معالجة عدم الاستجابة في المعاينة المزدوجة للطبقية

زين العابدين البشير

أستاذ مشارك ـ قسم الأساليب الكمية ـ كلية العلوم الإدارية ـ جامعة الملك سعود ـ الرياض ـ المملكة العربية السعودية

(قُدم للنشر في ١٤١٣/١١/٢٦هـ؛ وقبل للنشر في ١٤١٤/٥/٣هـ)

ملخص البحث. تتناول هذه الورقة تقدير مجموع المجتمع في المعاينة المزدوجة بغرض الطبقية في وجود عدم استجابة. تقدم الدراسة صيغًا للتباين ومقداره، وتوضح أن هذه الصيغ تأخذ الشكل المألوف عندما تكون الاستجابة مكتملة.

[4] El-Beshir, Z.A. "Nonresponse in Double Sampling for Regression." *J. King Saud University (Admin. Sci.)*, 7, No. 1 (1995), 19-36.

[5] Rao, J.N.K. On Double Sampling for Stratification and Analytical Surveys. *Biometrika*, 60, (1973), 125-133.

[6] Cochran, W.G. *Sampling Techniques, 3rd ed.* New York: Wiley, 1977.

where $\bar{y}'_{rh}$ is the mean of the y'_k in the responding set in s^*_{2h} and W^*_h the relative size of s^*_{2h}. If in addition sampling at phase one is simple random, (7.1) takes the form.

$$N\sum_h W^*_h \bar{y}_{r_h}$$

On the other hand, only the second and third terms in the variance formula (5.4) are affected by the type of stratification at phase two. Thus the second term reduces to

$$E_a\left[\frac{n_1(n_1-n_2)}{n_2}\sum_v W_v S^2\left(y'_{s_{1v}}\right)\right]$$

and the third term becomes

$$E_a E_{ST}\left[E_{\underline{m}}\left\{n_1^2\sum_h^{H_{s_2}} W^{*2}_h(1-f_{2h})S^2\left(y'_{s*2h}\right)/m_h\right\}\middle| s_1\right]$$

where $S^2(y'_{s_{2h}})$ is the variance of the y^*_k for the k's in s^*_{2h}.

As for variance estimator $\hat{V}(t_{c\pi^{**}})$ simplification resulting from proportional allocation is not appreciable. Some simplification is achieved in $\hat{V}_2$ by replacing f_{1v} by n_2/n_1. Also we have

$$\hat{V}_3 = n_1^2\sum_h W^{*2}_h(1-f_{2h})S^2\left(y'_{s*2h}\right)/m_h$$

for the third component. A more simple expression for this component is attained when the first-phase sample is obtained through simple random sampling.

References

[1] Neyman, J. "Contribution to the Theory of Sampling Human Populations." *J. Am. Statist. Assoc.*, (1983), 101-116.

[2] Särndal, C.E. and Swensson, B. "A General View of Estimation for Two-phases of Selection with Applications to Two-phase Sampling and Nonresponse." *Int. Statist. Rev.* 55, (1987), 279-294.

[3] Oh, H.L. and Scheuren, F.J. "Weighting Adjustment for Unit Nonresponse." In: *Incomplete Data in Sample Surveys,* 2nd ed. W.G. Madow. I. Olkin and D.B. Rubin, pp. 143-184. New York: Academic Press, 1983.

where $\hat{V}^*_{a1}$ and $\hat{V}^*_{a12}$ are given by (6.4) and (6.9) respectively. It is easy to see that (6.15) can be put in the form:

$$\hat{V}^*_{SI} = N(N-1)\sum_{v}^{H_{s_1}}\left\{\frac{n_{1v}-1}{n_1-1} - \frac{n_{2v}-1}{N-1}\right\}W_v S^2\left(y_{s_{2v}}\right)/n_{2v}$$

$$+ \frac{N(N-n_1)}{n_1-1}\sum_{v}^{H_{s_1}} W_v\left(\bar{y}_{s_{2v}} - \hat{\mu}\right)^2$$

This last form of variance estimator is the form given by Rao [5], Cochran [6, p.334] and Särndal and Swensson [2] for double sampling for stratification with complete response. We note, however, that Rao and Cochran gave their result for the conceptually different situation in which it is assumed that there exists a predetermined set of strata prior to the selection of the first-phase sample s_1.

In single-phase sampling, Särndal and Swensson pointed out that the approach of forming the strata in the light of the selected sample has the attractive feature of enabling the sampler to take into account the survey operations to which units in the selected sample are exposed. In double sampling, adopting this approach enables the sampler to take into account information in s_1 in forming the s_{1v} and survey operations to which units in s_2 are exposed in forming the s^*_{2h}.

Thus response rate may be affected by such factors as interviewer's, skill, age, sex or race. It may also be affected by the respondent's age, sex, profession ... etc. This can be reflected in the response model by partitioning s_2 into groups that correspond to interviewers or, if possible, to interviewers crossed with respondent age - sex-profession groups.

7. Proportional Allocation

If the stratified sample in phase two is performed with proportional allocation some simplification is gained in the formulae for the estimator $t_{c\pi^{**}}$, its variance and estimator of variance. Thus since in this case

$$f_{1v} = n_2/n_1 \qquad V = 1,\ldots,H_{s1}$$

the estimator $t_{c\pi^{**}}$ given by (4.1) becomes:

$$n_1\sum_{h} W^*_h \bar{y}'_{r_h} \tag{7.1}$$

$$\sum_{r}\sum \Delta_{kL|s_1 s_2 \underline{m}} C_k C_L = \sum_{h}\Big[\Delta_{k(h)} \sum_{h}\sum_{r_{vh}} C_k^2 + \Delta_{kL(h)} \sum_{v}\sum_{r_{vh}} C_k C_{L\neq k}$$

$$+ \Delta_{kL(h)} \sum_{v}\sum_{r_{vh}} C_k \sum_{v'\neq v}\sum_{r_{v'h}} C_k \Big] \qquad (6.10)$$

Using (6.10) together with the definition of the inclusion probabilities, the third component in (3.5) takes the simple form:

$$\hat{V}_3 = n_2^2 \sum_{h} W_h^{*2} (1 - f_{2h})\, S^2 y^{"}_{s_{rh}} / m_h \qquad (6.11)$$

Note that this term vanishes with full response.

Example 4.

If the first-phase sample is simple random of size n_1, (6.11) becomes

$$\hat{V}_{SI3} = n_2^2 f^{-2} \sum_{h} W_h^{*2} (1 - f_{2h})\, S^2 \left(y^{*}_{s_{rh}}\right) / m_h \qquad (6.12)$$

using the remark following (5.6).

The full estimator of the variance $Vt_{c\pi^{**}})$ in double sampling for stratification in the presence of nonresponse under the ERHG model and given $m_h \geq 1$ (all h) is then:

$$\hat{V}_{(\hat{t}_{c\pi^{*\theta}})} = \hat{V}_1 + \hat{V}_2 + \hat{V}_3 \qquad (6.13)$$

with $\hat{V}_1$, $\hat{V}_2$ and $\hat{V}_3$ given by (6.1), (6.7) and (6.11) respectively.

When sampling in the first phase is simple random (6.13) takes the form

$$\hat{V}_{s1} = \hat{V}_{s11} + \hat{V}_{s12} + \hat{V}_{s13} \qquad (6.14)$$

with the three terms given by (6.2), (6.8) and (6.12) respectively. If in addition we have complete response $\hat{V}_{s13}$ is zero and (6.14) becomes

$$\hat{V}^{*}_{SI} = \hat{V}^{*}_{S11} + \hat{V}^{*}_{S12} \qquad (6.15)$$

where $\Delta_{k(v)}$ denotes $\Delta_{kk|s_1}$ for $k \varepsilon s_{2v}$ and $\Delta_{kL(v)}$ denotes $\Delta_{kk|s_1}$ for $k, L \varepsilon s_{2v}$. Using (6.5) and remembering (2.1) and (3.2), the expression (6.5) becomes:

$$\hat{V}_2 = \sum_v f_{1v}^{-2} \frac{(1-f_{1v})}{(n_{2v}-1)} \left[\sum_h f_{2h}^{-1} \left\{ n_{2v} - 1 + \frac{(n_{2h}^{*}-1)}{(m_h-1)} \right\} \sum_{r_{vh}} y_k'^2 \right.$$

$$\left. - \sum_h f_{2h}^{-1} \left(\frac{(n_{2h}^{*}-1)}{(m_h-1)} - f_{2h}^{-1} \right) \left(\sum_{r_{vh}} y_k' \right)^2 - \left(\sum_h f_{2h}^{-1} \sum_{r_{vh}} y_k' \right)^2 \right] \qquad (6.7)$$

Example (3)

Let the first sample be obtained by simple random sampling. Then (6.7) takes the form

$$\hat{V}_{SI2} = \frac{N^2}{n_1^2} \hat{V}_2' \qquad (6.8)$$

where $\hat{V}_2'$ is of a form identical to that of $\hat{V}_2$ the only difference being that y_k' in $\hat{V}_2$ is replaced by y_k in $\hat{V}_2'$. If in addition we assume that reponse is complete and the strata in phase three are the s_{2v}'s then, noting (6.3) and the remark immediately preceding it, we see that (6.8) reduces to

$$\hat{V}_{SI2}^{*} = N^2 \sum_v W_v^2 (1-f_{1v}) S^2\left(y_{s_{2v}}\right)/n_{2v} \qquad (6.9)$$

a familiar form in the context of stratified sampling

6.3 Third component

The third component in (3.5) is an unbiased estimator of the third term in (3.4). It can be expressed in terms of stratifications of phase two and three given $\underline{m}$ and $m_h \geq 1$ (all h) in a way similar to that used for the second component. Thus observing that $\Delta_{kL|s_1 s_{2m}}$ is nonzero only for h = h' we have

So that (6.2) takes the relatively simple form:

$$\hat{V}^{*}_{SII} = \frac{N^2(1-f)}{n_1}\sum_{v} W_v\left[(1-Q_v)\, S^2_{(ys2v)} + \frac{n_1}{n_1-1}\left(\bar{y}_{s2v} - \hat{\mu}\right)^2\right] \qquad (6.4)$$

where:

$$Q_v = (n_1 - n_{1v})/\{(n_1-1)n_{2v}\}$$

and

$$\bar{y}_{s_{2v}} = \sum_{s_{2v}} y_k/n_{2v}\ ,\ \hat{\mu} = \sum_{v} W_v\, \bar{y}_{s_{2v}}$$

Here, $\bar{y}_{s_{2v}}$ is the sample mean for stratum s_{2v} and $\hat{\mu}$ is the weighted mean of strata sample means.

6.2 Second component

The estimator of the second component in (3.4) is the second component in (3.5). Given $\underline{m}$ and $m_h \geq 1$ for all h, this latter term is

$$\sum_{r}\sum \Delta_{kL|s_1}\, y''_k\, y''_L \Big/ \left(\pi_{kL|s_1} \quad \pi_{kL|s_1 s_2 \underline{m}}\right) \qquad (6.5)$$

Now, for $k \neq L$, $\Delta_{kL|s_1}$ is nonzero only if k and L belong to the same v. Hence we can put:

$$\sum_{r}\sum \Delta_{kL|s_1} c_k c_L = \sum_{v}\Big[\Delta_{k(v)}\sum_{h}\sum_{r_{vh}} c_k^2 + \Delta_{kL(v)}\sum_{h}\sum_{r_{vh}} c_k c_{L\neq k}$$

$$+ \Delta_{kL(v)}\sum_{h}\sum_{r_{vh}} c_k \sum_{h'\neq h}\sum_{r_{vh'}} c_k\Big] \qquad (6.6)$$

together with completing the square technique, and assuming the event $m_h \geq 1$ occurred, we see, after a straightforward but tedius algebra, that $\hat{V}_1$ takes the form:

$$\hat{V}_{SII} = \frac{N^2(1-f)}{n_1(n_1-1)} \sum_v \frac{W_v}{n_{2v}} \left[f_{1v}^{-1} \sum_h^{H_{s2}} f_{2h}^{-2} K_{vh} \left(\sum_{r_{vh}} y_k \right)^2 + \frac{n_{2v}}{n_1} \sum_h^{H_{s2}} f_{2h}^{-1} K_h G_{.h}^2 \right.$$

$$\left. + K_v G_{v.}^2 - \frac{n_{2v}}{n_1} \left(\sum_v f_{1v}^{-1} G_{v.} \right)^2 + \sum_h f_{2h}^{-1} D_{vh} \right] \qquad (6.2)$$

with:

$$K_{vh} = f_{2h} \frac{\left(n_{2h}^{*} - 1\right)}{\left(m_h - 1\right)} + f_{1v} \frac{\left(n_{1v} - 1\right)}{\left(n_{2v} - 1\right)} - 1$$

$$K_h = 1 - \left(n_{2h}^{*} - 1\right)/\left(m_h - 1\right), \quad K_v = f_{1v}^{-1} - \left(n_{1v} - 1\right)/\left(n_{2v} - 1\right)$$

It is of interest to see the form (6.2) takes when response is complete, and strata of phase three are themselves the s_{2v} drawn at phase two i.e. there is no subsampling at phase three and each s_{2h}^{*} is equal to one of the s_{2v} 's. This is the case of double sampling for stratification with full response. For such a situation we have, for any given h, $r_{vh} \equiv s_{vh}$ $(v = 1,\ldots,H_{s1})$ where s_{vh} is the empty set except for one v, say v', for which $s_{vh} \equiv s_{2v'}$. Hence $\sum_h \sum_{r_{vh}} (*)$ is equivalent to $\sum_{s_{2v}} (*)$ and we have:

$$G_{v.} = \sum_{s_{2v}} y_k$$

$$K_h = 0 \qquad (6.3)$$

$$K_{vh} = f_{1v}\left(n_{1v} - 1\right)/\left(n_{2v} - 1\right) - 1$$

6. Estimate of Variance

An expression for an unbiased estimator of $V(t_{c\pi}^{**})$ can be obtained through a term by term analysis of (3.5).

6.1 First component

From (3.5) an unbiased estimator of the first component in $(V\, t_{c\pi}^{**})$ given $\underline{m}$ and assuming $m_h \geq 1$ (all h) is

$$\hat{V}_1 = \sum_r \sum \Delta_{akL}\, y'_k\, y'_L \Big/ \left(\pi_{akL}\ \ \pi_{kL|s_1}\ \ \pi_{kL|s_1 s_2 \underline{m}} \right) \tag{6.1}$$

It is not possible to put (6.1) in a more explicit and more compact form. In fact example (2) below shows that even when the first-phase design is assumed, simple random sample of the form of (6.1) is still cumbersome.

Example (2)

Suppose the first-phase sample is a simple random sample of size n_1. Put

$$G_{v\cdot} = \sum_h^{H_{s_2}} f_{2h}^{-1} \sum_{r_{vh}} y_k \qquad G_{\cdot h} = \sum_v^{H_{s_1}} f_{1v}^{-1} \sum_{r_{vh}} y_k$$

and let

$$D_{vh} = (n_1 - 1) \sum_{r_{vh}} y_k^2 + \frac{(n_{1v} - 1)(n_{2h}^{*} - 1)}{(n_{2v} - 1)(m_h - 1)} \left\{ \sum_{r_{vh}} y_k^2 - \left(\sum_{r_{vh}} y_k \right)^2 \right\}$$

Then using the result

$$\sum_r \sum C_k C_L = \sum_v \sum_h \sum_{r_{vh}} C_k^2 + \sum_v \sum_h \sum_{r_{vh}} C_k C_{L \neq k} + \sum_v \sum_{v' \neq v} \sum_h \sum_{rv'h'} C_k \sum_{h \neq h'} \sum_{rvh} C_k$$

$$+ \sum_v \sum_{v' \neq v} \sum_h \sum_{r_{v'h}} C_k \sum_{r_{vh}} C_k + \sum_h \sum_{h' \neq h} \sum_v \sum_{rvh'} C_k \sum_{r_{vh}} C_k$$

The full variance expression is then, given $m_h \geq 1$ for all h,

$$V\left(\hat{t}_{c\pi^{**}}\right) = \sum_u \sum \Delta_{akL}\, y'_k\, y'_L + E_a\left[n_1^2 \sum_v^{H_{s_1}} W_v^2 (1-f_{1v}) S^2 y'_{s_{1v}} / n_{2v} \right]$$

$$+ E_a E_{ST}\left[E_{\underline{m}}\left\{ n_2^2 \sum_h^{H_{s_1}} W_h^{*2} (1-f_{2h}) S^2 y''_{s_{2h}} / m_h \right\} \middle| s_1 \right] \qquad (5.4)$$

where E_m indicates expectation over all realizations of $\underline{m}$, and E_{ST} denotes expectation over all stratified random samples in phase two given s_1. We recall the assumption made in §2 that exactly the same stratified design is used everytime the same s_1 is selected. Note that the second and third terms in (5.4) must be left in the form of an expected value since the conditional inclusion probabilities may depend on s_1 and s_2. The third term represents the nonresponse component of variance. It vanishes when $f_{2h} = 1$ i.e. when response is complete.

Example (1)

As an example, we consider double sampling for stratification with simple random sampling at phase one. In this case:

$$\pi_{ak} = \pi_{akk} = n_1/N = f \quad \pi_{akL} = f(n_1 - 1)/(N - 1) \quad k \neq L \qquad (5.5)$$

The variance (5.4) then becomes:

$$V\left(\hat{t}_{c\pi^{**}}\right) = N^2(1 - f)\, S^2(y_u)/n_1 + E_{S1}\left[N^2 \sum_v^{H_{s_1}} W_v^2 (1-f_{1v}) S^2\left(y_{s_{1v}}\right)/n_{2v} \right]$$

$$+ E_{S1} E_{ST}\left[E_{\underline{m}}\left\{ \frac{N^2 n_2^2}{n_1^2} \sum_h^{H_{s_2}} W_h^{*2} (1-f_{2h}) S^2\left(y^*_{s_{2h}}\right)/m_h \middle| s_1 \right\} \right] \qquad (5.6)$$

with $y^* = f_{1v}^{-1} y$ and E_{s1} denotes expectation over all simple random samples. A part from the third term, which disappears when response is complete, (5.6) is the usual form of variance of the total in double sampling for stratification with simple random sampling at phase one (see for example Särndal and Swensson [2].

so that using (2.1a) - (2.1c) and completing the square in the second term we finally get:

$$\sum_{s_1}\sum \Delta_{kL|s_1}\; y''_k\, y''_L \;=\; n_1^2 \sum_{v}^{H_{s_1}} W_v^2 (1-f_{1v}) S^2\, y'_{(s1v)} / n_{2v}$$

where in this paper we use $S^2(XA)$ to denote the variance of the numbers x_k in a set A of size n_A given by $\{\sum_A x_k^2 - (\sum_A x_k)^2/n_A\}/(n_A - 1)$.

The second term in the variance expression then becomes:

$$E_a \left\{ n_1^2 \sum_{v}^{H_{s_1}} W_v^2 (1-f_{1v}) S^2\, y'_{(s1v)} / n_{2v} \right\} \tag{5.1}$$

Let s_{vh} be the set of units in both s_{2v} and s^*_{2h}. The second-phase sample s_2 is the union of s_{vh} over all v and h. It is also the union of s^*_{2h} over all h. Noting that for any two units $k \varepsilon s_{vh}$ and $L \varepsilon s_{vh}$, $\Delta_{kL|s_1 s_2 \underline{m}}$ is nonzero only if $h = h'$ we can write assuming $m_h \geq 1$:

$$\sum_{s_2}\sum \Delta_{kL|s_1 s_2 \underline{m}}\; y'''_k\, y'''_L \;=\; \sum_{h}^{H_{s2}} \left\{ \sum_{v}^{H_{s_1}} \sum_{S_{vh}} \Delta_{kk|s_1 s_2 \underline{m}}\; y'''^2_k \;+\; \sum_{S^*_{2h}} \Delta_{kL|s_1 s_2 \underline{m}}\; y'''_k\, y'''_{L \neq k} \right\} \tag{5.2}$$

Using (3.2a), (3.2b), (3.2c) and (3.6) we can put the right hand side of (5.2) in the form:

$$n_2^2 \sum_{h}^{H_{s2}} W_h^{*2} (1-f_{2h}) S^2_{s^*_{2h}}(y''_*) / m_h \tag{5.3}$$

Where $W^*_h = n^*_{2h}/n_2$ and where $S^2(y''_{s^*_{2h}})$ can be expressed in terms of f_{1v} as

$$\left[\sum_{v}^{H_{s1}} \sum_{S_{vh}} \left(f_{1v}^{-1}\, y'_k \right)^2 - \left(\sum_{v}^{H_{s1}} \sum_{Svh} f_{1v}^{-1}\, y'_k \right)^2 \Big/ n^*_{2h} \right] \Big/ \left(n^*_{2h} - 1 \right)$$

$$\pi_{ak} = n_1 / N \tag{4.2}$$

and (4.1) becomes, remembering (3.6):

$$\hat{t}_{c\pi^{**}s1} = N \sum_{v}^{H_{s_1}} W_v \sum_{h}^{H_{s_2}} f_{2h}^{-1} \sum_{r_{vh}} y_k / n_{2v} \tag{4.3}$$

with $W_v = n_{1v}/n_1$ the relative size of stratum s_{1v}.

If in addition there is complete response, $f_{2h} = 1$ and $\sum_{h} \sum_{r_{vh}} (*)$ is equivalent to $\sum_{s_{2v}} (*)$ so that (4.3) reduces to the well known form for double sampling for stratification (with complete response):

$$\hat{t} = N \sum_{v}^{H_{s_2}} W_v \bar{y}_v \tag{4.4}$$

where $\bar{y}_v = \sum_{s_{2v}} y_k/n_{2v}$. Comparison of (4.3) and (4.4) shows that the adjustment to nonresponse takes the form of 'expanding' the totals of the responding sets r_{vh} ($v = 1,\ldots,H_{s_1}$) in the response homogeneity group s^*_{2h} by dividing each by the response probability of s^*_{2h}.

Variance Formula

To arrive at an expression for the variance of $t_{c\pi^{**}}$ that reflects the stratifications of phase two and three we first note that the first term in (3.4) depends only on the first-phase inclusion probabilities and hence will remain in the same form in the variance formula.

On the other hand, since for $k\varepsilon s_{1v'}$, $L\varepsilon s_{1v'}$, and $V \neq V'$, $\Delta_{kL|s_1} = 0$, the quantity within brackets in the second term of (3.4) can be written as:

$$\sum_{s_1}\sum \Delta_{kL|s_1}\, y''_k\, y''_L = \sum_{v}^{H_{s_1}} \Delta_{kk|s_1} \sum_{s_{1v}} y'^2_k / \pi^2_{k|s_1} + \sum_{v}^{H_{s_1}} \Delta_{kL|s_1} \sum\sum_{s_{1v}} y'_k\, y'_{L\neq k} \Big/ \left(\pi_{k|s_1}\, \pi_{L|s_1}\right)$$

and estimator of variance:

$$V(\hat{t}_{\pi^{**}}) = \sum_{r}\sum \Delta_{akL}\; y'_k\, y'_L/\pi^{**}_{kL} + \sum_{r}\sum \Delta_{kL|s_1}\, y''_k\, y''_L\Big/\left(\pi_{kL|s_1}\pi_{kL|s_1s_2}\right) + \sum_{r}\sum \Delta_{kL|s_1s_2}\, y'''_k\; y'''_L/\pi_{kL|s_1s_2} \tag{3.5}$$

where for a subset A of U we use $\sum\sum_{A} C_{kL}$ to denote $\sum_{k\varepsilon A}\sum_{L\varepsilon A} C_{kL}$. Also, in (3.3) - (3.5) we used "|" "||" "|||" "m" to refer to first-pahse, second-phase and third-phase expansion

respectively. Thus for example,

$$y'_k = y_k/\pi_{ak}\,(k \in U),\; y''_k = y'_k/\pi_{k|s_1}\;(k \in s_1),\; y'''_k = y''_k/\pi_{k|s_1s_2}\;(k \in s_2 \subset s_1) \tag{3.6}$$

The quantities $\pi_{k|s_1s_2}$ and $\pi_{kL|s_1s_2}$ indicate the third-phase inclusion probabilities of unit k ($k \varepsilon s_2 \subset s_1$) and units k&L ($k, L \varepsilon s_2 \subset s_1$) respectively. Also $\Delta_{kL|s_1s_2} = \pi_{kL|s_1s_2} - \pi_{k|s_1s_2}\,\pi_{L|s_1s_2}$. For convenience we put $\pi^{**}_{kL} = \pi_{akL}\,\pi_{kL|s_1}\,\pi_{kL|s_1{}^{s}{}_{2}}$. Finally E_a in (3.4) refers to expectation with respect to sampling distribution in phase one and E (./s_1) to expectation with respect to sampling distribution in phase two given s_1 is drawn in phase one.

A Conditional π^{**} – Expanded Sum Estimator for the Total

Given the set $\underline{m}$ of realized counts (or responses) and assuming $m_h \geq 1$ for h = 1,...,H_{s_2} an unbiased estimator of t can be obtained from (3.3) using (2.1a) and (3.2a) as

$$\hat{t}_{c\pi^{**}} = \sum_{v}^{H_{s_1}} f_{1v}^{-1} \sum_{h}^{H_{s_2}} f_{2h}^{-1} \sum_{r_{vh}} y'_k \tag{4.1}$$

where r_{vh} is the subset of units that belong to both stratum v (of phase two) and stratum h (of phase three).

If the first-phase sample is a simple random sample of size n_1 then

two and stratified Bernoulli sampling at phase three. This theory can directly be looked at as one for nonresponse in double sampling for stratification under the ERHG model, the only difference being that in the case of nonresponse the ERHG model is merely an assumption rather than an actually imposed randomization scheme.

Suppose now, that we have a stratified Bernoulli sampling at phase three with strata s^*_{2h} ($h = 1,\ldots,H_{s_2}$). Suppose further that the probability that unit k ($k \in s^*_{2h}$) is included in r_h is Θ_{hs_2} and the inclusions are independent. Given that a fixed set $\underline{m} = \{m_1, m_2,\ldots, m_h,\ldots,m_{H_{s_2}}\}$ of counts is realized, and that s_1 and s_2 are also fixed, r_h is a simple random sample of m_h units from n^*_{2h} units. We then have, for units in s^*_{2h}:

$$\pi_{k|s_1,s_2\underline{m}} = P\left(k \in r|s_1,s_2\underline{m}\right) = m_h / n^*_{n_{2h}} = f_{2h} \tag{3.2a}$$

and

$$\pi_{kL|s_1,s_2\underline{m}} = P\left(k,\ L \in r|s_1,s_2\underline{m}\right) = f_{2h}\ \left(m_h - 1\right)\Big/\left(n^*_{2h} - 1\right)\ ;\ k \neq L \tag{3.2b}$$

if k & L belong to the same stratum s^*_{2h} but equals $f_{2h'}$, $f_{2h'}$, if they belong to two different strata h and h'. (This implies that response is independent of the variable of interest y). **Also**

$$\pi_{kk|s_1,s_2\underline{m}} = \pi_{k|s_1,s_2\underline{m}} \tag{3.2c}$$

In what follows we shall repeatedly use a general result given by El-Beshir [4] for estimation of the total in three phases sampling with arbitrary design in all three phases. It states that for such a design the estimator.

$$\hat{t}_{\pi^{**}} = \sum_r y'''_k \tag{3.3}$$

is unbiased for t with variance:

$$V\left(\hat{t}_{\pi^{**}}\right) = \sum_u\sum \Delta_{akL}\ y'_k\, y'_L + E_a\left[\sum_{s_1}\sum \Delta_{kL|s_1}\, y''_k\, y''_L\right] + E_a E\left[\sum_{s_2}\sum \Delta_{kL|s_1s_2}\, y'''_k\, y'''_L\ |S_1\right] \tag{3.4}$$

The Extended Response Homogeneity Group Model

To estimate the population total in double sampling for stratification in the presence of nonresponse we adopt the approach of looking at the response set r say, as a (third-phase) subsample that is generated from S_2 according to a certain response model. This approach is adopted, within the framework of single-phase sampling, by a number of authors including Oh and Scheuren [3] and Särndal and Swensson [2]. El-Beshir [4] followed this approach in the treatment of nonresponse in double sampling for regression. The response model we shall assume is 'the response homogeneity groups' (RHG's) model of Särndal and Swensson [2].

To formulate the model in the present context, we assume that a first-phase sample s_1 is selected from U and a second-phase sample s_2 is selected from s_1 as in § 2. The second-phase sample s_2 is partitioned into H_{s_2} disjoint groups (strata) the hth, s^*_{2h} say, of size n^*_{2h} in such a way that units within each s^*_{2h} respond independently with the same response probability Θ_{hs_2}. Thus if we denote by r_h the response set in s^*_{2h} we have for k, L ε s^*_{2h} $(h = 1,\ldots,H_{s_2})$:

$$\begin{aligned} P(k \varepsilon r|s_1 s_2) &= P(k \varepsilon r_h|s_1 s_2) = \pi_{k|s_1 s_2} = \Theta_{hs_2} \\ P(k, L \varepsilon r|s_1 s_2) &= P(k, L \varepsilon r_h|s_1 s_2) = \pi_{kL|s_1 s_2} = \pi_{k|s_1 s_2}\pi_{L|s_1 s_2} \end{aligned} \tag{3.1}$$

where r is the union of the r_h 's and the suffix in Θ is meant to indicate that the response probability may differ from one group to the other and from one s_2 to the other. The model specified by (3.1) is the RHG model extended to include the case of double sampling. Because of this we may refer to it as the 'extended response homogeneity groups model' (ERHG).

Assuming the ERHG model holds, the situation is identical to a three-phase sampling set-up with arbitrary design at phase one, stratified random sampling at phase two and a stratified Bernoulli sampling at phase three. The groups s^*_{2h} in the ERHG model correspond to strata in a stratified Bernoulli sampling, and the response probability Θ_{hs_2} to Bernoulli probability. Note that the stratification principle used in phase two need not be the same as that used in phase three. In fact, in phase two we may use such factors as age, education,...etc. to demarcate strata, whereas for the nonresponse problem the factor of classification in phase three is always the response probability.

Our plan for the rest of the paper will therefore be to develop a theory of estimation for a three-phase sampling problem with stratified random sampling at phase

form strata from which a second smaller sample is selected by stratified random sampling. However, unlike what is usually encountered in sampling texts, we relax the (usually implicit) assumption that the second sample is complete.

Description of the Problem

We assume we have the population $U = \{1,2,...k,...N\}$. A first-phase sample s_1, of size n_1, is to be selected from U according to the probability distribution $\{P_1(s_1), s_1 \subset U\}$ with inclusion probabilities for unit k ($k \varepsilon U$) and units k and L ($k,L \varepsilon U$) given respectively by π_{ak} and π_{akL}. We define $\Delta_{akL} = \pi_{akL} - \pi_{ak}\,\pi_{aL}$. Data provided by s_1 is used to partition s_1 into H_{s_1} strata, the vth, s_{1v} say, is of size n_{1v}. A subsample s_{2v} of size n_{2v} is then selected from s_{1v} ($v = 1,...,H_{s_1}$) by simple random sampling. The union of the s_{2v} is s_2, the second-phase sample, and the union of the n_{2v} is n_2 its size. The n_{2v} are assumed to be specified by the sampler for the given set of strata. Furthermore, the same set of strata and the same sampling fractions within strata, are assumed used everytime the same first-phase sample s_1 is drawn, but that this need not be so for two different samples.

For all $k,L \varepsilon s_{1v}$ the second-phase (conditional) inclusion probability for unit k is

$$\pi_{k|s_1} = n_{2v}/n_{1v} = f_{1v} \tag{2.1a}$$

and for units k and L ($k \neq L$)

$$\pi_{kL|s_1} = f_{1v}\,(n_{2v} - 1)/(n_{1v} - 1) \tag{2.1b}$$

if k & L belong to the same stratum s_{1v}, but equals $f_{1v'}$ $f_{1v'}$, if they belong to two different strata v and v'. Also $\pi_{kk|s_1} = \pi_{k|s_1}$.

Define

$$\Delta_{kL|s_1} = \pi_{kL|s_1} - \pi_{k|s_1}\,\pi_{L|s_1} \tag{2.1c}$$

The value of a characteristic y, the variate of the study, is to be observed on units in s_2 the 'intended sample', and the resulting observations used to estimate the population total $t = \sum_u y_k$, where, with y_k the value of y for unit k, we used $\sum_A y_k$ to denote $\sum_{k \varepsilon A} y_k$ where A is a set such that $A \subseteq U$. Our problem is to estimate t assuming that some of the units in s_2 failed to respond.

J. King Saud Univ., Vol. 7, *Admin. Sci.* (2), pp. 129-145 (1415 A.H./1995)

The Treatment of Nonresponse in Double Sampling for Stratification

Z.A. El-Beshir
Associate Professor, Department of Quantitative Methods, College of Administrative Sciences, King Saud University, Riyadh, Saudi Arabia

(Received on 26/11/1413; accepted for publication on 3/5/1414 A.H.)

Abstract. This paper discusses estimation of the population total in double sampling for stratification in the presence of nonresponse. Formulae are provided for the variance and estimator of variance. These expressions are shown to reduce to familiar forms in double sampling when response is complete.

Introduction

In double (or two-phase) sampling a first large sample, s_1 say, is drawn. Information provided by this sample is then utilized in a ratio or a regression estimate or for stratification purposes. When the objective is a ratio or a regression estimate, s_1 is used to get measurement on an auxiliary variable x. In case of stratification, information supplied by s_1 is utilized in different ways. Neyman [1], who was the first to provide a theory for double sampling for stratification, considered the situation in which a predetermined set of strata exists in the population, and the reason behind selecting s_1 is to estimate the strata weights. In another application, (Särndal & Swensson [2]) no subdivision of the population into strata is assumed made prior to the taking of s_1 and information provided by this sample is used in dividing it into strata. In all cases, a second smaller sample s_2 is selected from s_1 and observations are made on a characteristic y for all units in s_2. This information is then used to estimate a population parameter such as the population mean or total through a ratio, regression or stratified random sampling estimator.

In this paper we consider the second application of double sampling for stratification. In other words, we consider the situation in which the first sample is used to