VIESADO (Hartley & Ross) NA AMOSTRAGEM DE CONGLOMERADOS QUE DIFEREM EM TAMANHO

BELMER GARCIA NEGRILLO

DISSERTAÇÃO APRE

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE EM

ESTATÍSTICA APLICADA

ORIENTADOR: PROF. DR. LINDO FAVA

BOLSA: ORGANIZAÇÃO DOS ESTADOS AMERICANOS

ABRIL DE 1974

SÃO PAULO

Dedico este trabalho

a meus pais

e

ã minha esposa

Í N D I C E

INTRODUÇÃO. I

CAPÍTULO I

1.1 - Estimador Razão na Amostragem Casual Simples 1

1.2 - Esperança Matemática e Variância do Estimador r. . . . 3

1.3 - Considerações Sobre a Magnitude do Viés do Estimador r 6

1.4 - Estimador $\bar{r}_H$ do Tipo Razão e Não Viesado de Hartley e Ross 9

1.5 - Comparação das Variâncias dos Estimadores r e $\bar{r}_H$ . . . 10

CAPÍTULO II

2.1 - Especificação da População – Parâmetros de Posição e de Variabilidade 13

2.2 - Amostragem de Conglomerados com Enumeração Completa dos seus Elementos 17

2.3 - Alguns Estimadores na Prática. 20

2.4 - O Estimador $\bar{\bar{y}}_H$ do Tipo Razão e Não Viesado 24

CONCLUSÕES. 29

BIBLIOGRAFIA. 31

I N T R O D U Ç Ã O

O problema de que tratamos nesta dissertação havia sido deixado em aberto pelo Prof. Dr. Lindo Fava em sua tese de Livre Docência Contribuição para o Estudo da Estimativa Razão. Foi-nos agora proposto para ser desenvolvido sob a forma de uma monografia de dissertação de Mestrado.

Para atingirmos o nosso objetivo, resolvemos dividir a monografia em dois capítulos. No Capítulo I, estudamos os principais resultados relativos aos estimadores razão e do tipo razão já consolidados na literatura pertinente ao assunto. Os resultados julgados relevantes foram sintetizados segundo as várias rubricas arroladas adiante.

No Capítulo II, ocupamo-nos propriamente com o problema proposto e após havermos conceituado a Amostragem de Conglomerados em um estágio e a notação a ser utilizada, os vários tipos de estimadores, viesados e não viesados e as respectivas precisões, demos ênfase ao estimador não viesado do tipo razão devido a Hartley e Ross. Acreditamos haver mostrado que esse estimador pode constituir-se num instrumento útil. Em alguns casos, ele permite uma avaliação aproximada da magnitude do viés e, em outros, dá-nos o acréscimo de precisão que pode haver em relação a estimadores mais simples, sabidamente não viesado.

Para terminar, gostaríamos de render nossas homenagens

aos que possibilitaram a concretização deste trabalho:

1. à *ORGANIZAÇÃO DOS ESTADOS AMERICANOS*, pela Bolsa de Estudos a nós proporcionada;
2. ao *PROF. DR. LINDO FAVA*, pela proposição do problema focalizado e pela orientação dele recebida;
3. ao Sr. *JOÃO BAPTISTA ESTEVES DE OLIVEIRA*, pelo trabalho de datilografia.

São Paulo, abril de 1974

Belmer Garcia Negrillo

C A P Í T U L O   1

1.1 - ESTIMADOR RAZÃO NA AMOSTRAGEM CASUAL SIMPLES

Consideremos uma população finita com N unidades de amostragem: U_1, U_2, ..., U_N. Em relação a esta população, podemos muitas vezes considerar simultaneamente os valores de duas variáveis: y e x. No caso de uma eunmeração completa dos elementos da população, denotamos os valores possíveis das varíaveis y e x por:

$$(Y_v, X_v), \ (v=1,2,\ldots,N).$$

Seja agora,

$$Y = \sum_{v=1}^{N} Y_v \tag{1.1.1}$$

a soma, ou o total, dos valores da variável y, e

$$X = \sum_{v=1}^{N} X_v, \tag{1.1.2}$$

a soma, ou o total, dos valores da variável auxiliar x. (Esse total,algumas vezes, é conhecido ou dado por uma pesquisa anterior.) Então,

$$R = \frac{Y}{X} = \frac{N\bar{Y}}{N\bar{X}} = \frac{\bar{Y}}{\bar{X}} \tag{1.1.3}$$

é o coeficiente angular da reta que passa pela origem e pelo baricentro $(\bar{Y}, \bar{X})$ da distribuição das variáveis y e x.

Em muitos casos, esse coeficiente angular tem um significado empírico imediato. Por exemplo, se a variável x denota áreas de fazendas e y as produções de um cereal nessas fazendas, a razão R será a produção média por unidade de área das N fazendas. Essa produção média por unidade de área é de compreensão e intuitiva por agrônomos, economistas e administradores em geral. Na prática, no entanto, o conhecimento desse parâmetro R, dado o tamanho N da população ou devido a razões outras de tempo e dinheiro, só é obtido através da análise dos valores das variáveis y e x de uma amostra casual a ser colhida da população originária.

Colhida uma amostra casual de tamanho $n < N$ da população originária, os n pares de valores observados das variáveis y e x, serão denotados por (y_i, x_i), $i=1,\dots,n$. Definiremos também a quantidade

$$r = \frac{\bar{y}}{\bar{x}} \qquad (1.1.4)$$

onde, $\bar{y}$ e $\bar{x}$ são as médias aritméticas dos valores das variáveis y e x na amostra obtida; r é denominado uma estatística ou um estimador do parâmetro R.

Do mesmo modo, sendo o total X conhecido,

$$\tilde{y}_r = r \cdot X \qquad (1.1.5)$$

é um estimador razão do total Y.

Igualmente, $\bar{X}$ sendo conhecida,

$$\bar{y}_r = r\bar{X} \qquad (1.1.6)$$

é um estimador razão da média aritmética $\bar{Y}$.

Cuidemos agora da caracterização da distribuição amostral do estimador r.

1.2 - ESPERANÇA MATEMÁTICA E VARIÂNCIA DO ESTIMADOR r

Consideremos o estimador r, como uma função de duas variáveis aleatórias $\bar{y}$ e $\bar{x}$, contínua e derivável num entorno do ponto $(\bar{Y}, \bar{X})$. Fazendo o desenvolvimento em série de Taylor dessa função de duas variáveis e considerando-se num desenvolvimento apenas os termos até 2ª ordem, tem-se:

$$f(\bar{y},\bar{x}) = f(\bar{Y},\bar{X}) + \frac{1}{1!}\left[(\bar{y}-\bar{Y})\frac{\partial f}{\partial \bar{y}} + (\bar{x}-\bar{X})\frac{\partial f}{\partial \bar{x}}\right]_{\substack{\bar{y}=\bar{Y}\\ \bar{x}=\bar{X}}} +$$

$$+ \frac{1}{2!}\left[(\bar{y}-\bar{Y})^2 \cdot \frac{\partial^2 f}{\partial \bar{y}^2} + 2(\bar{y}-\bar{Y})(\bar{x}-\bar{X}) \cdot \frac{\partial^2 f}{\partial \bar{y}\partial \bar{x}} + (\bar{x}-\bar{X})^2 \cdot \frac{\partial^2 f}{\partial \bar{x}^2}\right]_{\substack{\bar{y}=\bar{Y}\\ \bar{x}=\bar{X}}} + \text{Resto.}$$

ou seja, se $f(\bar{y},\bar{x}) = \dfrac{\bar{y}}{\bar{x}} = r$, virá,

$$r = R + \left[(\bar{y}-\bar{Y}) \cdot \frac{1}{\bar{X}} + (\bar{x}-\bar{X}) \cdot \left(-\frac{\bar{Y}}{\bar{X}^2}\right)\right] + \frac{1}{2}\left[2(\bar{y}-\bar{Y})(\bar{x}-\bar{X}) \cdot \left(-\frac{1}{\bar{X}^2}\right) + (\bar{x}-\bar{X})^2 \cdot \frac{2\bar{Y}}{\bar{X}^3}\right] +$$

$$+ \text{Resto.} \qquad (1.2.1)$$

Supondo-se agora amostragem sem reposição e tomando-se a esperança matemática de ambos os membros de *(1.2.1.)* virá:

$$E(r) \cong R + \frac{1}{\bar{X}^2}\left[\bar{X}E(\bar{y}-\bar{Y}) - \bar{Y}E(\bar{x}-\bar{X}) + R\,E(\bar{x}-\bar{X})^2 - E(\bar{y}-\bar{Y})(\bar{x}-\bar{X})\right]$$

$$= R + \frac{1-f}{n\bar{X}^2}\left[R\,S_x^2 - S_{xy}\right]$$

Donde se conclui que r é um estimador viesado de R, com viés, em primeira aproximação, dado por:

$$V_1 = \frac{1-f}{n\bar{X}^2}\left[R\,S_x^2 - S_{xy}\right] \qquad (1.2.2)$$

onde,

$$f = \frac{n}{N};\; S_x^2 = \frac{1}{N-1}\sum_{v=1}^{N}(X_v-\bar{X})^2 \quad e \quad S_{xy} = \frac{1}{N-1}\sum_{v=1}^{N}(X_v-\bar{X})(Y_v-\bar{Y}) \qquad (1.2.3)$$

Considerando no desenvolvimento em série de Taylor apenas os termos de ordem 0 e 1, o estimador r, para essa ordem de aproximação, seria não viesado, isto é, $E(r) = R$. Teríamos então, para essa ordem de aproximação, o seguinte valor para a variância aproximada de r:

$$\sigma^2(r) = E\left[r-R\right]^2 \cong E\left[(\bar{y}-\bar{Y})\cdot\frac{1}{\bar{X}} - (\bar{x}-\bar{X})\cdot\frac{\bar{Y}}{\bar{X}^2}\right]^2 =$$

$$= \frac{1}{\bar{X}^2}\,E\left[(\bar{y}-\bar{Y}) - R(\bar{x}-\bar{X})\right]^2$$

$$= \frac{1-f}{n\bar{X}^2}\left[S_y^2 - 2RS_{xy} + R^2S_x^2\right]$$

onde f, S_y^2, S_x^2 e S_{xy}, têm as definições dadas em *(1.2.3)*.

Esta variância também pode ser posta sob a forma:

$$\sigma^2(r) = \frac{1-f}{n\bar{X}^2}\cdot\frac{1}{N-1}\sum_{v=1}^{N}(Y_v - RX_v)^2 \qquad (1.2.4)$$

que mostra ser a dispersão de r função direta da dispersão dos valores da variável y ao redor da linha de regressão y = Rx da distribuição bidimensional das variáveis y e x, e função inversa do tamanho n da amostra casual.

Do mesmo modo, para o estimador $\tilde{y}_r$ do total populacional Y obtemos:

$$V_2 = \frac{N(1-f)}{n\bar{X}}\left[RS_x^2 - S_{xy}\right] \qquad (1.2.5)$$

como valor aproximado do viés, e:

$$\sigma^2(\tilde{y}_r) = \frac{N^2(1-f)}{n}\left[S_y^2 - 2RS_{xy} + R^2S_x^2\right] \qquad (1.2.6)$$

como valor aproximado para a variância.

Ainda, do mesmo modo,

$$V_3 = \frac{1-f}{n\bar{X}}\left[RS_x^2 - S_{xy}\right] \qquad (1.2.7)$$

será o valor aproximado do viés do estimador $\bar{y}_r$, e

$$\sigma^2(\bar{y}_r) = \frac{1-f}{n}\left[S_y^2 - 2RS_{xy} + R^2S_x^2\right] \qquad (1.2.8)$$

será o valor aproximado da respectiva variância.

Como vemos, os valores aproximados dos vieses e variâncias dos estimadores $\tilde{y}_r$ e $\bar{y}_r$ são facilmente deduzíveis, respectiva mente do viés e da variância do estimador r. Vamos a seguir fazer considerações apenas sobre o estimador r, mais diretamente ligado aos problemas que serão desenvolvidos no Capítulo II.

Estimadores praticamente não viesados das expressões *(1.2.2)* e *(1.2.3)*, a partir dos valores observados (x_i, y_i), são obtidos substituindo-se os parâmetros R, S_x^2, S_y^2 e S_{xy} pelos correspondentes amostrais dados por:

$$r = \frac{\bar{y}}{\bar{x}} \qquad s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Para se estimar a variância, no entanto, o melhor do ponto de vista operacional é partir do correspondente amostral da expressão *(1.2.4)*, ou seja:

$$s^2(r) = \frac{1-f}{\bar{X}^2 n(n-1)} \sum_{i=1}^{n} (y_i - rx_i)^2 =$$

$$= \frac{1-f}{\bar{X}^2 n(n-1)} \left[\sum_{i=1}^{n} y_i^2 - 2r \sum_{i=1}^{n} y_i x_i + r^2 \sum_{i=1}^{n} x_i^2 \right] \qquad (1.2.9)$$

onde, não sendo $\bar{X}$ conhecido, põe-se no seu lugar o valor $\bar{x}$, obtido da própria amostra.

1.3 - CONSIDERAÇÕES SOBRE A MAGNITUDE DO VIÉS DO ESTIMADOR *r*

Observamos inicialmente que, se a relação entre as variáveis y e x foi linear, passando pela origem, isto é, se $E(y|x) = R \cdot x$, então,

$$E(r) = E\frac{\sum y_i}{\sum x_i} = \frac{1}{\sum x_i} \sum_{i=1}^{n} E(y_i|x_i) =$$

$$= \frac{\sum Rx_i}{\sum x_i} = R \qquad (1.3.1)$$

ou seja, r será um estimador não viesado de R.

Nesse caso, ainda teremos que

$$\beta = \frac{\sum_{i=1}^{N} Y_i X_i}{\sum_{i=1}^{n} X_i^2},$$

coeficiente de regressão de mínimos quadrados de y sobre os x, também é igual a R, e consequentemente,

$$V_1 = \frac{1-f}{n\bar{X}^2}\left[RS_x^2 - S_{xy}\right] = 0 \qquad (1.3.2)$$

ou seja, o viés aproximado também é igual a zero.

Nos problemas de amostragem, no entanto, a hipótese de linearidade é difícil de ser posta à prova, preferindo-se investigar a magnitude do viés que pode estar afetando o estimador r.

Observamos inicialmente que o viés aproximado V_1 é de ordem $\frac{1}{n}$ enquanto que o erro padrão do mesmo estimador,

$$\sigma(r) = \sqrt{\frac{1-f}{n} \cdot \frac{1}{\bar{X}^2} (S_y^2 - 2RS_{xy} + R^2 S_x^2)}$$

é de ordem $1/\sqrt{n}$. Isto significa que para n (tamanho da amostra) crescente, o viés tende muito mais rapidamente a zero do que o respectivo erro padrão, ou dizendo de outro modo, para n grande, o viés que afeta r deve ser pequeno comparado com o seu erro padrão.

Mas ainda, fazendo-se o quociente do módulo do viés V_1 pelo erro padrão de r, obtemos:

$$\frac{|V_1|}{\sigma(r)} = \frac{\left|\frac{1-f}{n\bar{X}^2}(RS_x - S_{xy})\right|}{\sqrt{\frac{1-f}{n\bar{X}^2}(S_y^2 - 2RS_{xy} + R^2S_x^2)}} \leq \sqrt{\frac{1-f}{n}} \cdot \frac{S_x}{\bar{X}} = c.v.(\bar{x}) \tag{1.3.3}$$

ou seja, o quociente do módulo do viés V_1 pelo erro padrão $\sigma(r)$ é menor do que o coeficiente de variação da média aritmética $\bar{x}$ da variável auxiliar x, quando muito igual. Este resultado, aparentemente aproximado, porque V_1 e o $\sigma(r)$ são valores aproximados, é no entanto válido para o verdadeiro viés. De fato, denotando-se por V o verdadeiro viés, temos:

$$|V| = |E(r)-R| = \left|-\frac{1}{\bar{X}} E\left[\left[\frac{\bar{y}}{\bar{x}} - E\left(\frac{\bar{y}}{\bar{x}}\right)\right]\left[\bar{x}-E(\bar{x})\right]\right]\right| = \left|-\frac{1}{\bar{X}}\sigma_{r\bar{x}}\right| \leq \frac{1}{\bar{X}}\sigma(r)\sigma(\bar{x})$$

Donde,

$$\frac{|V|}{\sigma(r)} < \frac{\sigma(\bar{x})}{\bar{X}} = c.v.(\bar{x}) \tag{1.3.4}$$

Cochran, na hipótese de um estimador se distribuir normalmente, mostrou que a existência de um viés inferior em módulo, a 0,1 do erro padrão praticamente não altera o coeficiente de confiança a ele associado. Tendo então em conta este resultado e o obtido na expressão *(1.3.4)*, podemos fixar a seguinte regra prática relativamente ao tamanho n da amostra necessária para se obter um es timador r praticamente não viesado: determinar n tal que $c.v.(\bar{x}) \leq$ $\leq 0{,}1$. Isto será possível se em relação à variável auxiliar, também dispusermos de informações suplementares quanto à variabilidade.

1.4 - O ESTIMADOR $\bar{r}_H$ DO TIPO RAZÃO E NÃO VIESADO DE HARTLEY E ROSS

Hansem e Hurwitz (1943) mostraram que se a amostragem foi com reposição e proporcional aos valores X_i da variável auxiliar x, a expressão

$$\bar{r} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i} \tag{1.4.1}$$

é um estimador não não viesado da razão $R = \dfrac{\bar{Y}}{\bar{X}}$.

Hartley e Ross (1954), considerando este mesmo estimador $\bar{r}$ na amostragem simples com reposição, mostraram que o verdadeiro viés,

$$V(\bar{r}) = E(\bar{r}) - R = \bar{R} - R, \text{ onde } \bar{R} = \frac{1}{N} \sum_{v=1}^{N} \frac{Y_v}{X_v} \tag{1.4.2}$$

pode ser expresso como função da covariância σ_{rx} entre as variáveis $r = \dfrac{y}{x}$ e x, ou seja:

$$V(\bar{r}) = -\frac{1}{\bar{X}} \times \frac{1}{N} \sum_{v=1}^{N} \left(\frac{Y_v}{X_v} - R\right)\left(X_v - \bar{X}\right) = -\frac{1}{\bar{X}}\,\sigma_{rx} \tag{1.4.3}$$

Ora, um estimador não viesado da covariância σ_{rx} é obtido das observações amostrais calculando-se

$$s_{rx} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \bar{r}\right)(x_i - \bar{x}) = \frac{n}{n-1} \left[\bar{y} - \bar{r}\bar{x}\right] \qquad (1.4.4)$$

Então, se definirmos

$$\bar{r}_H = \bar{r} + \frac{n}{\bar{X}(n-1)} \left[\bar{y} - \bar{r}\bar{x}\right] \qquad (1.4.5)$$

segue-se que este estimador $\bar{r}_H$ é um estimador não viesado da razão R.

1.5 - COMPARAÇÃO DAS VARIÂNCIAS DOS ESTIMADORES r E $\bar{r}_H$

A variância exata do estimador $\bar{r}_H$ na amostragem casual com reposição foi obtida por Hartley e Goodman (1958); é dada pela expressão:

$$\sigma^2(\bar{r}_H) = \frac{1}{n\bar{X}^2}\left[\sigma^2 - 2R\,\sigma_{xy} + \bar{R}^2\sigma_x^2 + \frac{1}{n-1}(\sigma_x^2 \cdot \sigma_r^2 + \sigma_{xr}^2)\right] \qquad (1.5.1)$$

onde todas as variâncias e covariâncias que aparecem aqui são definidas como denominador N.

Sendo a amostragem sem reposição e para a ordem de aproximação $\frac{1}{n}$ considerada na determinação da variância de $r = \bar{y}/\bar{x}$, podemos escrever para o valor aproximado da variância de $\bar{r}_H$:

$$\sigma^2(\bar{r}_H) \doteq \frac{1-f}{n\bar{X}^2}\left[S_y^2 - 2\bar{R}S_{xy} + \bar{R}^2 S_x^2\right] \qquad (1.5.2)$$

Comparemos, agora, as variâncias dos estimadores r e $\bar{r}_H$. Inicialmente, observamos que, se a relação entre y e x for linear passando pela origem, isto é, se o modelo for do tipo:

$$E(y|x) = \beta x$$

r e $\bar{r}_H$ terão a mesma precisão.

2º) se a relação entre y e x for linear não passando necessariamente pela origem, isto é, se o modelo for do tipo:

$$E(y|x) = \alpha + \beta x$$

verificaremos facilmente que

$$R = \frac{\alpha}{E(x)} + \beta \quad \text{e} \quad \bar{R} = \alpha \cdot E(\frac{1}{x}) + \beta$$

onde
$$\beta = \frac{\sum_{v=1}^{N} (X_v - \bar{X})(Y_v - \bar{Y})}{\sum_{v=1}^{N} (X_v - \bar{X})^2}$$

A diferença entre as variâncias dos estimadores r e $\bar{r}_H$ pode ser posta sob a forma:

$$\sigma^2(\bar{r}_H) - \sigma^2(r) = \frac{\alpha^2 \sigma_x^2}{n\bar{X}} \left\{ E^2(\frac{1}{x}) - (\frac{1}{E(x)})^2 \right\} \qquad (1.5.3)$$

e, posto que

$$E(\frac{1}{x}) \geq \frac{1}{E(x)}$$

segue-se que, sob o modelo proposto, r é mais preciso do que $\bar{r}_H$.

3º) No caso geral, verificamos que a diferença entre as variâncias aproximadas de r e $\bar{r}_H$, dadas respectivamente pelas expressões *(1.2.3)* e *(1.5.2)*, pode ser posta sob a forma:

$$\sigma^2(\bar{r}_H) - \sigma^2(r) = \frac{1-f}{n} \frac{S_x^2}{\bar{X}^2} \left\{ (\bar{R} - \beta)^2 - (R - \beta)^2 \right\} \qquad (1.5.4)$$

Desse resultado *(1.5.4)* podemos inferir que, se $\bar{R}$ estiver mais próximo de β do que R, a diferença será negativa, indicando isso que o estimador $\bar{r}_H$ é mais preciso do que r.

Verificações empíricas nos casos práticos mais frequentes têm, no entanto, mostrado que R está mais próximo de β do que R. Na prática, portanto, é de se supor que r seja mais preciso do que $\bar{r}_H$. De qualquer modo, sendo n grande, a diferença *(1.5.4)* sempre pode ser estimada a partir da própria amostra.

CAPÍTULO II

2.1 - ESPECIFICAÇÃO DA POPULAÇÃO - PARÂMETROS DE POSIÇÃO E DE VARIABILIDADE

Conglomerado é uma unidade de amostragem, natural ou artificial, composta por outras unidades menores chamadas elementos. Re-lmente, poderíamos considerar os conglomerados constituídos por u… hierarquia de unidades de amostragem (unidades de 1ª ordem, 2ª ordem, 3ª ordem, etc.), as de maior ordem aninhadas dentro das de menor ordem. Neste trabalho iremos nos ocupar, apenas, de conglomerados (unidades de 1ª ordem) compostos por elementos (unidades de 2ª ordem).

Denotaremos por M o número de conglomerados, e por N_v ($v=1,2,\ldots,M$) o número de elementos de cada conglomerado. Denotaremos por

$$N = \sum_{v=1}^{M} N_v,$$

o número total de elementos da população originária. N é, então, o tamanho da população em termos de elementos de 2ª ordem e M é o tamanho a população em termos de conglomerados (ou elementos de 1ª ordem).

No caso de uma enumeração completa dos M conglomerados e dos N_v $(v=1,2,\ldots,M)$ elementos que compõem cada um deles, os valores possíveis da variável y (objeto de estudo) associada a cada elemento são denotados por

$$Y_{vu}, \quad (v=1,2,\ldots,M, \; u=1,2,\ldots,N_v) \qquad (2.1.1)$$

Denotemos por $\dot{y}$ a variável soma dos valores de y em relação aos elementos que compõem cada conglomerado, e os valores possíveis de $\dot{y}$ por $Y_{v\cdot}, (v=1,2,\ldots,M)$, sendo:

$$Y_{v} = \sum_{v=1}^{N_v} Y_{vu} \qquad (2.1.2)$$

Do mesmo modo, denotemos por $\mathring{n}$ a variável tamanho dos conglomerados, variável essa que pode assumir os valores N_v, $(v=1,2,\ldots,M)$.

O quadro I, a seguir, sintetisa a simbologia utilizada na apresentação dos valores possíveis da variável y, no caso de uma enumeração completa dos M conglomerados e de seus elementos constitutivos:

QUADRO I

CONGLOMERADOS	y	$\dot{y}$	$\mathring{n}$	$\dot{y}/\mathring{n}$
1	$Y_{11}, Y_{12},\ldots,Y_{1u},\ldots,Y_{1N_1}$	$Y_{1\cdot}$	N_1	$\bar{y}_1$
2	$Y_{21}, Y_{22},\ldots,Y_{2u},\ldots,Y_{2N_2}$	$Y_{2\cdot}$	N_2	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
v	$Y_{v1}, Y_{v2},\ldots,Y_{vu},\ldots,Y_{vN_v}$	$Y_{v\cdot}$	N_v	$\bar{y}_v$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M	$Y_{M1}, Y_{M2},\ldots,Y_{Mu},\ldots,Y_{MN_M}$	$Y_{M\cdot}$	N_v	$\bar{y}_v$

onde, na última coluna, estão arroladas as médias aritméticas por elementos de cada conglomerado, ou seja, os valores possíveis da variável razão $\dot{y}/\dot{n}$.

Em relação a esta população de valores possíveis da variável y, correspondente à enumeração completa, vamos definir alguns parâmetros de posição e de variabilidade que merecem algumas considerações porque reaparecerão logo mais adiante quando tratamos dos estimadores objetos da presente dissertação.

1º - Média aritmética por elemento, definida pela expressão:

$$\bar{\bar{Y}} = \frac{1}{N} \sum_{v=1}^{M} \sum_{u=1}^{N_v} Y_{vu} = \frac{1}{N} \sum_{v=1}^{M} N_v \bar{Y}_v = \frac{\sum_{v=1}^{M} Y_{v\cdot}}{\sum_{v=1}^{M} N_v} = \frac{\bar{Y}}{\bar{N}} = \bar{\bar{Y}}_R$$

onde,

$$\bar{Y} = \frac{1}{M} \sum_{v=1}^{M} Y_{v\cdot} \quad \text{e} \quad \bar{N} = \frac{1}{M} \sum_{v=1}^{M} N_v \tag{2.1.3}$$

Observamos que:

a) a média aritmética por elemento $\bar{\bar{Y}}$ é uma média aritmética ponderada das médias aritméticas por elemento $\bar{Y}_v$ dos conglomerados, com ponderação dada pelos tamanhos N_v destes.

b) a média $\bar{\bar{Y}}$ também o é quociente das somas das variáveis $\dot{y}$ e $\dot{n}$, sendo, nesse sentido, o coeficiente angular da reta que passa pela origem e pelo baricentro $(\bar{Y}, \bar{N})$, on de, como sabemos, $\bar{\bar{Y}}$ é a média aritmética por conglomerado da variável y e $\bar{N}$ é o tamanho médio dos conglomerados. Ela é, então, equivalente à razão R *(1.1.3)* do capítulo anterior, quando se fizer a substituição dos valores possíveis X_v da variável auxiliar x, pelos valores possíveis N_v da variável auxiliar $\dot{n}$ = tamanho dos conglomerados. Para lembrarmos dessa dupla significação de $\bar{\bar{Y}}$, vamos denotá-la também por $\bar{\bar{Y}}_R$.

2º - Definimos ainda como medida de posição por elemento a **média** aritmética da variável razão $\dot{y}/\dot{n}$, dada pela expressão:

$$\bar{\bar{Y}}_{\bar{R}} = \frac{1}{M}\sum_{v=1}^{M}\frac{Y_{v\cdot}}{N_v} = \frac{1}{M}\sum_{v=1}^{M}\bar{Y}_v \qquad (2.1.4)$$

Essa média aritmética corresponde à $\bar{R}$ *(1.4.2)* do capítulo anterior e, naturalmente, $\bar{\bar{Y}}_{\bar{R}} \neq \bar{\bar{Y}}_R = \bar{\bar{Y}}$, a não ser que os M conglomerados tenham, todos eles, o mesmo tamanho, ou seja, $N_v = \bar{N}$ (v=1,2,...,M), quando, então, $\bar{\bar{Y}}_{\bar{R}} = \bar{\bar{Y}}_R = \bar{\bar{Y}}$.

3º - Definimos a variância por elemento da variável y pela expressão

$$S_y^2 = \frac{1}{N-1}\sum_{v=1}^{M}\sum_{u=1}^{N_v}(Y_{vu} - \bar{\bar{Y}}_R)^2 \qquad (2.1.5)$$

Observamos que esta variância pode ser decomposta como segue:

$$S_y^2 = \frac{1}{N-1}\left[\sum_{v=1}^{M}(N_v-1)S_v^2 + \sum_{v=1}^{M}N_v(\bar{Y}_v-\bar{\bar{Y}}_R)^2\right] \qquad (2.1.6)$$

onde,

$$S_v^2 = \frac{1}{N_v-1} \sum_{u=1}^{N_v} (Y_{vu}-\bar{Y}_v)^2 \qquad (2.1.7)$$

são as chamadas variâncias dentro dos conglomerados.

4º - Em relação às variáveis $\dot{y}$ e $\dot{n}$, definimos as variâncias e covariâncias pelas expressões:

$$S^2_{(\dot{y})} = \frac{1}{M-1} \sum_{v=1}^{M} (Y_{v.}-\bar{Y})^2 \quad \text{e} \quad S^2_{(\dot{n})} = \frac{1}{M-1} \sum_{v=1}^{M} (N_v-\bar{N})^2 \quad \text{e}$$

$$S_{(\dot{y},\dot{n})} = \frac{1}{M-1} \sum_{v=1}^{M} (Y_{v.}-\bar{Y})(N_v-\bar{N}) \qquad (2.1.8)$$

5º - Finalmente, definimos a variância da variável razão $\dot{y}/\dot{n}$, pela expressão:

$$S^2_{\dot{y}/\dot{n}} = \frac{1}{M-1} \sum_{v=1}^{M} (\bar{Y}_v-\bar{\bar{Y}}_{\bar{R}})^2 \qquad (2.1.9)$$

2.2 - *AMOSTRAGEM DE CONGLOMERADOS COM ENUMERAÇÃO COMPLETA DOS SEUS ELEMENTOS*

A amostragem de $m < M$ conglomerados ao acaso difundiu-se na pesquisa aplicada, por razões várias, dentre as quais, a mais frequente é a inexistência de um sistema de referência ou listagem dos N elementos que compõem a população originária. Por outro lado, a organização ou construção de uma lista dos M conglomerados que irá servir como sistema de referência para a seleção da amostra de tamanho m é relativamente fácil, sem grandes dispêndios de tempo e /ou dinheiro.

O quadro II, a seguir, é uma reprodução do quadro I, em termos de valores observados da variável y, numa amostragem casual de tamanho m, com enumeração completa dos elementos constitutivos dos conglomerados selecionados:

<u>QUADRO II</u>

CONGLOMERADOS	y	$\dot{y}$	$\dot{n}$	$\dot{y}/\dot{n}$
1	$y_{11}, y_{12},\ldots,y_{1j},\ldots,y_{1n_2}$	$y_{1.}$	n_1	$\bar{y}_1$
2	$y_{21}, y_{22},\ldots,y_{2j},\ldots,y_{2n_2}$	$y_{2.}$	n_2	$\bar{y}_2$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$	$\vdots$	$\vdots$
i	$y_{i1}, y_{i2},\ldots,y_{ij},\ldots,y_{in_i}$	$y_{i.}$	n_i	$\bar{y}_i$
$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$	$\vdots$	$\vdots$	$\vdots$
m	$y_{m1}, y_{m2},\ldots,y_{mj},\ldots,y_{mn_m}$	$y_{m.}$	n_m	$\bar{y}_m$

Em relação a este quadro, observamos o mesmo princípio já adotado na exposição do Capítulo I, isto é, os valores possíveis da variável y, na amostra de tamanho m, foram denotados pelas minúsculas correspondentes, acompanhadas de dois índices. O primeiro, i (i=1,2,...,m), refere-se ao conglomerado selecionado na i-ésima extração, e o segundo j, refere-se ao elemento genérico dentro do conglomerado i. Os tamanhos n_i dos conglomerados, bem como os valores:

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij} \qquad e \qquad \bar{y}_i = \frac{y_{i.}}{n_i}$$

são variáveis aleatórias que poderão assumir, respectivamente, os valores:

$$n_i \; ; \; N_1 \; , \; N_2 \; , \; \ldots , \; N_v \; , \; \ldots , \; N_M$$
$$y_{i.} ; \; Y_{1.} , \; Y_{2.} , \; \ldots , \; Y_{v.} , \; \ldots , \; Y_{M.}$$
$$\bar{y}_i \; ; \; \bar{Y}_1 \; , \; \bar{Y}_2 \; , \; \ldots , \; \bar{Y}_v \; , \; \ldots , \; \bar{Y}_M$$

sendo, a função de probabilidade associada dada por:

$$P(n_i = N_v) = P(y_{i.} = Y_{v.}) = P(\bar{y}_i = \bar{Y}_v) = \frac{1}{M} \quad , \quad \text{para } v=1,2,\ldots,M.$$

Sendo o objetivo principal da amostragem, a estimação de um dos parâmetros focalizados na secção anterior — $\bar{\bar{Y}}$ por exemplo — é intuitivo que a seleção de uma amostra casual de tamanho $m < M$ conglomerados, com enumeração completa de seus elementos, ao invés de uma amostra casual do mesmo tamanho mas em termos de elementos simples coloca, desde logo, o problema de se saber a "quantidade de informação", ou precisão, que pode estar sendo perdida nesta substituição de elementos simples por conglomerados de elementos. O problema não é simples, dependendo:

a) do tipo da amostra casual;

b) da definição do estimador e das suas propriedades; e

c) do comportamento das componentes da variância da variável y (objeto de estudo, segundo a decomposição especifica pelo 2º membro da igualdade *(2.1.6)* da secção anterior).

2.3 - ALGUNS ESTIMADORES NA PRÁTICA

Numa amostragem casual sem reposição de m < M conglomerados com enumeração completa de seus elementos, um dos estimadores, às vezes utilizado, quando são conhecidos os tamanhos médios N dos conglomerados, é dado pela expressão:

$$\bar{\bar{y}} = \frac{\bar{y}}{\bar{N}}, \quad \text{onde} \quad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i. \qquad (2.3.1)$$

É fácil verificarmos que $\bar{\bar{y}}$ é um estimador não viesado da média aritmética por elemento $\bar{\bar{Y}}$ da população; basta notar $E(\bar{\bar{y}}) = \bar{\bar{Y}}$.

A variância de $\bar{\bar{y}}$ é dada pela expressão:

$$\sigma^2_{(\bar{\bar{y}})} = \frac{1}{\bar{N}^2}(1-f_1)\frac{S^2_y}{m} \qquad (2.3.2)$$

onde $f_1 = \frac{m}{M}$ é a fração de amostragem de conglomerados, e S^2_y é a variância dos totais dos conglomerados em relação à média $\bar{Y}$ por conglomerado.

Notando que

$$S^2_{\dot{y}} = \frac{1}{M-1} \sum_{v=1}^{M} (Y_{v.} - \bar{Y})^2$$

$$= \frac{1}{M-1} \sum_{v=1}^{M} (N_v \bar{Y}_v - \bar{N}\bar{\bar{Y}})^2$$

vemos que sua magnitude depende da relação existente entre os totais $Y_{v.}$ e os tamanhos N_v dos conglomerados, podendo ser muito grande, ainda que as médias por elemento $\bar{Y}_v$ dos conglomerados variem pouco ao redor da média por elemento $\bar{\bar{Y}}$. Portanto, se a variância dos N_v for grande, também será grande $S^2_{\dot{y}}$ e, consequentemente, a variância do estimador $\bar{\bar{y}}$. No caso especial dos tamanhos N_v dos conglomerados serem constantes, isto é, $N_v = \bar{N}$, então,

$$S^2_{\dot{y}} = \bar{N}^2 \frac{1}{M-1} \sum_{v=1}^{M} (\bar{Y}_v - \bar{\bar{Y}})^2 = \bar{N}^2 S^2_{\dot{y}/\dot{n}}$$

e a variancia do estimador $\bar{\bar{y}}$ será dada por:

$$\sigma^2_{(\bar{\bar{y}})} = (1-f_1) \cdot \frac{1}{m(M-1)} \sum_{v=1}^{M} (\bar{Y}_v - \bar{\bar{Y}})^2$$

$$\doteq (1-f_1) \cdot \frac{S^2_{\dot{y}/\dot{n}}}{m} \qquad (2.3.3)$$

No intuito de se reduzir a variância do estimador $\bar{\bar{y}}$, várias alternativas podem ser consideradas. Uma delas, muito utilizada na prática, consiste em se estratificar os M conglomerados em L classes contíguas e homogêneas em relação aos tamanhos N_v dos conglomerados. É claro, neste caso, que a amostragem subseqüente será uma amostragem estratificada de conglomerados. Outra alternativa consiste na construção de conglomerados artificiais, agregando conglomerados menores, naturais ou artificiais, ou mesmo elementos, de modo a se obter ao final uma população de M conglomerados, que não defíram muito em relação ao número de elementos que os compõem. O que orienta e determina a construção dessa população de M conglomerados artificiais é, no entanto, o comportamento das componentes da variância da variável y (objeto de estudo) especificadas pelo 2º membro da igualdade *(2.1.6)*. Devemos procurar fazer com que as chamadas variâncias dentro, S_v^2, sejam as maiores possíveis, já que não contribuem, como acabamos de ver, para a variância do estimador $\bar{\bar{y}}$ na amostragem de conglomerados com enumerações completa de seus elementos constitutivos.

Outra alternativa, consiste na utilização dos chamados estimadores razão ou do tipo razão, dos quais vamos nos ocupara a seguir.

O estimador razão da média populacional por elemento $\bar{\bar{Y}}$ é definido pela expressão:

$$\bar{\bar{y}}_r = \frac{\sum_{i=1}^{m} y_{i.}}{\sum_{i=1}^{m} n_i} = \frac{\bar{y}}{\bar{n}} \qquad (2.3.4)$$

Como vemos, $\bar{\bar{y}}_r$ é uma função quociente de duas variáveis aleatórias. Então, dentro da mesma linha de exposições seguida na secção (1.2), podemos escrever que o viés aproximado de $\bar{\bar{y}}_r$ é dado pela expressão:

$$V_1 = \frac{1-f_1}{\bar{N}^2 m}\left[\bar{\bar{Y}}_R S_{\dot{n}}^{\cdot} - S_{\dot{y}\dot{n}}\right] \tag{2.3.5}$$

Para a variância aproximada de $\bar{\bar{y}}_r$, podemos escrever:

$$\sigma^2(\bar{\bar{y}}_r) \cong \frac{1-f_1}{\bar{N}^2 m}\left[S_{\dot{y}}^2 - 2\bar{\bar{Y}}_R S_{\dot{y}\dot{n}} + \bar{\bar{Y}}_R^2 S_{\dot{n}}^2\right] \tag{2.3.6}$$

onde, naturalmente, as variâncias e covariâncias que aparecem nas expressões *(2.3.5)* e *(2.3.6)* foram definidas pelas expressões *(2.1.8)* e $f_1 = \frac{m}{M}$ é fração de amostragem de conglomerados.

Esta variância pode ainda ser escrita sob a forma:

$$\sigma^2(\bar{\bar{y}}_r) = \frac{1-f_1}{\bar{N}^2 m}\;\frac{1}{M-1}\sum_{v=1}^{M}(Y_{v.} - \bar{\bar{Y}}_R N_v)^2 \tag{2.3.7}$$

mostrando-nos que a dispersão de $\bar{\bar{y}}_r$, é função direta da **dispersão** dos valores da variável $\dot{y}$ ao redor da reta de regressão $\dot{y} = \bar{\bar{Y}}_R \dot{n}$ da distribuição bidimensional das variáveis $\dot{y}$ e $\dot{n}$, e função inversa do tamanho m da amostra de conglomerados.

Um estimador, praticamente não viesado, da variância de $\bar{\bar{y}}_r$ é dado pela expressão:

$$s^2(\bar{\bar{y}}_r) = \frac{1-f_1}{\bar{N}^2 m(m-1)}\sum_{i=1}^{m}(y_{i.} - \bar{\bar{y}}_r n_i)^2$$

$$= \frac{1-f_1}{\bar{N}^2 m(m-1)}\left[\sum_{i=1}^{m} y_{i.}^2 - 2\bar{\bar{y}}_r \sum_{i=1}^{m} y_{i.} n_i + \bar{\bar{y}}_r^2 \sum_{i=1}^{m} n_i^2\right] \tag{2.3.9}$$

facilmente calculável a partir das observações amostrais dispostas em um quadro semelhante ao da secção (2.2).

Do mesmo modo, as considerações que fizemos na secção (1.3), relativamente à amplitude do viés do estimador r em relação ao seu erro padrão $\sigma(r)$, podem ser feitas aqui em relação ao viés V_1 do estimador $\bar{\bar{y}}_r$ em relação ao erro padrão $\sigma(\bar{\bar{y}}_r)$. Em particular temos:

$$\frac{|V_1|}{\sigma(\bar{\bar{y}}_r)} \leq \frac{1-f_1}{m} \cdot \frac{S_{\bar{n}}}{\bar{N}} = s.v.(\bar{n}). \qquad (2.3.9)$$

colocando desse modo o viés do estimador $\bar{\bar{y}}_r$ sob controle estatístico.

2.4 - O ESTIMADOR $\bar{\bar{Y}}_H$ DO TIPO RAZÃO E NÃO VIESADO

É sabido que, uma amostragem casual com reposição e proporcionalmente aos tamanhos N_v dos conglomerados, o estimador definido pela relação:

$$\bar{\bar{y}}_{\bar{r}} = \frac{1}{m} \sum_{i=1}^{m} \frac{Y_{i.}}{n_i}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \bar{Y}_i \qquad (2.4.1)$$

é um estimador não viesado da média populacional por elemento $\bar{\bar{Y}}$. Este mesmo estimador, no entanto, se a amostragem for simples, isto é, não proporcional aos tamanhos dos conglomerados terá um viés da do pela relação:

$$V(\bar{\bar{y}}_{\bar{r}}) = E(\bar{\bar{y}}_{\bar{r}}) - \bar{\bar{Y}}$$
$$= \bar{\bar{Y}}_{\bar{R}} - \bar{\bar{Y}} \qquad (2.4.2)$$

No caso de $\bar{N}$ ser conhecido, este verdadeiro viés pode ser posto sob a forma:

$$V(\bar{\bar{y}}_{\bar{r}}) = -\frac{1}{\bar{N}} \cdot \frac{1}{M} \sum_{v=1}^{M} \left(\frac{Y_{v.}}{N_v} - \bar{\bar{Y}}_{\bar{R}} \right) \cdot (N_v - \bar{N})$$

$$= -\frac{1}{\bar{N}} \sigma_{(\dot{y}/\dot{n},\dot{n})} \qquad (2.4.3)$$

Ora, um estimador não viesado da covariância

$$\sigma_{(\dot{y}/\dot{n},\dot{n})} = \frac{1}{M} \sum_{v=1}^{M} \left(\frac{Y_{v.}}{N_v} - \bar{\bar{Y}}_{\bar{R}} \right) (N_v - \bar{N})$$

$$= \bar{Y} - \bar{\bar{Y}}_{\bar{R}} \bar{N} \qquad (2.4.4)$$

sempre pode ser obtido a partir da amostra selecionada

$$s_{(\dot{y}/\dot{n},\dot{n})} = \frac{1}{m-1} \sum_{i=1}^{m} \left(\frac{y_{i.}}{n_i} - \bar{\bar{y}}_{\bar{r}} \right) (n_i - \bar{n})$$

$$= \frac{m}{m-1} \left[\bar{y} - \bar{\bar{y}}_{\bar{r}} \bar{n} \right] \qquad (2.4.5)$$

Definindo-se então, o estimador $\bar{\bar{y}}_H$ pela relação

$$\bar{\bar{y}}_H = \bar{\bar{y}}_{\bar{r}} + \frac{m}{\bar{N}(m-1)} \left[\bar{y} - \bar{\bar{y}}_{\bar{r}}\bar{n}\right] \quad (2.4.6)$$

segue-se que $\bar{\bar{y}}_H$ é um estimador não viesado da média populacional por elemento $\bar{\bar{Y}}$.

Obtivemos a variância exata do estimador $\bar{\bar{y}}_H$ seguindo o desenvolvimento dado por Hartley e Goodman (1958). Podemos escrever essa variância sob a forma:

$$\sigma^2(\bar{\bar{y}}_H) = \frac{1}{\bar{N}^2 m} \left[\sigma^2_{\dot{y}} - 2\bar{\bar{Y}}_{\bar{R}}\sigma_{\dot{y}\dot{n}} + \bar{\bar{Y}}^2_{\bar{R}}\sigma^2_{\dot{n}} + \frac{1}{m-1}(\sigma^2_{\dot{n}} \circ \sigma^2_{\dot{y}/\dot{n}}, \quad + \sigma_{\dot{y}/\dot{n},\dot{n}}\right] \quad (2.4.7)$$

onde, aqui, todas as variâncias e covariâncias que aparecem no 2º membro são definidas com denominador M.

Na amostragem casual sem reposição e para a ordem de aproximação $\frac{1}{m}$ conservada na determinação da variância do estimador razão $\bar{\bar{y}}_r$, podemos escrever a variância aproximada de $\bar{\bar{y}}_H$ sob a forma:

$$\sigma^2(\bar{\bar{y}}_H) \cong \frac{1-f_1}{\bar{N}^2 m}\left[S^2_{\dot{y}} - 2\bar{\bar{Y}}_{\bar{R}}S_{\dot{y}\dot{n}} + \bar{\bar{Y}}_{\bar{R}}S_{\dot{n}}\right] \quad (2.4.8)$$

As considerações expendidas na secção (1.5), quando da comparação das variâncias dos estimadores r e $\bar{r}_H$, repetem-se aqui em relação aos estimadores $\bar{\bar{y}}_r$ e $\bar{\bar{y}}_H$. Em particular, sendo

$$\beta = \frac{\sum_{v=1}^{M} (Y_{v.} - \bar{\bar{Y}})(N_v - \bar{N})}{\sum_{v=1}^{M} (N_v - \bar{N})^2}$$

a diferença entre as variâncias aproximadas de $\bar{\bar{y}}_r$ e $\bar{\bar{y}}_H$ pode ser posta sob a forma:

$$\sigma^2(\bar{\bar{y}}_H) - \sigma^2(\bar{\bar{y}}_r) = \frac{1-f_1}{m} \frac{S_n^2}{\bar{N}^2}\left[(\bar{\bar{Y}}_{\bar{R}}-\beta)^2 - (\bar{\bar{Y}}_R-\beta)^2\right] \qquad (2.4.9)$$

que nos mostra que, se $\bar{\bar{Y}}_{\bar{R}}$ estiver mais próximo de β do que $\bar{\bar{Y}}_R$, a diferença será negativa, sendo neste caso, $\bar{\bar{y}}_H$ mais preciso do que $\bar{\bar{y}}_r$. Devemos assinalar, no entanto, que ainda não há determinações empíricas que elucidem a magnitude e o sinal dessa diferença. Pelo que sabemos relativamente aos estimadores r e $\bar{r}_H$ que têm como variável auxiliar X de natureza diversa dos tamanhos N_v dos conglomerados, tem-se verificado empiricamente que R está mais próximo de β do que $\bar{R}$.

De qualquer modo, acreditamos poder concluir que o estimador $\bar{\bar{y}}_H$ leva sobre o estimador $\bar{\bar{y}}_r$ a vantagem de ser não viesado, ainda que possa vir a apresentar, em algumas aplicações, variância maior. Em comparação com o estimador $\bar{\bar{y}}_{\bar{r}}$ (utilizado algumas vezes na prática), $\bar{\bar{y}}_H$ leva a vantagem de ser não viesado, sendo portanto a diferença entre eles, uma medida amostral da magnitude do viês implícito em $\bar{\bar{y}}_{\bar{r}}$. É claro, no entanto, que a utilização do estimador $\bar{\bar{y}}_H$ exige o conhecimento prévio de $\bar{N}$, conhecimento esse não exigido pelos estimadores $\bar{\bar{y}}_r$ e $\bar{\bar{y}}_{\bar{r}}$. Para finalisar observamos que, se na definição de $\bar{\bar{y}}_H$ substituirmos $\bar{N}$ por $\bar{n}$, obtido da própria amostra, es te estimador assumirá a forma:

$$\bar{\bar{y}}_H = \bar{\bar{y}}_{\bar{r}} + \frac{m}{(m-1)\bar{r}}(\bar{y} - \bar{\bar{y}}_{\bar{r}}\bar{n})$$

$$= \bar{\bar{y}}_{\bar{r}} + \frac{m}{m-1}(\bar{\bar{y}}_r - \bar{\bar{y}}_{\bar{r}}).$$

onde o verdadeiro viés de $\bar{\bar{y}}_{\bar{r}}$,(o 2º termo do 2º membro),é praticamente igual à diferença entre os estimadores razão $\bar{\bar{y}}_r$ e o estimador dado pela média das médias sem ponderação.

C O N C L U S Õ E S

1) - Se "Informações Suplementares", nos permitirem supor que $\bar{\bar{Y}}_R$ está mais próximo ao coeficiente de regressão β do que $\bar{\bar{Y}}_R$, então o estimador $\bar{\bar{y}}_H$, será preferido ao estimador $\bar{\bar{y}}_r$.

2) - Se "Informações Suplementares nos permitirem observar que o coeficiente de regressão β é maior que $\frac{1}{2}\bar{\bar{Y}}_{\bar{R}}$, então devemos preferir o estimador $\bar{\bar{y}}_H$ ao estimador $\bar{\bar{y}}$, e se o coeficiente de regressão é menor ou igual a $\frac{1}{2}\bar{\bar{Y}}_{\bar{R}}$ então o estimador $\bar{\bar{y}}$ será preferido ao estimador $\bar{\bar{y}}_H$, por ser mais preciso e mais simples de ser computado.

3) - Se "Informações Suplementares" relativas à variável $\dot{n}$ nos permitirem obter uma amostra com reposição e proporcional aos tamanhos N_v dos conglomerados, então o estimador $\bar{\bar{y}}_{\bar{r}}$ deve ser preferido aos demais estimadores estudados.

Das conclusões 1, 2 e 3 observamos a importância das "informações suplementares" relativas ao particular fenômeno, objeto da investigação. Quanto mais ricas e fidedignas forem, mais preciso poderá ser o plano amostral e o método de estimação a ser usado.

4) - O estimador $\bar{\bar{y}}_H$ leva sobre o estimador $\bar{\bar{y}}_r$ a vantagem de ser não-viesado, ainda que possa vir a apresentar, em algumas

aplicações, variância maior. Em relação ao estimador $\bar{\bar{y}}_{\bar{r}}$, é claro que $\bar{\bar{y}}_H$ é superior, se consideramos o viés de $\bar{\bar{y}}_{\bar{r}}$.

No entanto, a utilização do estimador $\bar{\bar{y}}_H$ exige o conhecimento prévio de $\bar{N}$ conhecimento este não exigido pelos estimadores $\bar{\bar{y}}_r$ e $\bar{\bar{y}}_{\bar{r}}$. Se, na definição de $\bar{\bar{y}}_H$, substituirmos $\bar{N}$ por $\bar{n}$, obtida da própria amostra, então, à medida que o tamanho da amostra cresce, o estimador $\bar{\bar{y}}_H$ se aproxima do estimador $\bar{\bar{y}}_r$.

5) O estudo que fizemos dos estimadores r, $\bar{r}$ e $\bar{r}_H$, em conexão com a amostragem simples, pode ser utilizado no estudo da subamostragem com unidades de grandezas desiguais. Se as unidades de segunda ordem também forem conglomerados, então o estudo que fizemos dos estimadores $\bar{\bar{y}}_r$, $\bar{\bar{y}}_{\bar{r}}$ e $\bar{\bar{y}}_H$ poderia ser utilizado no estudo da subamostragem, com unidades de primeira e segunda ordem, sendo ambas de grandezas desiguais.

B I B L I O G R A F I A

[1] Cochran, W.G., Sampling Techniques, John Willey and Sons, New York, 1953.

[2] Cook, M.B., "Bi-variate K-Statistics and Cumulants of their Joint Sampling Distributions", Biometrika, nº 38 1953, pág. 179-195.

[3] Cramer, H., Metodos Matematicos de Estadistica, Aguilar S.A., - Madrid, 1963.

[4] Deming, W.E., Some Theory of Sampling, John Willey and Sons, N. York, 1950.

[5] Fava, Lindo, Contribuição para o Estudo da Estimativa Razão, S. Paulo, 1959.

[6] Feller, W., An Introduction to Probability Theory and its Applications, Vol. I, John Wiley and Sons, N.York, 1968.

[7] Hansen, M.H., Hurwitz, W.N. e Madow, W.G., Sample Survey Methods and Theory, Vol. I, John Wiley and Sons, N. York, 1966.

[8] Hansen, M.H., Hurwitz, W.N. e Madow, W.G., Sample Survey Methods and Theory, Vol. II, John Wiley and Sons, N. York, 1966.

[9] Hartley, H.O e Goodman, L.A., "The Precision of Umbiased Ratio--type Estimators", in Journal of the American Statistical Association, Vol. 35, nº 282, 1958, pág 270-271.

[10] Hartley, H.O e Ross, A., "Unbiased Ratio Estimators", in *Nature*, Vol. 174, 1954, pg. 270-271.

[11] Mood, A.M. e Graybill, F.A., *Introduction to the Theory of Statistics*, McGraw-Hill Book, New York, 1963.

[12] Sukhatme, P.W., "Sampling Theory of Survey with Applications" *The Indian Society of Agricultural Statistics*, New Delhi, 1954.

ERRATA:

na pág.	linha	
1	6	leia-se: enumeração
3	13	onde se lê: $(\bar{\bar{x}}-\bar{X}).(\frac{\bar{\bar{Y}}}{\bar{X}^2})$, leia-se: $(\bar{\bar{x}}-\bar{X}).(-\frac{\bar{\bar{Y}}}{\bar{X}^2})$
7	7	leia-se: $\beta = \frac{\sum_{v=1}^{N} Y_v X_v}{\sum_{v=1}^{N} X_v^2}$
8	3	onde se lê: (RS_x-S_{xy}), leia-se: $(RS_x^2-S_{xy})$
9	16	onde se lê: R, leia-se: $\bar{R}$
10	10	onde se lê: $\sigma^2- 2R\sigma_{xy}$, leia-se: $\sigma_y^2-2\bar{R}\sigma_{xy}$
11	9	onde se lê: r e r_M, leia-se: r e $\bar{r}_M$
14	9	leia-se: $Y_{v.} = \sum_{u=1}^{N_v} Y_{vu}$
15	19	onde se lê: o ê, leia-se: ê o
20	14	onde se lê: S_y^2, leia-se: $S_{\hat{y}}^2$
22	17	onde se lê: $\bar{\bar{y}}$, leia-se : $\bar{\bar{Y}}$
23	3	onde se lê: $\bar{\bar{Y}}_R S_{\hat{n}}$, leia-se: $\bar{\bar{Y}}_R S_{\hat{n}}^2$

continuação ERRATA:

na pág.	linha	
24	8	onde se lê: $\frac{S_{\bar{\bar{n}}}}{\bar{\bar{N}}} = s.v.(\bar{\bar{n}})$, leia-se: $\frac{S_{\bar{\bar{n}}}}{\bar{\bar{N}}} = c.v.(\bar{\bar{n}})$
26	7	onde se lê: $\sigma_{\dot{y}/\dot{n},\dot{n}}$, leia-se: $\sigma^2_{\dot{y}/\dot{n},\dot{n}}$
26	13	onde se lê: $\bar{\bar{Y}}_R \, S_{\dot{n}}$, leia-se: $\bar{\bar{Y}}^2_{\bar{R}} \, S^2_{\dot{n}}$
28	2	onde se lê: $\frac{m}{(m-1)\,\bar{\bar{r}}}$, leia-se: $\frac{m}{(m-1)\,\bar{\bar{n}}}$
29	2	onde se lê: $\bar{\bar{Y}}_R$, leia-se: $\bar{\bar{Y}}_{\bar{R}}$
29	4	onde se lê: $\bar{\bar{\ }}_H$, leia-se: $\bar{\bar{Y}}_H$