Journal of Official Statistics
Vol. 5. No. 2, 1989, pp. 183–186

Miscellanea

Under the heading Miscellanea, essays will be published dealing with topics considered to be of general interest to the readers. All contributions will be refereed for their compatibility with this criterion.

The Variance of Direct Expansions From a Common Area Sampling Design

Phillip S. Kott[1]

Abstract: Many countries use a sample design in their agricultural surveys that is assumed, incorrectly, to be equivalent to stratified simple random sampling without replacement. This article discusses how the variance of a direct expansion based on that design should be estimated.

Key words: Area segment; cluster; probability proportional to size; model-based analysis; anticipated variance.

1. Introduction

Stratified simple random sampling (srs) without replacement is perhaps the most popular sampling design in the world. It allows for the simple and unbiased estimation of a population total with a variance that can be estimated in a straightforward and unbiased manner.

Survey statisticians, however, sometimes employ sampling designs that they believe to be equivalent to stratified srs without replacement but are not. This article points out a commonly used design that falls into this category. It then discusses the repercussions on variance estimation.

2. The Design

Many countries employ (at least in part) the following sampling design for their agricultural surveys. First, the land mass of a region of interest is stratified by putative land-use criteria. Then, a *with* replacement probability proportional to size (pps) sample of somewhat homogeneous primary sampling units or clusters is selected within each stratum; the measure of size being the number of equally-sized area segments in the cluster. Within clusters chosen m times for the sample, an srs of m segments is chosen without replacement (m is most often one, but it can be greater). Finally,

[1] Senior Mathematical Statistician, Division of Research and Applications, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250, U.S.A.

Acknowledgement: The author would like to thank Charles Perry, Ron Fecso, and the referees for their helpful comments.

there is complete enumeration of the farm values of interest within the sampled area segments. For a fuller description of this sampling design see Cotter and Nealon (1987).

This design, which is used by the United States, all of Central America, Pakistan, Morocco, Tunisia, Haiti[2], and a host of other countries, is generally believed to be effectively equivalent to srs without replacement but less expensive to conduct, since only area segments within sampled clusters need be delineated. It may be relatively inexpensive, but it is not equivalent to srs without replacement.

3. A Conventional Analysis

Let N be the population size of the stratum under study and n the sample size. Let the subscript j denote a cluster ($j = 1, 2, \ldots, L$) and the subscript ji ($i = 1, 2, \ldots, N_j$) a segment in cluster j. Finally, let n_{ji} be the number of times segment ji is in the sample (which can only be 0 or 1) and $n_j = \Sigma_i^{N_j}(n_{ji})$ be the number of sampled segments from cluster j. (For simplicity, we assume that all N_j exceed n.)

Since clusters are selected via pps with replacement sampling, $\mathrm{E}(n_j) = nN_j/N$. Each segment in j has the same probability of selection, call it p_{ji}. As a result, $p_{ji} = \mathrm{E}(n_{ji}) = \mathrm{E}(\Sigma_i[n_{ji}])/N_j = \mathrm{E}(n_j)/N_j = n/N$.

The probability of jointly choosing segments ji and gh for the sample is

$$p_{jigh} = \begin{cases} n/N & \text{if } ji = gh \\ \dfrac{nN_j(n-1)}{N^2(N_j-1)} & \text{if } j = g, \quad h \neq i \\ \dfrac{n(n-1)}{N^2} & \text{if } j \neq g. \end{cases} \tag{1}$$

[2] I would like to thank Roger Latham formerly of the National Agricultural Statistics Service's International Programs Office for providing me this abbreviated list of countries.

By contrast, for srs without replacement, p_{jigh} again equals n/N when $ji = gh$, but it equals $n(n-1)/[N(N-1)]$ otherwise.

Let y_{ji} be the farm value of interest for segment ji. The quantity $Y = \Sigma_j^L \Sigma_i^{N_j}(y_{ji})$ can be estimated in an unbiased fashion by the direct expansion $y = (N/n)\Sigma_j\Sigma_i(n_{ji}y_{ji})$, since $\mathrm{E}(n_{ji}) = n/N$.

Although y has the same form as a direct expansion estimator based on either srs *without* replacement or srs *with* replacement, its variance is another matter.

It simplifies the analysis to focus on an unbiased estimator for the true variance of y rather than on the variance itself. The Yates-Grundy estimator for the variance of y is (Cochran (1977, p. 261. eq. 9A.44)):

$$v_{YG} = \frac{1}{2}\sum_{g=1}^{L}\sum_{h=1}^{N_g}\sum_{j=1}^{L}\sum_{i=1}^{N_j}\left(\frac{p_{gh}p_{ji}}{p_{ghji}} - 1\right) \times n_{ji}n_{gh}\left(\frac{y_{ji}}{p_{ji}} - \frac{y_{gh}}{p_{gh}}\right)^2.$$

After some manipulation, we see that

$$v_{YG} = \frac{N^2}{n(n-1)}\left\{\sum_j\sum_i n_{ji}(y_{ji} - \bar{y})^2 - \sum_j (n_j/N_j)\sum_i n_{ji}(y_{ji} - \bar{y}_j)^2\right\}, \tag{2}$$

where $\bar{y} = y/N$, and $\bar{y}_j = \Sigma_i(n_{ji}y_{ji})/n_j$ when $n_j > 0$ and $\bar{y}$ otherwise.

There are two things immediately clear from equation (2). The Yates–Grundy variance estimator for y given the sampling design under study is not the same as the variance estimator for srs *without* replacement except in the degenerate case when $L = 1$ (making $n_j/N_j = n/N$ and $\bar{y}_j = \bar{y}$). On the other hand, v_{YG} is exactly equal to the standard variance estimator for srs *with* replacement (the first line of the right hand side of (2)) when no n_j exceeds unity;

i.e., when no cluster has more than one segment in the sample. Otherwise, the srs with replacement variance formula will be greater than v_{YG} when the sampled y_{ji} within a cluster with $n_j > 1$ are not all equal.

Since the Yates–Grundy estimator in (2) is an unbiased estimator of the variance of y, the standard srs with replacement variance estimator has (if anything) a slight upward bias as an estimator for the variance of y. Unfortunately, no similar statement can be made about the standard estimator for the variance of y under srs *without* replacement.

4. Model-Based Analysis

In order to relate the variance of y to the srs without replacement variance estimator,

$$v_{wtr} = \frac{N^2(1 - n/N)}{n(n-1)} \left\{ \sum_j \sum_i n_{ji}(y_{ji} - \bar{y})^2 \right\},$$

it is helpful to treat the y_{ji} as random variables. In particular, consider the stochastic structure of Scott and Smith (1969) which assumes

$$y_{ji} = \mu + \delta_j + \tau_{ji},$$

where δ_j and τ_{ji} are uncorrelated random variables with means of zero and variances of σ_B^2 and σ_W^2, respectively.

Let ε denote expectation with respect to the model. It is not difficult to show that $\varepsilon(v_{wtr})$ is always less than or equal to $\varepsilon\{E[(y - Y)^2]\}$, an expression Isaki and Fuller (1982) called the anticipated variance of y. Strict equality holds if and only if $\sigma_B^2 = 0$. Thus, treating v_{wtr} as the variance estimator for y is justified by the model when $\sigma_B^2 = 0$ but not when $\sigma_B^2 > 0$.

5. Discussion

The analysis presented here has shown that a very common sampling design for agricultural surveys is not – as is widely believed – equivalent to stratified simple random sampling without replacement. Nevertheless, using the standard stratified srs without replacement variance estimation formula for direct expansions can sometimes, but not always, be justified on model-based grounds. The use is justified when the farm values for segments in a particular stratum can be treated as uncorrelated random variables with a common mean and variance. The use is not justified, however, when the farm values for segments within the same cluster are correlated (i.e., when $\sigma_B^2 > 0$).

The Yates–Grundy variance estimator developed here is unbiased in the conventional sense but it is cumbersome to use and can be unstable. It may therefore be advisable to estimate variances of direct expansions using the somewhat conservative srs *with* replacement formula. This is especially appealing when the stratum sampling fractions are small (in U.S. agriculture surveys they are almost always less than 10%) so that the difference between the upwardly biased with replacement and downwardly biased (under the model) without replacement variance estimation formulae is also small.

One last point warrants mentioning. It is possible to modify the sampling design under discussion so that it is effectively equivalent to stratified srs without replacement. After a cluster is selected for the sample, its measure of size can be decreased by one before the next cluster (from the stratum) is chosen via pps sampling. Segments can be subsampled from singly or multiply selected clusters as they are now. The individual and joint segment selection probabilities could then be shown to be identical to the selection probabilities had a without replacement stratified simple random sample of segments been chosen.

6. References

Cochran, W.G. (1977): Sampling Techniques (3rd edition). New York: Wiley.

Cotter, J. and Nealon, J. (1987): Area Frame Design for Agricultural Surveys. U.S. Department of Agriculture, Washington, D.C.

Isaki, C.T. and Fuller, W.A. (1982): Survey Design Under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, pp. 89–96.

Scott, A. and Smith. T.M.F. (1969): Estimation in Multi-stage Surveys. Journal of the American Statistical Association, 64, pp. 830–840.

Received June 1988
Revised January 1989