AGRICULTURA EM SÃO PAULO
Boletim Técnico do Instituto de Economia Agrícola

Ano 36 Tomo 2 1989

ANÁLISE DO VIÉS EM ALGUNS PROCEDIMENTOS PARA FALTA DE RESPOSTA E PARA ERROS DE RESPOSTA EM LEVANTAMENTOS POR AMOSTRAGEM([1])

Francisco Alberto Pino([2])

RESUMO

Estudam-se viéses na estimação da média e da variância em amostras estratificadas com falta de resposta. Alguns procedimentos práticos para falta de resposta são analisados do ponto de vista do viés, a saber: levantamento dos não respondentes, subamostragem dos não respondentes, substituição pela média dos respondentes, substituição por elementos de outro estrato, eliminação dos não respondentes, anulação dos não respondentes, uso de modelos, alteração do espaço amostral e união de estratos. Conclui-se que o único procedimento que não traz problemas é o levantamento dos não respondentes.

ANALYSING BIAS IN SOME PROCEDURES FOR NONRESPONSE AND RESPONSE ERRORS IN SAMPLE SURVEYS

SUMMARY

Mean and variance estimation biases in stratified samples with nonresponse are studied. Some practical procedures for nonresponse are treated from bias viewpoint, namely: non-respondent survey, non-respondent subsampling, substitution by the respondent mean, substitution by other stratum elements, elimination of non-respondents, substitution of non-respondents by zeroes, use of models, sample space change and union of two strata. The non-respondent survey is the unique non-problematical approach.

([1]) Recebido em 16/05/89. Liberado para publicação em 28/08/89.
([2]) Pesquisador Científico do Instituto de Economia Agrícola (IEA).

1 - INTRODUÇÃO

O problema da falta de resposta aparece freqüentemente nos levantamentos por amostragem. Mesmo a ocorrência de erros de resposta pode ser considerada um caso de falta parcial de resposta. Em alguns casos, como em levantamentos agrícolas, a questão chega a ser preocupante PINO & CASER (3). O objetivo do presente trabalho é analisar diversos procedimentos utilizados para resolver ou contornar tais problemas, principalmente quanto aos viéses que possam surgir e quanto às suposições que estejam por trás de cada procedimento. Alguns dos procedimentos aqui apresentados não costumam ser recomendados pelos estatísticos, mas, são corriqueiramente usados na prática. Eles são aqui analisados exatamente para advertir seus usuários sobre as conseqüências e os perigos envolvidos em sua utilização.

1.1 - Notação

Considere-se a seguinte situação: numa população finita Ω de N elementos pretende-se estudar a variável aleatória X, com média μ e variância σ^2, com base numa amostra probabilística A de n elementos. Particiona-se o espaço amostral em H estratos:

$$\Omega = \bigcup_{h=1}^{H} \Omega_h, \text{ com } \Omega_i \cap \Omega_j = \emptyset \quad \text{para } i \neq j$$

Seja N_h o número de elementos de Ω_h:

$$N = \sum_{h=1}^{H} N_h. \qquad (1)$$

Sejam $A_h \subset \Omega_h$ o conjunto de elementos da amostra no h-ésimo estrato e n_h o número de elementos de A_h, de tal forma que

$$A = \bigcup_h A_h \quad e \quad n = \sum_h n_h \qquad (2)$$

Sendo I a função indicador,

$$I_T(\omega) = 1, \text{ se } \omega \in T$$

$$= 0, \text{ se } \omega \notin T,$$

definem-se as variáveis aleatórias

$$X_h(\omega) = X(\omega) \quad I_{\Omega_h}(\omega), h = 1, ..., H. \qquad (3)$$

com média μ_h e variância σ_h^2, de tal modo que

$$\mu = \frac{1}{N} \sum_h N_h \quad \mu_h \qquad (4)$$

A estimativa não viesada da média em cada estrato (em A_h) é dada por

$$\bar{X}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi}, \text{ com } E(\bar{X}_h) = \mu_h, \qquad (5)$$

onde E representa a esperança matemática.

A estimativa não viesada da média geral é dada por

$$\bar{X} = \frac{1}{N} \Sigma_h N_h \bar{X}_h, \text{ com } E(\bar{X}) = \mu \qquad (6)$$

A estimativa não viesada da variância dos elementos em cada estrato é dada por

$$s_h^2 = \frac{1}{n_h - 1} \left(\sum_{i=1}^{n_h} X_{hi}^2 - n_h \bar{X}_h^2 \right), \qquad (7)$$

com $E(s_h^2) = \sigma_h^2$.

O problema da falta de resposta pode ser expresso da maneira que segue. Supondo-se que no estrato k alguns elementos da amostra não forneçam respostas, sejam B_{k1} o conjunto dos respondentes e B_{k2} o conjunto dos não respondentes, com n_{k1} e n_{k2} elementos, respectivamente. Então,

$$A_k = B_{k1} \cup B_{k2}, \quad B_{k1} \cap B_{k2} = \emptyset \text{ e}$$
$$n_k = n_{k1} + n_{k2} \qquad (8)$$

Definem-se as médias amostrais

$$\bar{X}_{kj} = \frac{1}{n_{kj}} \sum_{i=1}^{n_k} X_{ki}(\omega) I_{B_{kj}}(\omega), j=1,2. \qquad (9)$$

Então, (5) pode ser escrita como

$$\bar{X}_k = \frac{1}{n_k}(n_{k1} \bar{X}_{k1} + n_{k2} \bar{X}_{k2}) \qquad (10)$$

onde $\bar{X}_{k2}$ não é conhecida.

Finalmente, a seguinte notação será utilizada:

$$E[Z] = E(Z \mid B_{k1}) \qquad (11)$$

para esperança condicional e

$$\text{Viés}[Z] = \nu - E[Z], \qquad (12)$$

onde ν é o valor a ser estimado por Z.

2 - PROCEDIMENTOS QUE NÃO ALTERAM A ESTRATIFICAÇÃO

A maior parte dos procedimentos procura não alterar a estratificação. Neste caso, uma vez que não se dispõe de $\bar{X}_{k2}$, procura-se substituí-lo por um valor $\bar{W}$, calculado sobre n_w elementos e tal que $E(\bar{W}) = \mu_w$.

Analisam-se, a seguir, as estimações da média do estrato, da média geral e da variância dos elementos do estrato no caso geral e depois aplicam-se procedimentos particulares.

2.1 - Estimação da Média do Estrato

Para estimar μ_k usa-se $\bar{Y}_k$, definido por

$$\bar{Y}_k = \frac{1}{n_k}(n_{k1}\bar{X}_{k1} + n_{k2}\bar{W}) \qquad (13)$$

Somando-se e subtraindo-se $n_{n2}\bar{X}_{k2}$ dentro dos parênteses, obtêm-se:

$$\bar{Y}_k = \bar{X}_k - \frac{n_{k2}}{n_k}(\bar{X}_{k2} - \bar{W}) \qquad (14)$$

Então:

$$E\left[Y_k\right] = \mu_k - \frac{n_{k2}}{n_k} E\left[\bar{X}_{k2} - \bar{W}\right] \text{ e} \qquad (15)$$

$$\text{Viés}\left[Y_k\right] = \frac{n_{k2}}{n_k} E\left[\bar{X}_{k2} - \bar{W}\right] \qquad (16)$$

De (16) conclui-se que, na estimação da média do estrato:

a) $\bar{Y}_k$ é estimador não viesado de μ_k se

$$E[\bar{W}] = E[\bar{X}_{k2}]$$

b) o viés é diretamente proporcional ao número de elementos faltosos, isto é, quanto maior for o número de não respondentes, maior será o viés;

c) fixado o número de faltosos, o viés é inversamente proporcional ao número de elementos da amostra no estrato, isto é, quanto menor for a amostra do estrato, maior será o viés; e

d) se $E[\bar{W}] < E[\bar{X}_{k2}]$, a média do estrato será subestimada, caso contrário, será superestimada.

2.2 - Estimação da Média Geral

Para estimar μ usa-se $\bar{Y}$, definido por

$$\bar{Y} = \frac{1}{N}\left(\sum_{h \neq k} N_h \bar{X}_h + N_k \bar{Y}_k\right) = \bar{X} - \frac{N_k}{N} \cdot \frac{n_{k2}}{n_k}(\bar{X}_{k2} - \bar{W}) \qquad (17)$$

Então,

$$E[Y] = \mu - \frac{n_k}{N} \cdot \frac{n_{k2}}{n_k} E[\bar{X}_{k2} - \bar{W}] \text{ e} \qquad (18)$$

$$\text{Viés}[\bar{Y}] = \frac{N_k}{N} \cdot \frac{n_{k2}}{n_k} E[\bar{X}_{k2} - \bar{W}] = \frac{N_k}{N} \text{Viés}[\bar{Y}_k] \leq \text{Viés}[\bar{Y}_k] \qquad (19)$$

De (19) conclui-se que, na estimação da média geral:

a) $\bar{Y}$ é estimador não viesado de μ se $E[\bar{W}] = E[\bar{X}_{k2}]$;

b) o viés é diretamente proporcional ao número de elementos faltosos;

c) fixado o número de faltosos, o viés é inversamente proporcional à fração amostral do estrato dada por n_k/N_k, isto é, quanto menor for essa fração amostral, maior será o viés;

d) fixado o número de faltosos, o viés é diretamente proporcional ao tamanho relativo do estrato, dado por N_h/N; e

e) o viés da estimativa da média geral é sempre menor ou igual ao viés da estimativa da média do estrato com falta de resposta.

2.3 - Estimação da Variância dos Elementos do Estrato

Para estimar σ_k^2 usa-se V_k, definido por

$$V_k = \frac{1}{n_k-1}\left[\sum_{i=1}^{n_k} X_{ki}^2\,(\omega)\; I_{B_{k1}}\,(\omega) + \sum_{m=1}^{n_w} W_m^2 - \right.$$

$$\left. - n_k\,\bar{Y}_k^2\right] \qquad (20)$$

Somando e subtraindo dentro dos colchetes

$$\sum_{i=1}^{n_k} X_{ki}^2(\omega)\;\; I_{B_{k2}}(\omega) - n_k\;\; \bar{X}_k^2 \qquad \text{em (20)}$$

e usando (10) e (13), obtém-se

$$V_k = s_k^2 + \frac{1}{n_k-1}\left\{\sum_m W_m^2 - \Sigma_i\, X_{ki}^2\,(\omega)\; I_{B_{k2}}\,(\omega) - \right.$$

$$- \frac{n_{k2}}{n_k}\;(\bar{W} - \bar{X}_{k2})\;[n_{k2}(\bar{W} + \bar{X}_{k2}) + 2n_{k1}$$

$$\left.\bar{X}_{k1}]\right\} \qquad (21)$$

De (21) conclui-se que, na estimação da variância dos elementos do estrato:

a) a condição para que a estimativa da variância seja não viesada é bem mais complexa que para a média;
b) mesmo que as estimativas das médias sejam não viesadas, ainda assim, a estimativa da variância poderá ser viesada; neste caso,

$$\text{Viés}[V_k] = \frac{1}{n_k-1}\sum_m E\left[(W_m^2)\right] - \sum_i E\left[(X_{ki}^2)\right] I_{B_{k2}}(\omega); \qquad (22)$$

c) V_k é estimador não viesado de σ_k^2

se $E[\bar{W}] = E[\bar{X}_{k2}]$ e

$$E\left[\sum_m W_m^2\right] = E\left[\sum_i (X_{ki}^2)\;(\omega)\; I_{B_{k2}}(\omega)\right]; \qquad (23)$$

A seguir analisam-se procedimentos particulares em relação a (16), (19) e (21).

2.4 – Levantamento dos Não Respondentes

Consiste em voltar ao campo e levantar todos os não respondentes. Neste caso, $n_w = n_{k2}$, $\bar{W} = \bar{X}_{k2}$ e não há vieses, porque a amostra se recompõe. É o melhor método a ser adotado, mas, às vezes é difícil ou caro demais para ser utilizado.

2.5 – Levantamento de Subamostra dos Não Respondentes

Quando o levantamento de todos os não respondentes for muito difícil ou muito caro, pode-se tomar uma subamostra BARTHOLOMEW(1). Assim, aumenta-se n_{k1} e diminui-se n_{k2} fazendo uma segunda visita. Se esta for aleatória, então $\bar{W}$ será não viesado para estimar $E[\bar{X}_{k2}]$, isto é,

$$\text{Viés}\,[\bar{Y}_k] = \text{Viés}\,[\bar{Y}] = 0. \qquad (24)$$

Porém, o viés da estimativa da variância será

$$\text{Viés}\left[V_k\right] = -\frac{1}{n_k-1}\sum_i E\left[X_{ki}^2\right] I_C(\omega) \qquad (25)$$

onde C é o conjunto de elementos do estrato que não respondem na primeira nem na segunda visita. Logo, a variância será superestimada. Se a suposição de aleatoriedade da subamostra não puder ser garantida, os viéses aparecerão.

Um procedimento para selecionar a subamostra de não respondentes com fração de subamostragem proporcional à razão amostral de falta de resposta foi apresentada por SRINATH (4).

2.6 – Substituição pela Média dos Respondentes

Neste caso, $\bar{W} = \bar{X}_{k1}$. Então,

$$\text{Viés}\,[\bar{Y}_k] = \frac{n_{k2}}{n_k}\;E[\bar{X}_{k2} - \bar{X}_{k1}]\;\text{e} \qquad (26)$$

$$V_K = \frac{1}{n_k-1}\sum_{i=1}^{n_k}\left[X_{ki}^2\,(\omega) - I_{B_{k1}}(\omega)\;\bar{X}_{k1}^2\right] \qquad (27)$$

Este procedimento produz estimativas das médias não viesadas se os valores a serem informados pelos não respondentes, em média, forem semelhantes aos dos respondentes. Tal

suposição é razoável quando a falta de resposta ocorre de forma aleatória em relação ao valor das respostas, isto é, quando a falta de resposta não depende do valor a ser informado. Por outro lado, note-se que sempre teremos $V_k \leq s_k^2$, isto é, a variância do estrato será sempre subestimada. Como conseqüência, a variância da estimativa da média também será subestimada, dando a falsa impressão de ter havido aumento de precisão da estimativa. Esta, talvez seja a conseqüência mais grave deste método para o usuário desavisado.

2.7 – Substituição por Elementos de Outro Estrato

Consiste em substituir os não respondentes por elementos de outro estrato, eventualmente sorteados aleatoriamente. Se o m-ésimo estrato fornecer os elementos para substituição, então $\overline{W} = \bar{X}_{m2}$.

Este procedimento só será razoável se existirem elementos semelhantes em dois estratos diferentes, o que é pouco provável que aconteça numa amostra estratificada. Entretanto, tal pode acontecer numa amostra em que pares de elementos sejam sorteados de maneira sistemática. Então,

$$\text{Viés}[\bar{Y}_k] = \frac{n_{k2}}{n_k} E[\bar{X}_{k2} - \bar{X}_{m2}] \qquad (28)$$

A estimativa da variância será viesada na maioria dos casos.

2.8 – Eliminação dos Não Respondentes

Consiste em diminuir o tamanho da amostra, utilizando somente os respondentes. Neste caso, as estimativas das médias e seus viéses são as mesmas do procedimento de substituição pela média dos respondentes (item 2.6). Entretanto, a estimativa da variância será diferente:

$$V_k = \frac{1}{n_{k1}-1}\left[\sum_{i=1}^{n_k} X_{ki}^2(\omega)\, I_{B_{k1}}(\omega) - n_{k1}\ \bar{X}_{k1}^2\right] \qquad (29)$$

Se os valores a serem informados pelos não respondentes forem, em média, semelhantes aos informados pelos respondentes, então, as estimativas das médias serão não viesadas, mas, possivelmente, menos precisas, por causa da diminuição do tamanho da amostra.

2.9 – Anulação dos Não Respondentes

Consiste em ignorar o problema, tomando os não respondentes como valores nulos. Neste caso, $\overline{W} = 0$.

As estimativas serão não viesadas se os não respondentes não tiverem a característica que está sendo levantada. Este procedimento poderá ser utilizado quando soubermos de antemão que no estrato k a média é nula: $\mu_k = 0$. É o que acontece, por exemplo, quando levantamos a produção de café no Estado de São Paulo, na região do Litoral.

2.10 – Uso de Modelos para Estimação

Sob certas suposições, casos de falta parcial de resposta podem ser resolvidos usando-se um modelo para sua estimação (PINO, 2). Mas, este procedimento somente será válido se suas suposições estiverem satisfeitas.

3 - PROCEDIMENTOS QUE ALTERAM A ESTRATIFICAÇÃO

Apresentam-se dois procedimentos, um que altera a estratificação a nível da população e outro que a altera a nível da amostra.

3.1 - Diminuição do Espaço Amostral

Consiste em utilizar a amostra para fazer inferências somente sobre a parte do espaço amostral onde não há falta de resposta. Divide-se o estrato k em duas partes:

$$\Omega_k = \Omega_{k1} \cup \Omega_{k2}, \text{ com } \Omega_{k1} \cap \Omega_{k2} = \emptyset.$$

Seja N_{k1} o número de elementos de Ω_{k1} e

$$\mu_k = \frac{1}{N_k}\left[N_{k1}\mu_{k1} + (N_k - N_{k1})\,\mu_{k2}\right] \quad (30)$$

Neste caso, $\bar{Y}_k = \bar{X}_{k1}$ é não viesado para estimar μ_{k1}. Entretanto, não podemos determinar exatamente que elementos constituem Ω_{k1} nem Ω_{k2}. Também não conhecemos N_{k1}, que pode ser estimado por

$$\hat{N}_{k1} = \left(1 - \frac{n_{k2}}{n_k}\right) N_k = \frac{n_{k1}}{n_k}\, N_k, \quad (31)$$

com $E\left[\hat{N}_{k1}\right] = N_{k1}$.

Então,

$$\bar{Y} = \frac{1}{N - N_k + \hat{N}_{k1}}\left(\sum_{h\neq k} N_h\, \bar{X}_h + \hat{N}_{k1}\bar{Y}_{k1}\right) \quad (32)$$

será não viesado para estimar

$$\mu^* = \frac{1}{N - N_k + N_{k1}}\left(\sum_{h\neq k} N_h\mu_h + N_{k1}\mu_{k1}\right) \quad (33)$$

se N_{k1} for conhecido.

Logo, este procedimento somente poderá ser considerado razoável se N_{k1} for muito próximo de N_k ou se N_{k1} for conhecido.

3.2 – União de Estratos

Consiste em unir os respondentes a outro estrato. É comumente usado quando n_{k1} é muito pequeno. Se unirmos o estrato k ao estrato m, então,

$$\bar{Y}_m = \frac{1}{n_m + n_{nk1}}\,(n_m\bar{X}_m + n_{k1}X_{k1}) \text{ com } (34)$$

$$\text{Viés}\,[\bar{Y}_m] = \frac{(n_mN_k - n_{k1}N_m)\,(\mu_m - \mu_k)}{(N_k+N_m)(n_{k1}+n_m)} -$$

$$- \frac{n_{k1}}{n_{k1} + n_m}\; E[\bar{X}_{k2}]; \quad (35)$$

$$\bar{Y} = \frac{1}{N}\left[\sum_{h\neq k,m} N_h\bar{X}_h + (N_k + N_m)\,\bar{Y}_m\right]$$

com (36)

$$\text{Viés}[\bar{Y}] = \frac{1}{N(n_m + n_{k1})}\left\{(n_mN_k - n_{k1}N_m)\right.$$

$$\left.(\mu_k - \mu_m) - n_{k1}(N_k + N_m)\,E[(\bar{X}_{k2})]\right\} \quad (37)$$

As estimativas das médias serão não viesadas se

$$E[\bar{X}_{k2}] = \frac{n_{k1}}{N_K + N_m}(n_mN_k - n_{k1}N_m)\,(\mu_m -$$

$$\mu_k), \quad (38)$$

isto é, se a média dos não respondentes na população for uma particular proporção da diferença entre as médias dos estratos m e k. Isso, eventualmente, poderá acontecer se os elementos forem ordenados pela variável em estudo e sorteados sistematicamente, em pares.

4 – CONCLUSÕES

O único procedimento que não traz problemas é o de levantamento dos não respondentes (2.4). Somente se tolera o uso de outro procedimento quando: a) o procedimento (2.4) for muito caro ou de difícil execução (por exemplo, se os dados foram levantados em passado distante); b) os pressupostos de alguns dos outros procedimentos forem válidos. Logo, a escolha do procedimento deverá recair sobre o mais apropriado para cada caso. Um estudo piloto dos não respondentes poderá indicar o caminho a seguir.

Finalmente, algumas comparações entre métodos podem ser feitas. Fixados os faltosos;

a) o procedimento de subamostra dos não respondentes (2.5) terá viés nulo na estimação das médias se a segunda visita for aleatória; os procedimentos de média (2.6) e eliminação (2.8), se os não respondentes forem semelhantes aos respondentes; o procedi-

dimento de outro estrato (2.7), se dois estratos forem semelhantes; o procedimento de anulação (2.9), se a média do estrato for nula; o procedimento de união de estratos (3.2), se os não respondentes forem intermediários entre dois estratos;

b) se os estratos forem construídos normalmente, isto é, diferentes entre si, então, o procedimento de outro estrato (2.7) deverá ter viés maior ou igual aos dos procedimentos de média (2.6) e de eliminação (2.8);

c) os procedimentos de outro estrato (2.7) e de união de estratos (3.2) fazem mais sentido em amostragem sistemática de pares;

d) nos procedimentos de levantamentos (2.4) de média (2.6) de anulação (2.9) de outros estratos, se $n_{m2} = n_{k2}$ (2.7), as variâncias das estimativas continuarão as mesmas, embora suas estimativas possam ser viesadas:

$$V(\bar{Y}_k) = \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \sigma_k^2; \qquad (39)$$

nos outros procedimentos as variâncias das estimativas, provavelmente aumentarão; no procedimento de média (2.6), a variância é sempre subestimada, causando a ilusão de que a precisão da estimativa melhorou;

e) as condições para que a estimativa da variância dos elementos do estrato seja não viesada são bem mais complexas que para a estimativa da média do estrato.

Finalmente, recomenda-se aos usuários muito cuidado ao estudar cada caso particular porque o efeito de procedimentos práticos não devidamente analisados podem ser bastante adversos.

LITERATURA CITADA

1. BARTHOLOMEW, D.J. A method of allowing for 'not-at-home' bias in sample surveys. **Applied Statistics**, 10:52-59, 1961.

2. PINO, Francisco A. Detecção e correção de erros em levantamentos agrícolas. **Pesquisa Agropecuária Brasileira**, Brasília, **21**(9):979-985, set. 1986

3. ——— & CASER, Denise V. **Falta de resposta em levantamentos por amostragem**: um estudo de caso. São Paulo, Secretaria de Agricultura e Abastecimento, IEA, 1984. 25p. (Relatório de Pesquisa, 08/84)

4. SRINATH, K.P. El muestreo multifásico en los problemas de falta de respuesta. **Estadística**, Santiago, **28**(107):196-203, jun. 1970.