estimador não tendencioso de $\bar{X}$.

Adaptando ao caso $R = \bar{\bar{Y}}$ obtém-se:

$X = \bar{N}$

$$V(\bar{y}) = V\left(\frac{1}{m}\sum_{i=1}^{m} N'_i \bar{y}_i\right) = V(\bar{y}_{Ac2}) = \frac{M-m}{M} \cdot \frac{S^2_{ey}}{m} +$$

$$+ \frac{1}{Mm}\sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i} \quad \text{com } S^2_{ey} = \frac{1}{M-1}\sum_{i=1}^{M}(Y_i - \bar{Y})^2$$

$$V(\bar{x}) = V\left(\frac{1}{m}\sum_{i=1}^{m} N'_i\right) = \frac{M-m}{M} \cdot \frac{S^2_{ex}}{m} \text{ com } S^2_{ex} = \frac{1}{M-1}\sum_{i=1}^{M}(N_i - \bar{N})^2$$

Para achar $C(\bar{y},\bar{x}) = C\left(\frac{1}{m}\sum_{i=1}^{m} N'_i \bar{y}_i, \frac{1}{m}\sum_{i=1}^{m} N'_i\right)$ usa-se a a covariância condicional, tendo em vista que $\bar{y}_i$, média da subamostra, depende da unidade primária selecionada. Desse modo, é necessário fixar uma *UP'* e depois fazer a variação entre as possíveis *UP*.

Tem-se:

$$C\left(\frac{1}{m}\sum_{i=1}^{m} N'_i\bar{y}_i,\ \frac{1}{m}\sum_{i=1}^{m} N'_i\right) =$$

$$= \underset{UP'_i}{C}\left[E\left(\frac{1}{m}\sum_{i=1}^{m} N'_i\bar{y}_i \mid UP'_i \text{ fix.}\right),\ E\left(\frac{1}{m}\sum_{i=1}^{m} N'_i \mid UP'_i \text{ fix.}\right)\right] +$$

$$+ \underset{UP'_i}{E}\left\{C\left[\left(\frac{1}{m}\sum_{i=1}^{m} N'_i\bar{y}_i,\ \frac{1}{m}\sum_{i=1}^{m} N'_i\right) \mid UP'_i \text{ fix.}\right]\right\}$$

Para a primeira parte do segundo membro, pode-se escrever:

$$\underset{UP'_i}{C}\left[\frac{1}{m}\sum_{i=1}^{m} Y'_i,\ \frac{1}{m}\sum_{i=1}^{m} N'_i\right] = \frac{M-m}{M}\,\frac{S_{exy}}{m}$$

$$\text{com } S_{exy} = \frac{\sum_{i=1}^{M}(Y_i - \bar{Y})(N_i - \bar{N})}{M-1}$$

Para a segunda parte, observe-se que se pode escrever $\frac{1}{m}\sum_{i=1}^{m} N'_i$ na forma $\frac{1}{m}\sum_{i=1}^{m} N'_i\bar{x}_i$ onde $\bar{x}_i$ é a média de uma característica que assume n'_i valores iguais a *1*. Nessa conformidade $C(\bar{y}_i, \bar{x}_i) = 0$ e a segunda parte também é nula.

Em resumo,

$$C\left(\frac{1}{m}\sum_{i=1}^{m} N'_i\bar{y}_i,\ \frac{1}{m}\sum_{i=1}^{m} N'_i\right) = \frac{M-m}{M}\cdot\frac{S_{exy}}{m}$$

Fazendo as substituições em (3.I) obtém-se:

$$V(\bar{\bar{y}}^{R}_{Ac2}) = \frac{1}{\bar{\bar{N}}^2}\cdot\frac{M-m}{Mm}\left[S^2_{ey} + \bar{\bar{Y}}^2 S^2_{ex} - 2\bar{\bar{Y}}\,S_{exy}\right] +$$

$$+ \frac{1}{\bar{\bar{N}}^2 Mm}\sum_{i=1}^{M} N_i^2\,\frac{N_i - n_i}{N_i}\cdot\frac{S_i^2}{n_i}$$

Observe-se que:

$$S_{ey}^2 + \overline{\overline{Y}}^2 S_{ex}^2 - 2\overline{\overline{Y}}\, S_{exy} = \frac{1}{M-1} \sum_{i=1}^{M} [(Y_i - \overline{Y})^2 +$$

$$+ \overline{Y}^2 (N_i - \overline{N})^2 - 2\overline{\overline{Y}} (Y_i - \overline{Y})(N_i - \overline{N})] =$$

$$= \frac{1}{M-1} \sum_{i=1}^{M} [(Y_i - \overline{Y}) - \overline{\overline{Y}} (N_i - \overline{N})]^2$$

$$= \frac{1}{M-1} \sum_{i=1}^{M} (Y_i - \overline{\overline{Y}} N_i)^2 =$$

$$= \frac{1}{M-1} \sum_{i=1}^{M} N_i^2 (\overline{Y}_i - \overline{\overline{Y}})^2 = S_{eR}^2$$

Substituindo em $V(\overline{\overline{y}}_{Ac2}^{R})$ vem:

$$V(\overline{\overline{y}}_{Ac2}^{R}) = \frac{M-m}{M\overline{N}^2} \frac{S_{eR}^2}{m} + \frac{1}{Mm} \sum_{i=1}^{M} \left(\frac{N_i}{\overline{N}} \right)^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

4 – ESTIMADOR CONSISTENTE DE $V(\overline{\overline{y}}_{Ac2})$

Sabe-se que um estimador consiste de (3.I) é:

$$\hat{V}(\hat{R}) = \frac{1}{\overline{\overline{X}}^2} [\hat{V}(\overline{y}) + \hat{R}^2 \hat{V}(\overline{x}) - 2\hat{R}\ \hat{C}(\overline{y}, \overline{x})] \qquad (4.I)$$

onde $\hat{V}(\overline{y})$, $\hat{V}(\overline{x})$, $\hat{C}(\overline{y}, \overline{x})$ são estimadores não tendenciosos de $V(\overline{y})$, $V(\overline{x})$, $C(\overline{y}, \overline{x})$ respectivamente e R é estimador consistente de R.

Adaptando ao caso, tem-se:

$$\hat{V}(\overline{\overline{y}}) = \hat{V}(\overline{\overline{y}}_{Ac2}) = \frac{M-m}{M} \cdot \frac{s_{ey}^2}{m} + \frac{1}{Mm} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \cdot \frac{s_i^2}{n_i'}$$

expressão já conhecida do Capítulo 4, onde

$$s_{ey}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (N_i' \bar{y}_i - \bar{y}_{Ac2})^2 \quad \text{e} \quad s_i^2 \text{ é a variância da subamostra.}$$

$$\hat{V}(\bar{x}) = \frac{M-m}{M} \cdot \frac{s_{ex}^2}{m} \quad \text{onde} \quad s_{ex}^2 = \frac{1}{m-1} \sum_{i=1}^{m} (N_i' - \bar{N}')^2$$

$$\hat{C}(\bar{y}, \bar{x}) = \frac{M-m}{M} \cdot \frac{s_{eyy}}{m} \quad \text{onde} \quad s_{exy} =$$

$$= \frac{1}{m-1} \sum_{i=1}^{m} (N_i' \bar{y}_i - \bar{y}_{Ac2})(\bar{N}_i' - \bar{N}')$$

Fazendo as substituições em (4.1) obtém-se:

$$\hat{V}(\bar{y}_{Ac2}^R) = \frac{1}{\bar{N}^2} \left[\frac{M-m}{M} \cdot \frac{s_{ey}^2}{m} + \frac{1}{Mm} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \cdot \frac{s_i^2}{n_i'} + \right.$$

$$\left. + (\bar{\bar{y}}_{Ac2}^R)^2 \frac{M-m}{M} \cdot \frac{s_{ex}^2}{m} - 2\, \bar{\bar{y}}_{Ac2}^R \frac{M-m}{M} \frac{s_{exy}}{m} \right] =$$

$$= \frac{1}{\bar{N}^2} \frac{M-m}{Mm} [s_{ex}^2 + (\bar{\bar{y}}_{Ac2}^R)^2 s_{ex}^2 - 2\, \bar{\bar{y}}_{Ac2}^R s_{exy}] +$$

$$+ \frac{1}{\bar{N}^2 Mm} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \cdot \frac{s_i^2}{n_i'} =$$

$$= \frac{1}{\bar{N}^2} \frac{M-m}{Mm(m-1)} \sum_{i=1}^{m} [(N_i' \bar{y}_i - \bar{y}_{Ac2})^2 + (\bar{\bar{y}}_{Ac2}^R)^2 (N_i' - \bar{N}')^2 -$$

$$- 2\, \bar{\bar{y}}_{Ac2}^R (N_i' \bar{y}_i - \bar{y}_{Ac2}) (N_i' - \bar{N}')] + \frac{1}{\bar{N}'^2 Mm} \sum_{i=1}^{m} N_i'^2 \frac{N_i' - n_i'}{N_i'} \frac{s_i^2}{n_i'}$$

donde,

$$\hat{V}(\bar{\bar{y}}_{Ac2}^R) = \frac{M-m}{Mm(m-1)} \sum_{i=1}^{m} \left(\frac{N_i'}{\bar{N}} \right)^2 (\bar{y}_i - \bar{\bar{y}}_{Ac2}^R)^2 +$$

$$+ \frac{1}{Mm} \sum_{i=1}^{m} \left(\frac{N_i'}{\bar{N}} \right)^2 \frac{N_i' - n_i'}{N_i'} \frac{s_i^2}{n_i'}$$

5 – ESTIMADOR DE RAZÃO DO TOTAL Y

Representando por y^{*R}_{Ac2} o estimador consistente de Y, tem-se:

$$y^{*R}_{Ac2} = M\bar{N}\bar{\bar{y}}^{R}_{Ac2}$$

Dessa igualdade, seguem-se as expressões dos estimadores e variâncias

$$y^{*R}_{Ac2} = M\bar{N} \cdot \frac{\sum_{i=1}^{m} N'_i \bar{y}_i}{\sum_{i=1}^{m} N'_i}$$

$$V(y^{*R}_{Ac2}) = M^2 \frac{M-m}{M} \cdot \frac{S^2_{eR}}{m} + \frac{M}{m} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

Observe-se que $V(y^{*R}_{Ac2})$ depende da variação das médias e não dos totais, variação essa que sofre pouca influência da variação de tamanho das *UP*.

$$\hat{V}(y^{*R}_{Ac2}) = M^2 \frac{M-m}{Mm(m-1)} \sum_{i=1}^{m} N'^2_i (\bar{y}_i - \bar{\bar{y}}^{R}_{Ac2})^2 + \\ + \frac{M}{m} \sum_{i=1}^{m} N'^2_i \frac{N'_i - n'_i}{N'_i} \cdot \frac{s^2_i}{n'_i} \qquad (5.I)$$

De (5.I) obtém-se:

$$\hat{V}(y^{*R}_{Ac2}) = \frac{M^2}{m^2} \left[\frac{M-m}{M} \frac{m}{m-1} \sum_{i=1}^{m} N'^2_i (\bar{y}_i - \bar{\bar{y}}^{R}_{Ac2})^2 + \right. \\ \left. + \frac{m}{M} \sum_{i=1}^{m} N'^2_i \frac{N'_i - n'_i}{N'_i} \frac{s^2_i}{n'_i} \right]$$

Supondo $M \gg m$, a expressão acima se aproxima para:

$$\hat{V}(y^{*R}_{Ac2}) \doteq \frac{M^2}{m} \frac{1}{m-1} \sum_{i=1}^{m} N_i'^2 (\bar{y}_i - \bar{\bar{y}}^R_{Ac2})^2$$

ou,

$$V(y^{*R}_{Ac2}) \doteq M^2 \frac{s^2_{eR}}{m} \text{ com } s^2_{eR} = \frac{1}{m-1} \sum_{i=1}^{m} N_i'^2 (\bar{y}_i - \bar{\bar{y}}^R_{Ac2})^2$$

6 – AMOSTRA AUTOPONDERADA

Sabe-se (Capítulo 4) que a condição de amostra autoponderada é dada pela igualdade:

$$\frac{m}{M} \cdot \frac{n}{N} = \frac{n}{N}$$

ou seja, o produto das frações de amostragem em cada estágio é constante e igual a fração geral de amostragem $\frac{n}{N}$. Em outras palavras, todas as *US* têm a mesma probabilidade $\frac{n}{N}$ de pertencer à amostra.

Nessa condição,

$$\bar{\bar{y}}^R_{Ac2} = \frac{\sum_{i=1}^{m} N_i' \bar{y}_i}{\sum_{i=1}^{m} N_i'} = \frac{\bar{N}}{\bar{n}} \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i'} y_{ij}}{\sum_{i=1}^{m} N_i'} =$$

$$= \frac{1}{f^2} \cdot \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i'} y_{ij}}{\sum_{i=1}^{m} N_i'} \text{ sendo } f_2 \text{ a fração de amostragem de 2.º estágio.}$$

Para o estimador da variância aproximada de $V(\bar{y}^R_{Ac2}) = \frac{s^2_{eR}}{\bar{N}^2 m}$ tem-se:

$$s^2_{eR} = \frac{1}{m-1} \sum_{i=1}^{m} \frac{N_i'^2}{n_i'^2} \left(\sum_{j=1}^{n_i'} y_{ij} - \frac{\sum_{i=1}^{m} N_i' \sum_{j=1}^{n_i'} y_{ij}}{\sum_{i=1}^{m} N_i'} \right)^2 =$$

$$= \frac{1}{m-1} \left(\frac{mN}{nM} \right)^2 \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i'} y_{ij} - \frac{\sum_{i=1}^{m} N_i' \sum_{j=1}^{n_i'} y_{ij}}{\sum_{i=1}^{m} N_i'} \right)^2$$

Conseqüentemente,

$$V(\bar{y}^R_{Ac2}) = \frac{m}{(m-1)n^2} \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i'} y_{ij} - \frac{\sum_{i=1}^{m} N_i' \sum_{j=1}^{n_i'} y_{ij}}{\sum_{i=1}^{m} N_i'} \right)^2$$

7 — EXEMPLO

Considere-se o Exemplo 6 — Capítulo 2.

Suponha-se que se deseja estimar o consumo médio semanal por domicílio (em unidades do produto) fazendo subamostragem mas mantendo a fração geral de amostragem $\frac{650}{26000} = \frac{1}{40}$. Para isso, fixou-se a fração de amostragem de 1.º estágio em $\frac{1}{8}$ e a fração de amostragem de 2.º estágio em $\frac{1}{5}$.

Desse modo, considerando a existência de 400 quarteirões, foram selecionados $400 \frac{1}{8} = 50$ arteirões e, em cada quarteirão da amostra, foram selecionados $\frac{1}{5}$ dos domicílios.

Desse modo, obteve-se o quadro (parcialmente apresentado):

QUARTEIRÕES DA AMOSTRA	N.º DE DOMICÍLIOS (N_i')	N.º DE DOMICÍLIOS DA AMOSTRA (n_i')	TOTAL DE UNIDADES CONSUMIDAS NOS DOMICÍLIOS DA AMOSTRA $\sum_{j=1}^{n_i'} y_{ij}$
1	68	14	53
2	54	11	30
3	50	10	27
.	.	.	.
.	.	.	.
.	.	.	.
50	72	14	41
Soma	3 152	710	1 910

Estimativa do número de unidades consumidas, por domicílio:

$$\bar{\bar{y}}_{Ac2}^{R} = \frac{1}{f^2} \frac{\sum_{i=1}^{m} \sum_{j=i}^{n_i'} y_{ij}}{\sum^{m} N_i'} = 5 \cdot \frac{1910}{3152} = 3{,}03$$

Estimativa da variância de $\bar{\bar{y}}_{Ac2}^{R}$

$$V(\bar{\bar{y}}_{Ac2}^{R}) = \frac{50}{49(710)^2}\, 4500 = 0{,}0091$$

$$\sqrt{V(\bar{\bar{y}}_{Ac2}^{R})} = 0{,}095$$

8 – ESTIMADOR DE RAZÃO, EM RELAÇÃO A UMA CARACTERÍSTICA AUXILIAR QUE NÃO SEJA O TAMANHO

Partindo do estimador de razão de duas características correlacionadas, propõe-se, como exercício, desenvolver os estimadores de $\bar{\bar{Y}}$: estimador consistente, variância (com tendenciosidade desprezível) e respectivo estimador.

CAPÍTULO 6 — Amostragem de conglomerados em 2 — estágios — Controle da variação de tamanho:

Probabilidade desigual de seleção das unidades primárias, com reposição.

1 — CONFIGURAÇÃO DA AMOSTRA

As unidades de π_N são grupadas em M UP e observa-se uma característica y:

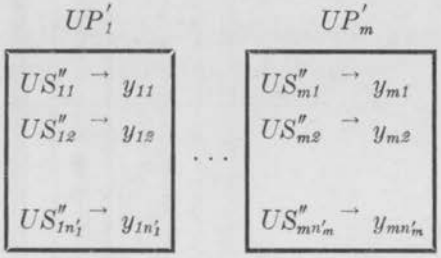
UP_1		UP_M
$US_{11} \rightarrow Y_{11}$		$US_{M1} \rightarrow Y_{M1}$
$US_{12} \rightarrow Y_{12}$	...	$US_{M2} \rightarrow Y_{M2}$
$\vdots$		$\vdots$
$US_{1N_1} \rightarrow Y_{1N_1}$		$US_{MN_M} \rightarrow Y_{MN_M}$

Seja P_i a probabilidade de seleção de UP_i $(i = 1,2,\ldots,M)$

Seleciona-se uma amostra de m UP de acordo com as probabilidades de seleção P_i e com reposição.

UP'_i		UP'_m
$US'_{11} \rightarrow Y'_{11}$		$US'_{m1} \rightarrow Y'_{m1}$
$US'_{12} \rightarrow Y'_{12}$	...	$US'_{m2} \rightarrow Y'_{m2}$
$US'_{1N'_1} \rightarrow Y'_{1N'_1}$		$US'_{mN'_m} \rightarrow Y'_{mN'_m}$

Em cada uma dessas unidades primárias da amostra de 1.º estágio, seleciona-se uma subamostra com igual probabilidade de seleção e sem reposição:



Observe-se que uma *UP* pode ser selecionada mais de uma vez, na amostra de 1.º estágio. Cada vez que é selecionada, seleciona-se uma nova subamostra, independente da anterior, posto que há reposição de cada subamostra selecionada.

A amostra de y é:

$$\{y_{11}, y_{12}, \ldots y_{1m'_1}; \ldots\ldots\ldots; y_{m1}, y_{m2}, \ldots y_{mn'_m}\}$$

2 – PARÂMETROS DE Y:

Total em UP_i: $\quad Y_i = \sum_{j=1}^{N_i} Y_{ij}$

Média em UP_i: $\quad \overline{Y}_i = \dfrac{Y_i}{N_i}$

Variância em UP_i: $S_i^2 = \dfrac{T}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \overline{Y}_i)^2$

Média por UP: $\quad \overline{\overline{Y}} = \dfrac{Y}{M}$

Média por US: $\quad \overline{\overline{\overline{Y}}} = \dfrac{Y}{MN}$ com $N = \dfrac{N}{M}$

3 – ESTATÍSTICAS

Total em UP_i':

$$Y_i' = \sum_{j=1}^{N_i'} Y_{ij}'$$

Média em UP_i':

$$\overline{Y}_i' = \frac{Y_i'}{N_i'}$$

Variância em UP_i':

$$S_i'^2 = \frac{1}{N_i' - 1} \sum_{j=1}^{N_i'} (Y_{ij}' - \overline{Y}_i')^2$$

Média da subamostra em UP_i':

$$\overline{y}_i = \frac{1}{n_i'} \sum_{j=1}^{n_i'} y_{ij}$$

Variância da subamostra em UP_i':

$$s_i^2 = \frac{1}{n_i' - 1} \sum_{j=1}^{n_i'} (y_{ij} - \overline{y}_i)^2$$

4 – TEOREMA

Um estimador não tendencioso de y é:

$$y_{Ac2}^{*P} = \frac{1}{m} \sum_{i=1}^{m} \frac{N_i' \overline{y}_i}{P_i'}$$

Prova

$$E(y_{Ac2}^{*P}) = E\left(\frac{1}{m} \sum_{i=1}^{m} \frac{N_i' \overline{y}_i}{P_i'}\right) =$$

$$= \underset{UP'_i}{E}\left[E\left(\frac{1}{m}\sum_{i=1}^{m}\frac{N'\,\bar{y}_i}{P'_i} \mid UP'_i \text{ fixado}\right)\right] =$$

$$= \underset{UP'_i}{E}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{N'_i\,\bar{Y}'_i}{P'_i}\right] =$$

$$= \underset{UP'_i}{E}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{Y'_i}{P'_i}\right] =$$

$$= \sum_{i=1}^{M}\frac{Y_i}{P_i}P_i = Y$$

4.1 — *Corolário*

Um estimador não tendencioso da média $\bar{\bar{Y}}$ é:

$$\bar{\bar{y}}^{P}_{Ac2} = \frac{1}{Nm}\sum_{i=1}^{m}\frac{N'_i\,\bar{y}_i}{P'_i}$$

Prova

Fazer como exercício.

5 — VARIÂNCIA DE y^{*P}_{Ac2}

$$V(y^{*P}_{Ac2}) = \underset{UP'_i}{V}\left[E(y^{*P}_{Ac2} \mid UP'_i \text{ fixado})\right] +$$

$$+ \underset{UP'_i}{E}\left[V(y^{*P}_{Ac2} \mid UP'_i \text{ fixado})\right] =$$

$$= \underset{UP'_i}{V}\left[E\left(\frac{1}{m}\sum_{i=1}^{m}\frac{N'_i\,\bar{y}_i}{P'_i} \mid UP'_i \text{ fixado}\right)\right] +$$

$$+ \underset{UP'_i}{E}\left[V\left(\frac{1}{m}\sum_{i=1}^{m}\frac{N'_i\,\bar{y}_i}{P'_i} \mid UP'_i \text{ fixado}\right)\right] \qquad (5.I)$$

Para a primeira parcela do segundo membro, tem-se:

$$\underset{UP'_i}{V}\left[E\left(\frac{1}{m}\sum_{i=1}^{m}\frac{N'_i\,\bar{y}_i}{P'_i} \mid UP'_i \text{ fixado}\right)\right] =$$

$$= \underset{UP'_i}{V}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{N'_i\,\bar{Y}'_i}{P'_i}\right] = \frac{1}{m}\sum_{i=1}^{M}\left(\frac{Y_i}{P_i} - Y\right)^2 P_i \quad (5.II)$$

de acordo com a expressão semelhante da Ac_1.

Para a segunda parcela, tem-se:

$$\underset{UP'_i}{E}\left[V\left(\frac{1}{m}\sum_{i=1}^{m}\frac{N'_i\,\bar{y}_i}{P'_i} \mid UP'_i \text{ fixado}\right)\right] =$$

$$= \underset{UP'_i}{E}\left[\frac{1}{m^2}\sum_{i=1}^{m}\left(\frac{N'_i}{P'_i}\right)^2 \frac{N'_i - n'_i}{N'_i}\,\frac{S'^2_i}{n'_i}\right] =$$

$$= \frac{1}{m}\sum_{i=1}^{M}\left(\frac{N_i}{P_i}\right)^2 \frac{N_i - n_i}{N_i}\,\frac{S_i^2}{n_i}\,P_i =$$

$$= \frac{1}{m}\sum_{i=1}^{M}(N_i)^2\,\frac{1}{P_i}\,\frac{N_i - n_i}{N_i}\,\frac{S_i^2}{n_i} \quad (5.III)$$

Substituindo (5.II) e (5.III) em (5.I) obtém-se:

$$V(y^{*P}_{Ac2}) = \frac{1}{m}\sum_{i=1}^{M}\left(\frac{Y_i}{P_i} - Y\right)^2 P_i +$$

$$+ \frac{1}{m}\sum_{i=1}^{M}(N_i)^2\,\frac{1}{P_i}\,\frac{N_i - n_i}{N_i}\,\frac{S_i^2}{n_i}$$

5.1 — *Exercício*

Achar $V(\bar{\bar{y}}^{P}_{Ac2})$

6 — TEOREMA

Um estimador não tendencioso de $V(y^{*P}_{Ac2})$ é:

$$\hat{V}(y^{*P}_{Ac2}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{N'_i\,\bar{y}_i}{P'_i} - y^{*P}_{Ac2}\right)^2$$

Prova

$$E[\hat{V}(y_{Ac2}^{*P})] = \frac{1}{m(m-1)} E\left[\sum_{i=1}^{m}\left(\frac{N_i' \bar{y}_i}{P_i'}\right)^2 - m(y_{Ac2}^{*P})^2 = \right]$$

$$= \frac{1}{m(m-1)}\left[\sum_{i=1}^{m} E\left(\frac{N_i' \bar{y}_i}{P_i'}\right)^2 - mE(y_{Ac2}^{*P})^2\right] \qquad (6.I)$$

Para a primeira expectância no segundo membro, tem-se:

$$E\left(\frac{N_i' \bar{y}_i}{P_i'}\right)^2 = \underset{UP_i'}{E}\left\{E\left[\left(\frac{N_i' \bar{y}_i}{P_i'}\right)^2 \mid UP_i' \text{ fixado}\right]\right\} =$$

$$= \underset{UP_i'}{E}\left\{V\left[\frac{N_i' \bar{y}_i}{P_i'}\right) \mid UP_i' \text{fixado}\right] + \left[E\left(\frac{N_i' y_i}{P_i'} \mid UP_i' \text{ fixado}\right)\right]^2\right\} =$$

$$= \underset{UP_i'}{E}\left[\left(\frac{N_i'}{P_i'}\right)^2 \frac{N_i' - n_i'}{N_i'} \frac{S_i'^2}{n_i'} + \left(\frac{N_i' \bar{Y}_i'}{P_i'}\right)^2\right] =$$

$$= \sum_{i=1}^{M}\left(\frac{N_i}{P_i}\right)^2 \cdot \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i} P_i + \sum_{i=1}^{M}\left(\frac{N_i \bar{y}_i}{P_i}\right)^2 P_i \qquad (6.II)$$

Para a segunda expectância no segundo membro, tem-se:

$$E(y_{Ac2}^{*P})^2 = V(y_{Ac2}^{*P}) + [E(y_{Ac2}^{*P})]^2 = V(y_{Ac2}^{*P}) - Y^2 \qquad (6.III)$$

Substituindo (6.II) e (6.III) em (6.I) vem:

$$E[\hat{V}(y_{Ac2}^{*P})] = \frac{1}{m-1}\left[\sum_{i=1}^{M}\left(\frac{N_i}{P_i}\right)^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} P_i + \right.$$

$$\left. + \sum_{i=1}^{M}\left(\frac{N_i \bar{Y}_i}{P_i}\right)^2 P_i - V(y_{Ac2}^{*P}) - Y^2\right] =$$

$$= \frac{1}{m-1}\left[\sum_{i=1}^{M}\left(\frac{N_i Y_i}{P_i} Y\right)^2 + \sum_{i=1}^{M}\left(\frac{N_i}{P_i}\right)^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} P_i - \right.$$

$$\left. V(y_{Ac2}^{*P})\right] = \frac{1}{m-1}[m\ V(y_{Ac2}^{*P}) - V(y_{Ac2}^{*P})] = V(y)_{Ac2}^{*P}$$

6.1 — *Exercício*

Achar $\hat{V}(\bar{y}_{Ac2}^{P})$

7 – PROBABILIDADE DE SELEÇÃO PROPORCIONAL AO TAMANHO

Valem as considerações feitas na *Ac1*, Capítulo 3. Define-se a probabilidade proporcional ao tamanho por,

$$P_i = \frac{N_i}{N} \quad (i = 1, 2, \ldots M)$$

e a probabilidade proporcional a uma medida de tamanho por,

$$P_i = \frac{X_i}{X} \quad (i = 1, 2, \ldots M)$$

8 – AMOSTRA AUTOPONDERADA

A probabilidade de uma *US* qualquer pertencer a amostra é:

$$mP_i \frac{n_i}{N_i}$$

A amostra é autoponderada se essa probabilidade é constante e igual a fração de amostragem geral $\frac{n}{N}$. Tem-se, então:

$$mP_i \frac{n_i}{N_i} = \frac{n}{N} \tag{8.I}$$

ou,

$$mP_i \frac{n_i}{N_i} = f$$

Observe-se que, em média, $\sum_{i=1}^{m} n_i'$ dá o tamanho n prefixado, De fato, de (8.I) obtém-se:

$$E\left(\sum_{i=1}^{m} n_i'\right) = \frac{n}{mN} E\left(\sum_{i=1}^{m} \frac{N_i'}{P_i'}\right) = \frac{nmN}{mN} = n$$

8.1 — *Adequação da expressão de* y^{*P}_{Ac2}

Substituindo (8.I) em,

$$y^{*P}_{Ac2} = \frac{1}{m} \sum_{i=1}^{m} \frac{N'_i}{P'_i} \sum_{j=1}^{n'_i} \frac{y_{ij}}{n'_i}$$

obtém-se:

$$y^{*P}_{Ac2} = \frac{1}{f} \sum_{i=1}^{m} \sum_{j=1}^{n'_i} y_{ij}$$

mesma expressão já encontrada com probabilidade igual de seleção

8.2 — *Adequação da expressão de* $\hat{V}(y^{*P}_{Ac2})$

Da expressão,

$$\hat{V}(y^{*P}_{Ac2}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{N'_i \bar{y}_i}{P'_i} - y^{*P}_{Ac2} \right)^2 =$$

$$= \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{N'_i}{P'_i n'_i} \sum_{j=1}^{n'_i} y_{ij} - y^{*P}_{Ac2} \right)^2$$

obtém-se, por substituição de acordo com (8.I):

$$\hat{V}(y^{*P}_{Ac2}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{m}{f} \sum_{j=1}^{n'_i} y_{ij} - y^{*P}_{Ac2} \right)^2 =$$

$$= \frac{m}{(m-1) f^2} \sum_{i=1}^{m} \left(\sum_{j=1}^{n'_i} y_{ij} - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n'_i} y_{ij} \right)^2$$

8.3 — *Exercício*

Achar $\bar{y}^{*P}_{Ac2}$ e $\hat{V}(\bar{y}^{*P}_{Ac2})$ com amostra autoponderada.

9 — EXEMPLO

Uma determinada localidade tem 53 povoados, dos quais se selecionam 14, com reposição e probabilidade de seleção proporcional à população do Censo.

No povoado i da amostra, faz-se uma listagem das N_i' fazendas de gado e seleciona-se uma subamostra de fazendas, com tamanho suficiente para se obter uma fração geral de amostragem $f = \frac{1}{100}$ das fazendas, com o objetivo de estimar o número total de cabeças de gado.

Da igualdade,

$$mP_i' \frac{n_i'}{N_i'} = \frac{1}{100}$$

obtém-se a fração de amostragem de 2.º estágio,

$$\frac{n_i'}{N_i'} = \frac{1}{100mP_i'} = \frac{1}{1400P_i'}$$

Feita a seleção dos 14 povoados e a contagem das fazendas, aplicou-se a fração de amostragem de 2.º estágio, obtendo-se as fazendas da subamostra e levantando, em cada uma, o número de cabeças de gado.

POVOADOS DA AMOSTRA	PROBABILIDADE DE SELEÇÃO (P_i') DO POVOADO.	N.º DE FAZENDAS NO POVOADO (N_i')	FRAÇÃO DE AMOSTRAGEM DE 2.º ESTÁGIO. $\frac{n_i'}{N_i'}$	N.º DE FAZENDAS NA SUBAMOSTRA (n_i')	N.º TOTAL DE CABEÇAS DE GADO $\sum_{j=1}^{n_i'} y_{ij}$
1	0,0026	19	0,2747	5	2 200
2	0,0098	23	0,0729	2	820
3	0,0146	31	0,0489	2	760
4	0,0167	40	0,0428	2	1 100
5	0,0187	54	0,0382	2	600
6	0,0187	54	0,0382	2	510
7	0,0220	39	0,0325	1	300
8	0,0249	55	0,0385	2	1 200
9	0,0258	46	0,0277	1	500
10	0,0298	83	0,0240	2	880
11	0,0362	74	0,0197	1	300
12	0,0370	70	0,0193	1	410
13	0,0465	60	0,0154	1	570
14	0,0465	60	0,0154	1	350
Soma	—	—	—	—	10 500

Estimativa do número total de cabeças de gado:

$$y^{*P}_{Ac2} = \frac{1}{f} \sum_{j=1}^{m} \sum_{i=1}^{n'_i} y_{ij} = 100(10500) = 1050000 \text{ cabeças}$$

Estimativa da variância de y^{*P}_{Ac2}

$$\hat{V}(y^{*P}_{Ac2}) = \frac{m}{(m-1)f^2} \sum_{i=1}^{m} \left(\sum_{j=1}^{n'_i} y_{ij} - \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n'_i} y_{ij} \right)^2 =$$

$$= \frac{14}{13} (100)^2 (3312100) = 35668{,}77\ (1000)^2$$

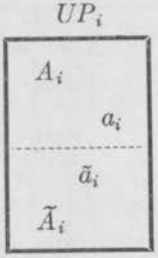
$$\sqrt{\hat{V}(y^{*P}_{Ac2})} = 188861$$

10 – ESTIMAÇÃO DE PROPORÇÃO

Suponha-se a população dividida nas classes A e $\tilde{A}$.

A UP_i fica dividdia nas classes, com A_i e $\tilde{A}_i$ unidades, respectivamente.

A subamostra de tamanho n_i fica também dividida nas duas classes, com a_i e $\tilde{a}_i$ unidades.



10.1 – *Teorema*

Um estimador não tendencioso de $\frac{1}{N} \sum_{i=1}^{M} A_i$ proporção de A é:

$$p_{Ac2} = \frac{1}{mN} \sum_{i=1}^{m} \frac{N'_i}{P'_i} p_i \qquad \text{onde } p_i = \frac{a'_i}{n'_i} \text{ é a proporção de } A \text{ na subamostra}$$

Prova

Basta substituir $\bar{y}_i$ na expressão de $\bar{\bar{y}}_{Ac2}$ por p_i

Se a amostra é autoponderada, ocorre a condição:

$$mP_i \frac{n_i}{N_i} = \frac{n}{N}$$

donde se obtém:

$$p_{Ac2} = \frac{1}{n} \sum_{i=1}^{m} a_i'$$

10.2 — *Teorema*

Um estimador não tendencioso de $V(p_{Ac2})$ é:

$$\hat{V}(p_{Ac2}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{N_i'}{NP_i'} p_i - p_{Ac2} \right)^2$$

Prova

Fazer como exercício

Se a amostra é autoponderada,

$$\hat{V}(p_{Ac2}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{m}{n} a_i' - p_{Ac2} \right)^2$$

10.3 — *Exemplo*

Considere-se o Exemplo 2.9. Suponha-se que se deseja estimar o número total de empregados do sexo masculino que trabalham nas fazendas. Os valores obtidos na subamostra foram:

POVOADOS DA AMOSTRA	NÚMERO DE EMPREGADOS DO SEXO MASCULINO NA SUBAMOSTRA
1	75
2	42
3	31
4	26
5	40
6	32
7	18
8	41
9	20
10	38
11	16
12	10
13	15
14	15
Soma	419

Estimativa do número total de empregados do sexo masculino:

$$y^*_{Ac2} = \frac{1}{f} \sum_{i=1}^{m} a'_i = 100 \ (41\ 9) \ = 41900 \text{ empregados}$$

Como exercício, achar $\hat{\gamma}(y^*_{Ac2})$

11 – TAMANHO MÉDIO DA AMOSTRA DE 2.º ESTÁGIO

Suponha-se fixado o número m de unidades primárias. O número médio de unidades secundárias na amostra autoponderada de 2.º estágio é:

$$E(n'_i) = \frac{n}{N_m} E\left(\frac{N'_i}{P'_i}\right) = \frac{n}{N_m} \sum_{i=1}^{M} N_i = \frac{n}{m}$$

isto é, fixando em m o número de unidades primárias, o número de unidades secundárias na amostra é, em média,

$$\bar{n} = \frac{n}{m}$$

12 – EXEMPLO

Certa localidade está dividida em 101 regiões. Cada região está dividida em subregiões, com um total de 3 232 subregiões. Finalmente, as subregiões estão divididas em domicílios, com total de 1 717 200 domicílios.

Trata-se de selecionar uma amostra de domicílios em 2 – estágios, utilizando-se como unidade primária a região e como unidade secundária o domicílio.

Sabe-se, de estudos anteriores, que a relação de custos $\frac{C_1}{C_0}$ é da ordem de 20. Desse modo e considerando que o coeficiente de

correlação intraclasse é 0,05 para a característica a estimar, tem-se que o tamanho médio da subamostra é:

$$\overline{n} = \sqrt{20 \frac{0,95}{0,05}} = 19$$

Conseqüentemente, o número de regiões a selecionar é 20, para se obter uma amostra final de 380 domicílios.

A seleção das regiões será feita com probabilidade proporcional ao número de domicílios. Para esse objetivo, dispõe-se de uma listagem das regiões com os respectivos números de domicílios. Considerando que a característica de estimação é a despesa com aluguel, as regiões foram listadas em ordem crescente do imposto predial mediano, para permitir alguma estratificação ao ser feita uma seleção sistemática:

REGIÕES	N.º DE DOMICÍLIOS	N.º ACUMULADO
1	2700	2700
2	1100	3800
3	4400	8200
4	1800	10000
5	3200	13200
6	1800	15000
7	2600	17600
8	14700	32300
9	9600	41900
10	6000	47900
11	4500	52400
12	5200	57600
13	2800	60400
14	1900	62300
15	5100	67400
16	6000	73400
17	6800	80200
18	2700	82900
19	10300	93200
20	6700	99900
21	2000	101900
22	5800	107700
23	5200	112900
24	6000	118900
25	13400	132300
26	2900	135200
27	4500	139700

(Continuação)

REGIÕES	N.º DE DOMICÍLIOS	N.º ACUMULADO
28	10400	150100
29	4300	154400
30	6000	160400
31	4100	164500
32	4800	169300
33	7600	176900
34	4600	181500
35	6900	188400
36	8000	196400
37	16000	212400
38	4400	216800
39	7500	224300
40	3400	227700
41	5700	233400
42	14900	248300
43	2800	251100
44	2500	253600
45	6900	260500
46	5200	265700
47	10000	275700
48	5700	281400
49	7100	288500
50	12100	300600
51	3000	303600
52	8000	311600
53	6500	318100
54	14100	332200
55	5600	337800
56	10300	348100
57	9600	357700
58	23500	381200
59	12200	393400
60	4200	397600
61	15200	412800
62	5400	418200
63	6300	424500
64	4100	428600
65	11000	439600
66	5900	445500
67	10100	455600
68	5800	461400
69	1700	463100
70	32400	495500
71	6100	501600
72	20000	521600
73	83600	607900
74	17200	625100
75	12400	637500
76	11000	648500
77	5100	653600

(Conclusão)

REGIÕES	N.º DE DOMICÍLIOS	N.º ACUMULADO
78	35200	688800
79	8300	697100
80	15900	713000
81	19300	732300
82	47300	779600
83	13500	793100
84	31900	825000
85	78500	903500
86	54300	957800
87	40800	998600
88	27900	1026500
89	53600	1080100
90	63300	1143400
91	19600	1163000
92	36400	1199400
93	8300	1207700
94	7700	1215400
95	76700	1292100
96	70700	1362800
97	7500	1370300
98	74600	1444900
99	33000	1477900
100	103000	1580900
101	136300	1717200

Seleção das $m = 20$ regiões:

O intervalo de amostra é $\frac{N}{m} = \frac{1\,717\,200}{20} = 85\,860$ domicílios. Observe-se que há 3 regiões com mais de 85 860 domicílios: as regiões 73, 100 e 101. Essas regiões são autorepresentadas, isto é, são incluídas certamente na amostra. Com a retirada dessas 3 regiões, o número total de domicílios fica reduzido para 1 717 200 - 325 600 = 1 391 600. Por conseguinte, tem-se um novo intervalo de amostra $\frac{1\,391\,600}{17} = 81\,859$. Com esse novo intervalo de amostra $k = 81\,859$, não se tem mais regiões autorepresentadas. Portanto, basta selecionar-se 17 regiões não autorepresentadas, com probabilidade de seleção proporcional ao número de domicílios. Para isso, retiram-se da listagem as 3 regiões autorepresentadas e seleciona-se um ponto de partida θ no intervalo dos números inteiros de 1 a 81 859.

Seja $\theta = 14\,191$ o número selecionado.

Esse ponto, locado na coluna "Número acumulado" corresponde à região 6. Somando k = 81 859 a $\theta = 14\,191$, obtém-se 95 050 que, locado na mesma coluna de acumulado corresponde à região 20. Continuando a somar k = 81 859, vão sendo obtidas as demais regiões não autorepresentadas. Os números dessas 17 regiões são:

6, 20, 34, 45, 56, 63, 72, 78, 82, 85, 86, 88, 90, 92, 95, 96 e 98.

Seleção dos domicílios:

Para as regiões autorepresentadas

Considere-se a expressão,

$$mP_i \frac{n_i}{N_i} = \frac{n}{N}$$

que caracteriza a amostra autoponderada com probabilidade desigual de seleção das unidades primárias. Para as autorepresentadas, a probabilidade da UP pertencer a amostra é $mP_i = 1$, *donde* $\frac{n_i}{N_i} = \frac{n}{N}$, ou seja, a fração de amostragem de 2.º estágio é $\frac{n}{N} = \frac{380}{1717200} = 0{,}000221$. Na região autorepresentada 73, o número esperado de domicílios é 0,000221 (86300) = 19; na região autorepresentada 100 é 0,000221 (103000) = 23 e na região autorepresentada 101 o número esperado de domicílios é 0,000221 (136300) = 30. Ao todo, a subamostra nas regiões autorepresentadas abrange 72 domicílios.

Para as regiões não autorepresentadas

Para as regiões não autorepresentadas, restam 380 — 72 = 308 domicílios, que distribuídos nas 17 regiões da amostra, dão, em média, 18 domicílios por região.

A seleção dos domicílios é feita com probabilidade igual de seleção. Observe-se que basta listar os domicílios das regiões da amostra, evitando-se o trabalho de listagem dos 1717200 domicílios.

A seleção de domicílios também pode ser feita sistematicamente. Por exemplo, considere-se a região 73. Feita a listagem dos 86300 domicílios, o intervalo de amostra é $\frac{86300}{19} = 4542$. Seleciona-se um número inteiro no intervalo 1 a 4542 como ponto de partida θ e, depois, vai-se somando 4542 a esse número e aos seguintes, até obter 19 números. Esses números correspondem aos domicílios da amostra.

13 – EXERCÍCIO

Achar as expressões do Estimador de Razão, nos moldes do Capítulo 5 mas com probabilidade desigual de seleção.

CAPÍTULO 7 — Amostragem de Conglomerados em 2 — estágios. Controle da variação de tamanho: estratificação das unidades primárias com probabilidade desigual de seleção.

1 — INTRODUÇÃO

A estratificação das UP é feita grupando em um mesmo estrato as unidades primárias de tamanhos (ou medidas de tamanho) aproximadamente iguais. A seleção das UP, dentro de cada estrato, é feita com probabilidade proporcional ao tamanho (ou a uma medida de tamanho).

O processo para achar os estimadores é simples. Basta considerar as expressões do Capítulo 6 e adaptá-las a um estrato genérico h, acrescentando aos símbolos um índice h.

2 — ESTIMADOR NÃO TENDENCIOSO DE Y

Recorde-se que o estimador de Y com probabilidade desigual de seleção e sem estratificação é:

$$y^{*P}_{Ac2} = \frac{1}{m} \sum_{i=1}^{m} N'_i \frac{P'_i}{\bar{y}_i}$$

No estrato h, o estimador de Y_h, total do estrato, é:

$$y^{*P}_{h.Ac2} = \frac{1}{m_h} \sum_{i=1}^{m_h} N'_{hi} \frac{\bar{y}_{hi}}{P'_{hi}}$$

Conseqüentemente, o estimador de Y é:

$$y_{Ac2}^{*P.Ets} = \sum_{h=1}^{L} y_{h.Ac2}^{*P} = \sum_{h=1}^{L} \frac{1}{m_h} \sum_{i=1}^{m_h} N'_{hi} \frac{y_{hi}}{P'_{hi}}$$

3 – VARIÂNCIA DE $y_{Ac2}^{*P.Est}$

Recorde-se que a variância de y_{Ac2}^{*P} é:

$$V(y_{Ac2}^{*P}) = \frac{1}{m} \sum_{i=1}^{M} \left(\frac{Y_i}{P_i} - Y \right)^2 P_i + \frac{1}{m} \sum_{i=1}^{M} \frac{N_i^2}{P_i} \frac{N_i - n_i}{N_i} \frac{S_2^i}{n_i}$$

No estrato h, a variância de $y_{h.Ac2}^{*P}$ é, então:

$$V(y_{h.Ac2}^{*P.Est}) = \frac{1}{m_h} \sum_{i=1}^{M_h} \left(\frac{Y_{hi}}{P_{hi}} - Y_h \right)^2 P_{hi} +$$

$$+ \frac{1}{m_h} \sum_{i=1}^{M_h} \frac{N_{hi}^2}{P_{hi}} \frac{N_{hi} - n_{hi}}{N_{hi}} \frac{S_{hi}^2}{n_{hi}}$$

Conseqüentemente, a variância de $y_{Ac2}^{*P.Est}$ é:

$$V(y_{Ac2}^{*P.Est}) = \sum_{h=1}^{L} V(y_{h.Ac2}^{*P})$$

4 – ESTIMADOR NÃO TENDENCIOSO DE $V(y_{Ac2}^{*P.Est})$

Da expressão,

$$V(y_{Ac2}^{*P}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{N'_i \bar{y}_i}{P'_i} - y_{Ac2}^{*P} \right)^2$$

obtém-se,

$$V(y_{h.Ac2}^{*P}) = \frac{1}{m_h(m_h-1)} \sum_{i=1}^{m_h} \left(\frac{N'_{hi} \bar{y}_{hi}}{P'_{hi}} - y_{h.Ac2}^{*P} \right)^2$$

donde,

$$\hat{V}\left(y_{Ac2}^{*P.Est}\right) = \sum_{h=1}^{L} \frac{1}{m_h\,(m_h - 1)} \sum_{i=1}^{m_h} \left(\frac{N'_{hi}\,\bar{y}_{hi}}{P'_{hi}} - y_{h.Ac2}^{*P} \right)^2$$

5 – AMOSTRA AUTOPONDERADA

A probabilidade de uma *UP* do estrato h pertencer a amostra é:

$$m_h\,P_{hi}\,\frac{n_h}{N_{hi}}$$

Essa probabilidade pode ser constante no estrato mas variando de estrato para estrato,

$$m_h\,P_{hi}\,\frac{n_{hi}}{N_{hi}} = \frac{n_h}{N_h}\;(h = 1,\ 2,\ \ldots\ L)$$

ou ser constante para todos os estratos,

$$m_h\,P_{hi}\,\frac{n_{hi}}{N_{hi}} = \frac{n}{N}\;(h = 1,\ 2,\ \ldots\ L)$$

No primeiro caso, a amostra é autoponderada no estrato e no segundo caso é autoponderada em geral.

5.1 – *Exercícios*

a) Achar as expressões de $y_{Ac2}^{*P.Est}$, da respectiva variância e do estimador da variância, para o caso de amostra autoponderada.

b) Achar as expressões dos estimadores, conjugando a estratificação com o estimador de razão e probabilidade desigual de seleção.

6 – ESTRATOS GRUPADOS

6.1 – *Introdução*

Em pesquisa de âmbito nacional, pode ocorrer que se façam tantos estratos que basta selecionar uma *UP* em cada estrato. O problema que ocorre nesse caso, é a impossibilidade de estimar a variância entre as *UP* dentro de cada estrato.

A técnica dos "estratos grupados" consiste em grupar os estratos em pares, formando novos estratos. Desse modo, já se tem 2 observações nos novos estratos, permitindo a estimação da variância entre as *UP*.

6.2 – *Configuração da amostra*

Denomine-se estrato-componente E_{h1} e estrato-componente E_{h2} os estratos que compõem o novo estrato E_h $(h = 1, 2, \ldots L/2)$

E_h

E_{h1}	E_{2h}
UP_{h11}	UP_{h21}
$\vdots$	$\vdots$
$UP_{h1M_{h1}}$	$UP_{h2M_{h2}}$

O estrato-componente E_{h1} contém as unidades primárias $UP_{h11} \ldots UP_{h1M_{h1}}$ e o estrato-componente E_{h2} contém as unidades primárias $UP_{h21} \ldots UP_{h2M_{h2}}$.

6.2.1 – Parâmetros

No estrato-componente E_{h1}:

N.º de unidades primárias: M_{h1}

Probabilidade de seleção da unidade primária U_{h1i}: P_{h1i}

N.º de unidades secundárias em U_{h1i}: N_{h1i}

Total de y em U_{h1i}: Y_{h1i}

Total de y em E_{h1}:

$$Y_{h1} = \sum_{i=1}^{M_{h1}} Y_{h1i}$$

De modo análogo, definem-se parâmetros para o estrato-componente E_{h2}

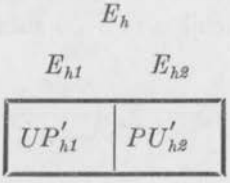
Conseqüentemente, no estrato E_h $(h = 1, 2, \ldots L/2)$ tem-se:

N.º de unidades primárias: $M_h = M_{h1} + H_{h2}$

Total de y: $Y_h = Y_{h1} + Y_{h2}$

6.2.2 — *Seleção da amostra*

Selecione-se uma *UP* em cada estrato-componente, de acordo com as probabilidades de seleção. Represente-se por UP'_{h1} e UP'_h as unidades selecionadas em E_h $(h = 1, 2, \ldots L/2)$



Desse modo, obtém-se a amostra de 1.º estágio:

$$\{(UP_{11} ; UP_{12}) \ldots (UP_{L/2,1} ; UP_{L/2,2})\}$$

A amostra de 2.º estágio é obtida selecionando n'_{h1} unidades secundárias em UP_{h1} com probabilidade igual de seleção. Seja $\bar{y}_{h1}$ a média de y nessa amostra. De modo análogo, selecionam-se n'_{h2} unidades secundárias na UP'_{h2} com média $\bar{y}_{h2}$. Repete-se o procedimento para $h = 1, 2, \ldots L/2$.

6.3 — *Estimadores*

A partir da expressão do estimador não tendencioso do total com probabilidade desigual de seleção, obtém-se:

a) Estimador não tendencioso do total Y_{h1} de E_{h1}:

$$y^{*}_{h2.\,Ac2} = \frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}}$$

b) Estimador não tendencioso do total Y_{h2}:

$$y^{*}_{h2.\,Ac2} = \frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h2}}$$

c) Estimador não tendencioso do total Y_h de E_h:

$$y^{*}_{h.\,Ac2} = y^{*}_{h1.\,Ac2} + y^{*}_{h2.\,Ac2} = \frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}} + \frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h2}}$$

6.4 — *Teorema*

Um estimador não tendencioso do total geral Y é:

$$y^{*G}_{Ac2} = \sum_{h=1}^{L/2} y^{*}_{h.\,Ac2} = \sum_{h=1}^{L/2} \left(\frac{N'_{h1}\,\bar{y}_{h1}}{N'_{h1}} + \frac{N'_{h2}\,\bar{y}_{h2}}{N'_{h2}} \right)$$

Prova

Imediata

6.5 — *Variância de* y^{*G}_{Ac2}

Considerando a independência das amostras nos estratos-componentes, tem-se:

$$V(y^{*}_{h.\,Ac2}) = V(y^{*}_{h1.\,Ac2}) + V(y^{*}_{h.\,Ac2})$$

donde,

$$V(y^{*G}_{Ac2}) = \sum_{h=1}^{L/2} V(y^{*}_{h.\,Ac2}) =$$

$$= \sum_{h=1}^{L/2} V(y^{*}_{h1.\,Ac2}) + \sum_{h=1}^{L/2} V(y^{*}_{h2.\,Ac2} =$$

$$= \sum_{h=1}^{L/2} V\left(\frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}}\right) + \sum_{h=1}^{L/2} V\left(\frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h2}}\right)$$

6.6 – *Teorema*

Um estimador de $V(y^{*G}_{Ac2})$ é:

$$\hat{V}(y^{*G}_{Ac2}) = \sum_{h=1}^{L/2} \left(\frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}} + \frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h2}}\right)^2$$

com tendenciosidade,

$$T = \sum_{h=1}^{L/2} (Y_{h1} - Y_{h2})^2$$

que se anula quando são iguais os totais dos estratos-componentes em cada par.

Prova

$$E[\hat{V}(y^{*G}_{Ac2})] = \sum_{h=1}^{L/2} E\left(\frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}} - \frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h2}}\right)^2 =$$

$$= \sum_{h=1}^{L/2} \left[E\left(\frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}} - Y_{h1}\right) - \left(\frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h2}} - Y_{h2}\right) + (Y_{h1} - Y_{h2})\right] =$$

$$= \sum_{h=1}^{L/2} \left[E\left(\frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}} - Y_{h1}\right)^2 + E\left(\frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h1}} - Y_{h2}\right)^2 + (Y_{h1} - Y_{h2})^2\right]$$

posto que se anulam as expectâncias dos duplos produtos. Conseqüentemente,

$$E[\hat{V}(y^{*G}_{Ac2})] = \sum_{h=1}^{L/2} V\left(\frac{N'_{h1}\,\bar{y}_{h1}}{P'_{h1}}\right) + \sum_{h=1}^{L/2} V\left(\frac{N'_{h2}\,\bar{y}_{h2}}{P'_{h2}}\right) +$$

$$+ \sum_{h=1}^{L/2} (Y_{h1} - Y_{h2})^2 = V(y^{*G}_{Ac2}) + \sum_{h=1}^{L/2} (Y_{h1} - Y_{h2})^2$$

Se $Y_{h1} = Y_{h2}$ $(h = 1, 2, \ldots L/2)$

$$E[\hat{V}(y^{*G}_{Ac2})] = V(y^{*G}_{Ac2})$$

6.7 — *Estratos grupados, com amostra autoponderada*

A probabilidade de seleção de uma *US* em E_{h1} é

$$P_{h1}\frac{n_{h1}}{N_{h1}} \text{ e em } E_{h2} \text{ é } P_{h2}\frac{n_{h2}}{N_{h2}}$$

Se essas probabilidades $(h = 1, 2, \ldots L/2)$ são iguais a fração geral de amostragem $\frac{n}{N}$, tem-se:

$$P_{h1}\frac{n_{h1}}{N_{h1}} = \frac{n}{N}$$

e

$$P_{h2}\frac{n_{h2}}{N_{h2}} = \frac{n}{N}$$

Nesse caso, as expressões de y^{*G}_{Ac2} e $\hat{V}(y^{Ac2}_{*G})$ sofrem as seguintes modificações. Representando por y_{h1} o total da amostra em UP_{h1} e por y_{h2} o total em UP_{h2}, pode-se escrever:

$$y^{*G}_{Ac2} = \sum_{h=1}^{L/2}\left(\frac{N'_{h1}\,y_{h1}}{P'_{h1}\,n'_{h1}} + \frac{N'_{h2}\,y_{h2}}{P'_{h2}\,n'_{h2}}\right) = \frac{N}{n}\sum_{h=1}^{L/2}(y_{h1} + y_{h2})$$

e para $\hat{V}(y^{*G}_{Ac2})$,

$$V(y^{*G}_{Ac2}) = \sum_{h=1}^{L/2} \left(\frac{N'_{h1}\, y_{h1}}{P'_{h1}\, n'_{h1}} - \frac{N'_{h2}\, y_{h2}}{P'_{h2}\, n'_{h2}} \right)^2 = \left(\frac{N}{n} \right)^2 \sum_{h=1}^{L/2} (y_{h1} - y_{h2})^2$$

6.8 — *Exemplo*

Para estimar o número de moradores de determinada localidade, estabeleceu-se uma amostra de conglomerados em 2 — estágios, em que as unidades primárias são os setores censitários e as unidades secundárias os domicílios. A cada domicílio, associou-se o número de moradores.

Os setores foram estratificados em 20 estratos. Depois os estratos foram grupados em pares de modo que os estratos-componentes tivessem populações de valor próximo, de acordo com dados do último Censo. De cada estrato-componente selecionou-se um Setor, com probabilidade de seleção proporcional à população.

Em cada Setor da amostra, selecionou-se uma subamostra de domicílios, com tamanho suficiente para gerar uma amostra auto-ponderada, com fração de amostragem em cada estrato igual a fração geral de amostragem de 1%. Desse modo, a fração de amostragem de 2.º estágio é:

$$\frac{n_{h1}}{N_{h1}} = \frac{n}{P_{h1}\, N} = \frac{1}{100\, P_{h1}} \text{ em } E_{h1}$$

e,

$$\frac{N_{h2}}{n_{h2}} = \frac{n}{P_{h2}\, N} = \frac{1}{100\, P_{h2}} \text{ em } E_{h2}$$

Em E_{h1}, por exemplo, selecionou-se um setor com probabilidade proporcional à população do Censo. Desse modo, foi selecionado o setor com população de 210 moradores. Nesse setor, foi selecionada uma subamostra de domicílios, com fração de amostragem 56/210.

ESTRATOS	POPULAÇÃO DOS ESTRATOS COMPONENTES: DE E_{h1}	DE E_{h2}	POPULAÇÃO DO SETOR: SELECIONADO DE U'_{h1}	DA U'_{h2}	FRAÇÃO DE AMOSTRAGEM DE 2-º ESTÁGIO: $\frac{n'_{h1}}{N'_{h1}}$	$\frac{n'_{h2}}{N'_{h2}}$
1	5 600	5 100	210	180	56/210	51/180
2	3 400	2 900	120	110	34/120	29/110
3	4 500	4 100	180	160	45/180	41/160
4	7 300	7 000	340	320	73/340	70/320
5	5 400	5 100	240	200	54/240	51/200
6	2 300	2 000	90	80	23/90	20/80
7	6 300	6 200	310	280	63/310	62/280
8	4 600	4 000	190	130	46/190	40/130
9	3 100	3 000	140	100	31/140	30/100
10	2 200	1 900	100	80	22/100	19/80

Feita a seleção das amostras de domicílios em todos os setores selecionados no 1.º estágio, procedeu-se à contagem do número de moradores nesses setores, obtendo-se o seguinte quadro:

ESTRATOS	N.º OBSERVADO DE MORADORES NO DOMICÍLIOS DA SUBAMOSTRA EM UP'_{h1} (y_{h1})	EM UP'_{h2} (y_{h2})	$y_{h1} + y_{h2}$	$(y_{h1} - y_{h2})^2$
1	71	65	136	36
2	40	33	73	49
3	48	52	100	16
4	79	83	162	16
5	60	56	116	16
6	35	22	57	169
7	65	71	136	36
8	56	50	106	36
9	37	39	76	4
10	28	21	50	49
Soma			1 012	427

Estimativa do número atual de moradores:

$$y^{*G}_{Ac2} = 100(1012) = 101200 \quad \text{moradores}$$

Estimativa da variância de y^{Ac2}_{*G}

$$V(y^{*G}_{Ac2}) = (100)^2 (427)$$

$$\hat{\gamma}(y^{*G}_{Ac2}) = \frac{\sqrt{427}}{1012} = 0{,}0204 \quad \text{ou} \quad 2{,}04\%$$

CAPÍTULO 8 — Amostragem de Conglomerados em 3 — estágios. Amostras replicadas.

1 — INTRODUÇÃO

A dificuldade de cadastramento para seleção da amostra se reduz a medida em que aumenta o número de estágios.

Suponha-se que se deseja selecionar uma amostra nacional de domicílios. Em princípio, a amostra pode ser feita em 2 — estágios, considerando os Setores Censitários como unidades primárias constituídas de unidades secundárias, os domicílios. No entanto, o processo pode se tornar muito trabalhoso, considerando o grande número de Setores Censitários e a necessidade de serem todos relacionados para, posteriormente, serem alguns selecionados para a amostra.

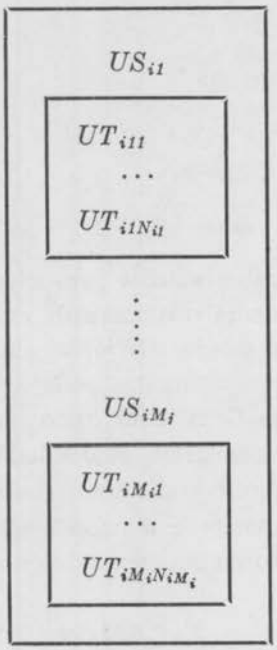
Com uma amostra m 3 — estágios, definem-se os municípios como unidades primárias, constituídas das unidades secundárias — os Setores Censitários — os quais, por sua vez, são constituídos das unidades terciárias — os domicílios. Inicialmente selecionam-se municípios, de modo que a listagem dos Setores Censitários só é feita nos municípios da amostra.

No entanto, a medida em que aumenta o número de estágios, mais se torna complicada a expressão da variância do estimador, o que induz à procura de um processo mais simples de cálculo da variância. Esse processo de cálculo é chamado "amostras replicadas" e será estudado neste capítulo para aplicação em amostras autoponderadas em vários estágios.

2 – CONFIGURAÇÃO DA AMOSTRA

Considere-se a i-ésima unidade primária:

UP_i $\qquad (i = 1, 2, \ldots R)$



Associe-se à unidade terciária UT_{ijk} a observação Y_{ijk}.

Total da US_{ij} na UP_i: $Y_{ij} = \sum_{k=1}^{N_{ij}} Y_{ijk}$

Média por UT na US_{ij}: $\overline{Y}_{ij} = \frac{Y_{ij}}{N_{ij}}$

Total da UP_i: $Y_i = \sum_{j=1}^{M_i} Y_{ij}$

Média por US na UP_i: $\overline{Y}_i = \frac{Y_i}{M_i}$

Média por UT na UP_i: $\overline{\overline{Y}}_i = \frac{Y_i}{N_i}$ com $N_i = \sum_{j=1}^{M_i} N_{ij}$

Total geral: $Y = \sum_{i=1}^{R} Y_i$

Média por UP: $\overline{Y} = \frac{Y}{R}$

Média por US: $\overline{\overline{Y}} = \frac{Y}{M}$ com $M = \sum_{i=1}^{R} M_i$

Média por UT: $\overline{\overline{\overline{Y}}} = \frac{Y}{N}$ com $N = \sum_{i=1}^{R} \sum_{j=1}^{M_i} N_{ij} = \sum_{i=1}^{R} N_i$

Quanto às variâncias, tem-se:

Variância entre as UP: $S_e^2 = \frac{1}{R-1} \sum_{i=1}^{R} (Y_i - \overline{Y})^2$

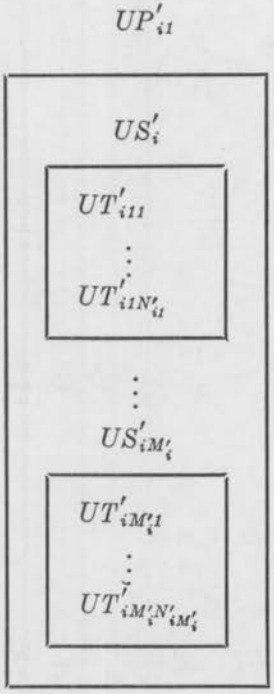
Variância entre as US na UP_i: $S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (Y_{ij} - \overline{Y}_i)^2$

Variância entre as UT na US_{ij}:

$$S_{ij}^2 = \frac{1}{N_{ij} - 1} \sum_{k=1}^{N_{ij}} (Y_{ijk} - \overline{Y}_i)^2$$

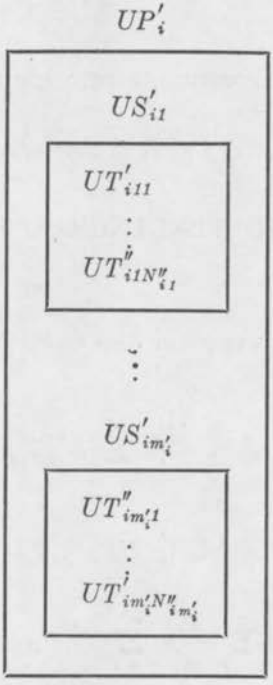
Amostra de 1.º estágio:

Selecione-se uma *Als* de r unidades primárias. Seja UP'_i e i-ésima UP da amostra. Os parâmetros anteriormente definidos são, agora, variáveis aleatórias.



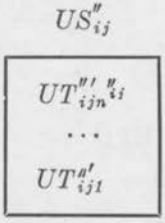
Amostra de 2.º estágio:

De cada UP' da amostra, selecione-se uma Als de unidades secundárias. Desse modo, na UP'_i seleciona-se uma Als de m'_j unidades secundárias.



Amostra de 3.º estágio:

De cada US'' da amostra de unidades secundárias, selecione-se uma Als de unidades terciárias. Assim, na US_{ij} selecione-se uma Als de n''_{ij} unidades terciárias.



$(i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, m'_i)$

Associado a UT_{ijk} a observação y_{ijk} obtém-se:

Tamanho da amostra: n''_{ij}

Média da amostra: $\bar{y}_{ij} = \frac{1}{n''_{ij}} \sum_{k=1}^{n''_{ij}} y_{ijk}$

Variância da amostra: $s \quad s^2_{ij} = \frac{1}{n''_{ij} - 1} \sum_{k=1}^{n''_{ij}} (y_{ijk} - \bar{y}_{ij})^2$

A amostra final, é constituída pelo conjunto

$$\{y_{ijk} \mid i = 1, 2, \ldots r; j = 1, 2, \ldots m'_i; k = 1, 2, \ldots n''_{ij}\}$$

3 – ESTIMADOR NÃO TENDENCIOSO DO TOTAL Y

Considerando o processo em 2 – estágios, o estimador do total da UP'_i é:

$$y^*_i = \frac{M'_i}{m'_i} \sum_{j=1}^{m'_i} N''_{ij} \bar{y}_{ij}$$

donde o total da amostra $\{UP'_1, UP'_2, \ldots UP'_r\}$ é,

$$\sum_{i=1}^{r} \frac{M'_i}{m'_i} \sum_{j=1}^{m'_i} N''_{ii} \bar{y}_{ij}$$

Conseqüentemente, o estimador de Y é,

$$y^*_{Ac3} = \frac{R}{r} \sum_{i=1}^{r} \frac{M'_i}{m'_i} \sum_{j=1}^{m'_i} N''_{ij} \bar{y}_{ij}$$

São imediatos os estimadores que se seguem.

3.1 – *Estimador não tendencioso da média por* UP:

$$\bar{y}_{Ac3} = \frac{y^*_{Ac3}}{R} = \frac{1}{r} \sum_{i=1}^{r} \frac{M'_i}{m'_i} \sum_{j=1}^{m'_i} N''_{ij} \bar{y}_{ij}$$

3.2 — *Estimador não tendencioso da média por* US:

$$\overline{\overline{y}}_{Ac3} = \frac{y^*_{Ac3}}{M} = \frac{R}{rM} \sum_{i=1}^{r} \frac{M'_i}{w'_i} \sum_{j=1}^{m'_i} N''_{ij} \bar{y}_{ij}$$

$$\text{com } M = \sum_{i=1}^{R} M_i$$

Pondo, ainda, $\overline{M} = \frac{M}{R}$ número médio de *US* por *UP*, pode-se escrever:

$$\overline{\overline{y}}_{Ac3} = \frac{1}{r\overline{M}} \sum_{i=1}^{r} \frac{M'_i}{m'_i} \sum_{j=1}^{m'_i} N''_{ij} \bar{y}_{ij}$$

3.3 — *Estimador não tendencioso da média por* UT:

$$\overline{\overline{\overline{y}}}_{Ac3} = \frac{y^*_{Ac3}}{N} = \frac{R}{rN} \sum_{i=1}^{r} \frac{M'_i}{m'_i} \sum_{j=1}^{m'_i} N''_{ij} \bar{y}_{ij}$$

Pondo, $\overline{N} = \frac{N}{R}$ número médio de *UT* por *UP*, tem-se:

$$\overline{\overline{\overline{y}}}_{Ac3} = \frac{1}{r\overline{N}} \sum_{i=1}^{r} \frac{M'_i}{m'_i} \sum_{j=1}^{m'_i} N''_{ij} \bar{y}_{ij}$$

4 — AMOSTRA AUTOPONDERADA

A probabilidade de uma *UT* pertencer à amostra é:

$$\frac{r}{R} \cdot \frac{m_i}{M_i} \cdot \frac{n_{ij}}{N_{ij}}$$

Fixada uma fração geral de amostragem, $\frac{n}{N}$, a amostra é autoponderada se:

$$\frac{r}{R} \frac{m_i}{M_i} \cdot \frac{n_{ji}}{N_{ij}} = \frac{n}{N}$$

Nesse caso, o estimador não tendencioso y^{*}_{Ac3} pode ser escrito:

$$y^{*}_{Ac3} = \frac{N}{n} \sum_{i=1}^{r} \sum_{j=1}^{m'_i} \sum_{k=1}^{n''_{ij}} y_{ijk} = \frac{N}{n}\, y$$

com $y = \sum_{i=1}^{r} \sum_{j=1}^{m'_i} \sum_{k=1}^{n''_{ij}} y_{ijk}$ total da amostra.

5 – FRAÇÕES DE AMOSTRAGEM CONSTANTES NOS 3 ESTÁGIOS

Suponham-se fixadas as frações de amostragem de 1.º e 2.º estágios, nos valores f_1 e f_2 respectivamente, com amostra autoponderada.

Nesse caso, a probabilidade que uma UT_{ijk} pertença à amostra é:

$$f_1 f_2 \frac{n_{ij}}{N_{ij}} = \frac{n}{N} = f$$

Então, a fração de amostragem de 3.º estágio é constante e igual a,

$$f_3 = \frac{f}{f_1 f_2}$$

6 – EXEMPLO

Considerem-se 100 escolas de 1.º grau, com aproximadamente 21 000 alunos de 1.ª série. Trata-se de selecionar uma amostra de alunos em 3 – estágios: seleção de escolas, seleção de turmas e seleção de alunos.

A finalidade é estimar a despesa média com condução para a escola, por aluno.

Pretende-se que a amostra tenha 100 alunos, o que corresponde à fração geral de amostragem:

$$\frac{n}{N} = \frac{100}{21000} = \frac{1}{210}$$

No 1.º estágio, fixou-se a fração de amostragem em $f_1 = \frac{1}{10}$; no 2.º estágio em $f_2 = \frac{1}{3}$. No 3.º estágio, a fração de amostragem resulta da condição de amostra autoponderada,

$$f_1 f_2 f_3 = \frac{n}{N}$$

$$\text{donde } f_3 = \frac{30}{210} = \frac{1}{7}$$

O quadro resultante da amostra é o seguinte:

ESCOLAS DA AMOSTRA	N.º DE TURMAS (M_i')	N.º DE TURMAS DA AMOSTRA (m_i')	N.º DE ALUNOS NAS TURMAS DA AMOSTRA $\sum_{j=1}^{m_i'} N_{ij}''$	TAMANHO DA AMOSTRA DE ALUNOS. $\sum_{j=1}^{m_i'} n_{ij}''$	DESPESA MENSAL TOTAL. $\sum_{i=1}^{m_i'} \sum_{j=1}^{n_{ij}''} y_{ijk}$
1	7	2	65	9	20
2	8	3	84	12	28
3	9	3	92	13	26
4	6	2	58	8	18
5	6	2	59	8	15
6	9	3	89	13	23
7	8	3	98	14	34
8	7	2	60	9	20
9	6	2	62	9	25
10	5	2	59	8	13
Soma	71	24	726	103	222

Estimativa da despesa média por aluno

$$\bar{\bar{y}}_{AcS} = \frac{1}{n}\, y = \frac{1}{10}\, 222 = 22{,}2$$

Estimativa da variância de $V(\bar{\bar{\bar{y}}}_{Ac3})$

$$V(\bar{\bar{\bar{y}}}_{Ac3}) = r\left(\frac{1}{n^2}\right)\frac{1}{r-1}\sum_{i=1}^{r}\left[\sum_{j=1}^{m'_i}\sum_{k=1}^{n''_{ij}} y_{ijk} - \frac{1}{r}\,y\right]^2 =$$

$$= 10\left(\frac{1}{103}\right)\frac{1}{9}\,(356{,}72) = 3{,}848$$

$$\sqrt{\hat{V}(\bar{\bar{\bar{y}}})} = 1962$$

$$\gamma(\bar{\bar{\bar{y}}}_{Ac3}) = 0{,}088 \text{ ou } 8{,}8\%$$

7 – SELEÇÃO COM PROBABILIDADE DESIGUAL

7.1 – *Configuração da amostra*

Seja P_i a probabilidade de seleção da unidade primária UP_i $(i = 1, 2, \ldots R)$. Selecionem-se r unidades primárias, de acordo com essas probabilidades. De cada unidade primária UP'_i da amostra, selecionem-se m'_i unidades secundárias, tendo a unidade secundária US_{ij} probabilidade de seleção P_{ij}. Finalmente, da unidade secundária US''_{ij} da amostra, selecionem-se n''_{ij} unidades terciárias, com probabilidade igual de seleção.

7.2 – *Estimador não tendencioso de Y*

Considerando o processo em 2 – estágios, o estimador do total da UP'_i é:

$$y_i^* = \frac{1}{m'_i}\sum_{j=1}^{m'_i}\frac{N''_{ij}\,\bar{y}_{ij}}{P''_{ij}}$$

donde, o estimador não tendencioso de Y é,

$$y_{Ac3}^{P*} = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{m'_i P'_i}\sum_{j=1}^{m'_i}\frac{N''_{ij}\,\bar{y}_{ij}}{P''_{ij}}$$

7.3 — *Exercício*

Achar as expressões dos estimadores da média por *UT*, por *US* e por *UP*.

7.4 — *Amostra autoponderada*

A probabilidade de UT_{ijk} pertencer a amostra é:

$$rP_i\, m_i\, P_{ij}\, \frac{n_{ij}}{N_{ij}}$$

A amostra é autoponderada se essa probabilidade é constante e igual a fração geral de amostragem, isto é, se:

$$rP_i\, m_i\, P_{ij}\, \frac{n_{ij}}{N_{ij}} = \frac{n}{N}$$

Nesse caso, o estimador y^{*P}_{Ac3} assume a mesma forma do estimador y^{*}_{Ac3} :

$$y^{*P}_{Ac3} = \frac{N}{n}\, y$$

7.5 — *Tamanho constante da amostra nos 3 estágios*

Suponham-se fixados o número de unidades primárias e o número de unidades secundárias. Trata-se de achar o número de unidades terciárias que conduza à amostra autoponderada.

A probabilidade de seleção de uma UT_{ijk} que pertença a amostra autoponderada é:

$$rP'_i\, mP''_{ij}\, \frac{n''_{ij}}{N''_{ij}} = \frac{n}{N}$$

donde,

$$E(n''_{ij}) = \underset{UP'_i}{E}\left[E\left(\frac{nN''_{ij}}{NrmP'_i} \mid UP'_i \; fix.\right)\right] =$$

$$= \underset{UP'_i}{E}\left[\frac{n}{NrmP'_i}\sum_{j=1}^{M'_i}\frac{N'_{ij}}{P'_{ij}}P'_{ij}\right] =$$

$$= \underset{UP'_i}{E}\left[\frac{nN'_i}{NrmP'_i}\right] =$$

$$= \frac{n}{Nrm}\sum_{i=1}^{R}\frac{N_i}{P_i}P_i = \frac{nN}{Nrm} = \frac{n}{rm}$$

Portanto, fixados r e m, o número de unidades terciárias, em média, é:

$$\bar{n} = \frac{n}{rm}$$

7.6 — *Exemplo*

Considere-se o Exemplo 11 do Capítulo 6. Recorde-se que a localidade está dividida em 101 regiões, cada uma dividida em subregiões, dando um total de 3232 subregiões ou uma média de $\frac{3232}{101} = 32$ subregiões por região. As subregiões, por sua vez, estão divididas em domicílios, com um total de 1717200 domicílios. Para obter-se uma amostra de 380 domicílios foram selecionadas 20 regiões, com probabilidade proporcional ao número de domicílios. As regiões de números 73, 100 e 101 são autorepresentadas e deles, foram selecionados 19, 23 e 30 domicílios, respectivamente, com um total de 72 domicílios.

Os 308 domicílios restantes, foram distribuídos nas regiões não autorepresentadas, com uma média de $\frac{308}{17} = 18$ domicílios por região. Para completar 380 domicílios, foram selecionados 20 domicílios na última região da amostra.

O problema, agora, é selecionar uma amostra em 3 — estágios: regiões, subregiões e domicílios, mantendo 380 domicílios na amostra.

Com as 20 regiões já selecionadas e fixando 3 subregiões por região, o número de domicílios por subregião não autorepresentada é:

$$n = \frac{308}{17\,(3)} \doteq 6$$

Nas regiões autorepresentadas, o número de domicílios já foi fixado. Desse modo, tem-se a seguinte distribuição da amostra:

a) Na região autorepresentada 73, 3 subregiões com 6 domicílios cada.

b) Na região autorepresentada 100, 3 subregiões com 8 domicílios cada.

c) Na região autorepresentada 101, 3 subregiões com 10 domicílios cada.

d) Em 16 regiões não autorepresentadas, 3 subregiões, com 6 domicílios cada.

e) Em uma região não autorepresentada, 3 subregiões, com 7 domicílios cada.

Amostragem de 1.º estágio — Seleção de 20 regiões.

Suponham-se selecionadas as mesmas regiões do Exemplo 11 — Capítulo 6.

Amostragem de 2.º estágio — Seleção de subregiões.

Para ilustrar a possibilidade de variar a probabilidade de seleção, suponha-se a seleção com probabilidade proporcional à população.

Para exemplificar, considere-se a região autorepresentada 73. Esta região está dividida em 18 subregiões, de acordo com o seguinte quadro:

Região 73

SUBREGIÕES	POPULAÇÃO	ACUMULADO
73.1	3 400	3 400
73.2	2 100	5 500
73.3	3 000	8 500
73.4	1 800	10 300
73.5	4 100	14 400
73.6	1 200	15 600
73.7	3 400	19 000
73.8	2 000	21 000
73.9	2 400	23 400
73.10	1 700	25 100
73.11	5 000	30 100
73.12	1 100	31 200
73.13	3 200	34 400
73.14	2 300	36 700
73.15	1 900	38 600
73.16	5 100	43 700
73.17	1 300	45 000
73.18	2 400	47 400
Soma	47 400	—

O intervalo de amostra é $\frac{47400}{3} = 15800$. Selecionando-se um ponto de partida de 1 a 15800 obteve-se 12700 que corresponde à subregião 73.5; somando 15800 a 12700 obtém-se 28500 que corresponde à subregião 73.11; somando 15800 a 28500 obtém-se 44300 que corresponde à subregião 73.17. Portanto, na região 73 são selecionadas as subregiões 73.5, 73.11 e 73.17. De cada subregião da amostra selecionam-se 6 domicílios, com probabilidade igual de seleção e sem reposição. Procedendo-se de modo análogo para as demais regiões da amostra, obtém-se o quadro que se segue, onde consta a renda domiciliar.

REGIÕES DA AMOSTRA		N.º DE DOMICÍLIOS DA AMOSTRA	SOMA DAS RENDAS DOMICILIARES NA AMOSTRA (em S.M.)
Autoponderadas	Não autoponderadas		
73	—	18	112
100	—	24	408
101	—	30	570
	6	18	36
	20	18	50
	34	18	54
	45	18	65
	56	18	72
	63	18	79
	72	18	108
	78	18	117
	82	18	130
	85	18	144
	86	18	158
	88	18	171
	90	18	180
	92	18	216
	95	18	234
	96	18	252
	98	21	370
Soma		381	3526

Estimativa da renda média por domicílio:

$$\bar{\bar{y}}_{Ac3}^{P} = \frac{3526}{381} = 9{,}25$$

Estimativa da renda total da localidade:

$$y_{*P}^{Ac3} = 9{,}25\,(1717200) = 15884100$$

8 — AMOSTRAS REPLICADAS

Em vez de se selecionar uma amostra autoponderada com m unidades primárias e fração geral de amostragem $f = \frac{n}{N}$, selecio-

nam-se g grupos de $\frac{m}{g}$ unidades primárias, dos quais obtém-se amostras autoponderadas, com o mesmo número de estágios e fração de amostragem geral $\frac{f}{g}$, ou seja, com $\frac{n}{g}$ unidades da população em cada grupo. O objetivo desse processo é facilitar a estimação da variância no caso de vários estágios.

8.1 — *Teorema*

Seja y_i o total da amostra final no grupo i. Seja $\sum_{i=1}^{g} y_i$ o total dos g grupos.

Então $\bar{y}_{Rep} = \frac{y}{n}$ é estimador não tendencioso de $\bar{Y}$, média por unidade da população.

Prova

Considere-se o conjunto dos totais das amostras $\{y, y, \ldots y_g\}$. Esse conjunto pode ser entendido como uma *Als* de uma população de totais de $\frac{N}{k}\left(k = \frac{n}{g}\right)$ grupos de unidades. Sejam $Y_1, Y_2, \ldots Y_{N/k}$ esses totais. Então $\bar{y}_g = \frac{1}{g} \sum_{i=1}^{g} y_i$ é estimador não tendencioso de

$$k\bar{Y} = \frac{\sum_{i=1}^{N/k} Y_i}{N/k}$$

Conseqüentemente, $\bar{y}_{Rep} = \frac{y_g}{k}$ é estimador não tendencioso de $\bar{Y}$.

8.2 — *Variância de* $\bar{y}_{Rep}$

$$V(\bar{y}_{Rep}) = V\left(\frac{\bar{y}_g}{k}\right) = \frac{1}{k^2} V(\bar{y}_g)$$

Porém,

$$V(\bar{y}) = \frac{\frac{N}{k} - \frac{n}{k}}{\frac{N}{k}} \frac{S_g^2}{g} = \frac{N-n}{N} \frac{S_g^2}{g}$$

$$\text{onde} \quad S_g^2 = \frac{\sum_{i=1}^{N/k} Y_i^2 - \frac{\left(\sum_{i=1}^{N/k} Y_i\right)^2}{N/k}}{N/k - 1}$$

donde,

$$V(\bar{y}_{Rep}) = \frac{N-n}{N} \frac{S_g^2}{kn}$$

8.3 — *Teorema*

Um estimador não tendencioso de $V(\bar{y}_{Rep})$ é:

$$V(\bar{y}_{Rep}) = \frac{N-n}{N} \frac{s_g^2}{kn}$$

$$\text{onde} \quad s_g^2 = \frac{\sum_{i=1}^{g} Y_i^2 - \frac{\left(\sum_{i=1}^{g} y_i\right)^2}{g}}{g-1}$$

Prova

Imediata.

8.4 — *Exemplo*

Considere-se a seguinte modificação no Exemplo 11: em vez de selecionar-se uma de amostra de 1.º estágio de 20 regiões, autoponderada e com fração geral de amostragem,

$$f = \frac{380}{1717200}$$

formem-se $g = 4$ grupos de 5 regiões cada um. Em cada grupo, a fração geral de amostragem é $\frac{f}{g}$, ou seja, de cada grupo seleciona-se uma amostra de 95 domicílios, autoponderada.

Formação dos grupos de regiões.

Exemplifica-se com a formação do 1.º grupo. O intervalo de amostra para uma seleção sistemática é $\frac{1717200}{5} = 343400$. Selecionando um ponto de partida no intervalo 1 a 343440, obteve-se $\theta = 314836$. Os demais pontos são: 658276, 1001716, 1345156, 1688596. Esses pontos correspondem às regiões: 53, 78, 88, 96, 101.

Procedendo de modo análogo para formar os outros 3 grupos, cada um com 5 regiões, obteve-se:

Para o 2.º grupo, as regiões 36, 59, 73, 82, 87

Para o 3.º grupo, as regiões 38, 65, 76, 85, 89

Para o 4.º grupo, as regiões 42, 70, 81, 87, 94

Seleção das subregiões e dos domicílios.

Em cada região, foram selecionadas 3 subregiões, dando um total de 60 subregiões na amostra de 2.º estágio. Em cada subregião foram selecionados 6 domicílios mas uma região de cada grupo contribuiu com 2 subregiões com 8 domicílios e 1 subregião com 7 domicílios, de modo a dar 95 domicílios por grupo.

Desse modo, obteve-se:

GRUPO	REGIÕES	N.º DE DOMICÍLIOS NA AMOSTRA	SOMA DAS RENDAS DOS DOMICÍLIOS DA AMOSTRA (S.M)
1	53	18	72
	78	18	117
	88	18	171
	96	18	252
	101	23	500
	Soma	95	1112
2	36	18	63
	59	18	79
	73	18	117
	82	18	130
	87	23	249
	Soma	95	638
3	38	18	63
	65	18	108
	76	18	117
	85	18	144
	89	23	270
	Soma	95	702
4	42	18	65
	70	18	108
	81	18	130
	87	18	170
	94	23	340
	Soma	95	813

Estimativa da renda média domiciliar

$$\bar{y}_{Rep} = \frac{1112 + 638 + 802 + 813}{95\,(4)} = \frac{3265}{380} = 8{,}59$$

$$\sum_{i=1}^{4} y_i^2 = 1112^2 + 638^2 + 702^2 + 813^2 = 2797361$$

$$s^2 = \frac{2797361 - \frac{3265^2}{4}}{3} = 44101{,}6$$

$$\hat{V}(\bar{y}_{Rep}) = \frac{44101{,}6}{380(95)} = 1{,}222 \text{ donde } \hat{\gamma}(\bar{y}_{Rep}) = 0{,}129 \text{ ou } 12{,}9\%$$

Composto e impresso no
Centro de Serviços Gráficos
do IBGE, Rio de Janeiro, RJ

IBGE
DIRETORIA DE ADMINISTRAÇÃO
CENTRO DE SERVIÇOS GRÁFICOS