DISCUSSION

Richard Bolstein, George Mason University
6118 Mountain Springs Lane, Clifton, VA 22024

Key Words: probability proportional to size

Probability proportional to size sampling is used to improve accuracy when the study variables of interest are believed to be correlated with a measure of size (MOS) of the sampling units. This occurs often in cluster sampling with the MOS usually chosen as the number of second stage units in the cluster. Sampling can be done with replacement (denoted pps) or without replacement (πps) but is technically simpler in the *with replacement* case. This session contained three papers which applied standard methodology and one technical paper in which a new estimator was proposed with pps sampling. The latter paper by Warde and Belgacem will be discussed first. We begin with a brief review of unequal probability sampling with replacement.

1. Background

Let N denote the size of a finite population. It is desired to estimate the total $Y = \sum_{i=1}^{N} y_i$ of a variable y which has value y_i at population unit $i = 1,\cdots N$. Let $z_i > 0$ be a known positive number assigned to unit $i = 1,\cdots,N$ such that $\sum_{i=1}^{N} z_i = 1$. If a sample of size n is chosen with replacement, such that unit i has probability z_i of selection on each draw (called ppz sampling), let s(i) denote the unit selected on the i^{th} draw. Then it is well-known (Cochran, p. 293) that

$$(1) \qquad \hat{Y}_{ppz} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_{s(i)}/z_{s(i)}$$

is an unbiased estimator of Y with variance

$$(2) \qquad V[\hat{Y}_{ppz}] = \frac{1}{n} \cdot \sum_{i=1}^{N} z_i \cdot (y_i/z_i - Y)^2$$
$$= \frac{1}{n} \cdot \sum_{i=1}^{N} \sum_{j>i} z_i \cdot z_j \cdot (y_i/z_i - y_j/z_j)^2$$

In pps sampling, the z_i are given by

$$(3) \qquad z_i = X_i/X_0$$

where X_i is a measure of size of unit i and

$$X_0 = \sum_{i=1}^{N} X_i.$$

Now let $z_i^* > 0$ be fixed but unknown numbers associated with each population unit i. Assume these numbers can be measured without error for each unit in the sample, and define

$$(4) \qquad \hat{Y}_{ppnz} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_{s(i)}/z^*_{s(i)}$$

In ppz sampling,

$$E[y_{s(j)}/z^*_{s(j)}] = \sum_{i=1}^{N} z_i \cdot (y_i/z_i^*)$$

for each $j = 1,\cdots n$. Therefore

$$(5) \qquad E[\hat{Y}_{ppnz}] = \sum_{i=1}^{N} y_i \cdot (z_i/z_i^*)$$

so the bias of $\hat{Y}_{ppnz}$ as an estimator of Y is

$$(6) \qquad B = B[\hat{Y}_{ppnz}] = \sum_{i=1}^{N} y_i \cdot (z_i/z_i^* - 1)$$

It is straight-forward to show that

$$(7)\quad V[\hat{Y}_{ppnz}] = \sum_{i=1}^{N} z_i \cdot \left(\frac{y_i}{z_i^*} - Y - B\right)^2$$
$$= \frac{1}{n} \cdot \sum_{i=1}^{N} \sum_{j>i} z_i \cdot z_j \cdot (y_i/z_i^* - y_j/z_j^*)^2$$

2. The Warde-Belgacem Paper.

This paper studies the use of the estimator $\hat{Y}_{ppnz}$ in pps sampling where the z_i are given by (3) and the z_i^* by

$$(8)\quad z_i^* = X_i^*/X_0$$

where X_i^* is the revised measure of size of unit i and is known only for units in the sample. Note that $\sum_{i=1}^{N} z_i^*$ is not necessarily equal to 1 because $X_0^* = \sum_{i=1}^{N} X_i^*$ is unknown and may not equal X_0. However, formulas (5)-(7) only require that $z_i^* > 0$ for all $i = 1,\cdots N$.

The authors have an interesting idea here but they did not develop the theory. They state the second equation for the variance in (7) but do not give a proof or reference, and they were apparently unaware of the bias formula (6). Moreover, in their examples they allow X_i^*, and hence z_i^*, to have the value zero. This is important since it represents the very real situation that the sampling unit may no longer belong in the sample frame: for example, X_i may be the number of employees in establishment i from a listing several months old, but the company has since gone out of business. But formulas (4)-(7) are no longer correct in this situation. The authors claim that since $X_i^* = 0$ implies $y_i = 0$, the ratio $y_{s(i)}/z_{s(i)}^*$ in (4) can be set equal to 1, but this would result in (4) having a substantial positive bias. The natural way to proceed is to discard unit i from the sample if $X_i^* = 0$ (since this means it is ineligible) and continue sampling until n units are obtained with $X_i^* > 0$ for each i. Formula (4) then makes sense. Partition the set of units $\{1,\cdots N\}$ into disjoint and exhaustive sets $\mathcal{I}$ (for in-frame) and $\mathcal{O}$ (for out-of-frame) on which $X_i^* > 0$ and $X_i^* = 0$ respectively. Then it is easy to show that (5)-(7) should be modified by replacing z_i by $z_i/[1 - \sum_{k \in \mathcal{O}} z_k]$ and restricting the summations to the index set $\mathcal{I}$.

The authors simulate the MOS X_i for each of ten populations of size $N = 100$ from a beta distribution with parameters 1.5 and 6, which has a mean of 0.2 and standard deviation of 0.13 and define the z_i by (3). Then they first model the

$$(9)\quad z_i^* = z_i + e_{1i}$$

where the e_{1i} are independent, normally distributed with mean 0 and standard deviation .001. So it may happen that some $z_i^* < 0$, which they then reset to 0. (While this may not seem to be the best way to model the MOS change, there are other points to address here.) The study variable is modeled in turn by

$$(10)\quad y_i = \beta z_i^* + e_i$$

where the e_i are independent and normally distributed with mean 0 and variance σ^2. It is also assumed that the e_i are independent of the e_{1i}. Again we must assume that y_i is not included in the total Y if $z_i^* = 0$ since unit i should be excluded from the population in this case. It does not appear that it is excluded in the paper.

To compare the accuracy of (4) with (1), from each of the ten populations the authors generated 100 pps samples of size 10 each and observed that the absolute error (which they inappropriately call bias) of estimating Y by (4) is less than the error of estimating Y by (1) in 72% of the samples across populations. They also estimate the variances of (1) and (4) for each of the ten populations from the 100 samples, and conclude that the variance of the new estimator is less in nine populations. We wish to make two points. First, **computation of the mean square errors can be calculated directly from (2), (6), and (7), which renders sampling unnecessary.** Second, **the results of sampling are predictable anyway because they are forced by the model (9) and (10).**

To understand this second point, we assume for simplicity that $z_i^* > 0$ for all $i = 1,\cdots 100$. Let $\mathbf{e}$ and $\mathbf{e_1}$ denote the vector of the (e_i) and (e_{1i}) respectively. By (6), (9), and (10), the bias of $\hat{Y}_{ppnz}$ given $\mathbf{e}$ and $\mathbf{e_1}$ is

$$B[\hat{Y}_{ppnz} \mid \mathbf{e}, \mathbf{e_1}] = \beta \sum_{i=1}^{N} e_{1i} + \sum_{i=1}^{N} (e_{1i}/z_i^*) \cdot e_i$$

On taking the expectation first with respect to $\mathbf{e}$ and then with respect to $\mathbf{e_1}$ we see that the new estimator $\hat{Y}_{ppnz}$ is model-unbiased. Substitution of (10) into (7) yields

(11) $V[\hat{Y}_{ppnz} \mid \mathbf{e}, \mathbf{e_1}] =$

$$\frac{1}{n} \cdot \sum_{i=1}^{N} \sum_{j>i} z_i z_j (e_i/z_i^* - e_j/z_j^*)^2$$

$$= \frac{1}{n} \sum_{i=1}^{N} z_i (1 - z_i) e_i^2/(z_i^*)^2 - \frac{2}{n} \sum_{i=1}^{N} \sum_{j>i} \frac{z_i z_j e_i e_j}{z_i^* z_j^*}$$

Taking the expectation with respect to **e**:

(12) $$\mathcal{E}V[\hat{Y}_{ppnz} \mid \mathbf{e_1}] = \frac{\sigma^2}{n} \sum_{i=1}^{N} z_i (1 - z_i)/(z_i^*)^2$$

We want to compare this to $V[\hat{Y}_{ppz} \mid \mathbf{e_1}]$. By (9) and (10),

(13) $$y_i = \beta z_i + \beta e_{1i} + e_i = \beta z_i + e_{2i}$$

where the e_{2i} are independent and normally distributed with mean 0 and variance equal to $\sigma^2 + \beta^2/10^6$. Substitution into (2) gives a formula analogous to (11):

$V[\hat{Y}_{ppz} \mid \mathbf{e}, \mathbf{e_1}] =$

$$\frac{1}{n} \cdot \sum_{i=1}^{N} \sum_{j>i} z_i z_j (e_{2i}/z_i - e_{2j}/z_j)^2$$

$$= \frac{1}{n} \sum_{i=1}^{N} (1 - z_i) e_{2i}^2/z_i - \frac{2}{n} \sum_{i=1}^{N} \sum_{j>i} e_{2i} e_{2j}.$$

Taking the expectation with respect to **e** and then $\mathbf{e_1}$ gives

(14) $$\mathcal{E}_1 \mathcal{E}V[\hat{Y}_{ppz}] = \frac{\sigma^2 + \beta^2/10^6}{n} \sum_{i=1}^{N} (1 - z_i)/z_i$$

Since $\beta = 1{,}395{,}000$, comparison of (12) and (14) shows that the model (9)-(10) forces $\hat{Y}_{ppnz}$ to have a substantially smaller variance and mean square error on average than $\hat{Y}_{ppz}$. Consequently, the simulation results are predictable.

3. Applications Papers.

The Vasectomy Survey paper presented by Gargiullo is a textbook example of stratified pps sampling. In this application, the units of analysis are *physician practices*, which are clusters of physicians. The practices are stratified into three groups: urologists, general surgeons, and family practitioners. The sample frame is the AMA Physician master file, and stratified simple random sampling of physicians with replacement from this frame constitutes pps sampling of practices within strata where the measure of size of a practice is the number of physicians in it. One fact that causes some loss in precision is that in non-urology strata there are many practices which perform no vasectomies. However, pps sampling is still preferable to simple random sampling with ratio or regression estimators in this situation. One interesting connection with the Warde-Belgacem paper is this: If a physician is selected and is no longer affiliated with the practice listed in the frame, then if that practice is kept in the sample, the measures of size z_i are out of date and the estimator $\hat{Y}_{ppnz}$ could possibly be used to advantage.

Finally, the Rental Assistance Subsidies paper by Errecart et al is an interesting example of a stratified two-stage systematic selection process proportional to size. Here there are geographic strata and the PSU's are clusters of housing projects. But each cluster contains two types of projects: PHA and MF. They really want to draw two separate samples from each geographic stratum: a sample of PHA housing units and a sample of MF housing units. To do this, they selected the PSU's by using a measure of size derived from both types of projects, namely the simple average of the proportion of PHA tenants in the region. Then a systematic sample of PSU's was taken proportional to size. Now from each chosen PSU, two samples were drawn independently: one of PHA projects and one of MF projects. In other words, two distinct samples were drawn with the first stage in common. Their choice of measure of size for the first stage was clever because it lowered the probability of selection of a cluster which was dominated by one type of housing. This approach should have a wide variety of other applications, and indeed this discussant has done something similar.

Reference

Cochran, W.G. (1977) *Sampling Techniques*, Wiley