9-1962

The Variance of an Estimator with Post-Stratified Weighting.pdf

William (Bill) H. Williams
CUNY Hunter College

THE VARIANCE OF AN ESTIMATOR WITH POST-STRATIFIED WEIGHTING*

W. H. WILLIAMS
Bell Telephone Laboratories, Incorporated

A straightforward procedure is presented which uses the variance formula of a simple estimator to obtain an approximate formula for the variance of a post-stratified estimator. The approximation is the well-known one for the variance of a ratio estimator. The technique is applied to samples which have been selected in stratified and multi-stage designs and can be extended to other ratio estimators.

1. INTRODUCTION

IN MOST APPLICATIONS of stratified sampling, the stratum to which a unit belongs can be identified in advance, and a sample of fixed, predetermined size can be drawn from each of the strata. However, there are numerous situations in which the population strata sizes are known, but it is impossible to determine the exact stratum to which a unit belongs until the unit has actually been sampled. For example, the religious affiliation of a person cannot usually be ascertained until the individual has been interviewed, but the distribution of religions may well be known. Hence, although it is known that there are ${}_iM$ units in the i-th stratum, it is not known which particular units make up the ${}_iM$ and it is impossible to draw a sample of fixed size. It would be possible to continue sampling until a fixed number in each stratum is selected, and then to ignore any excess observations, but this method might often be expensive and is not considered here. A stratified estimator may still be used by placing each unit in its appropriate stratum as it is drawn. This is known as post-stratification. However, the variable sample size in each stratum means that the usual variance formulas for stratified samples may often be regarded as inappropriate.

In simple sampling the post-stratified estimator of the total and mean are known and can be found, for example, in Cochran [1] and Hansen, *et al.* [2]. These results have been obtained by first considering the sample number in each stratum as fixed and then allowing for the variation over possible stratum sample sizes by using a result of Stephan [4]. The resultant formulas, as would be expected, are approximately the same formulas as stratified sampling with proportional allocation.

A method for obtaining the variance of a post-stratified estimator in any type of sampling is derived in this paper.

2. THE GENERAL APPROACH

Consider a sample of size m that has been drawn according to any specified sampling scheme from a population of size M. Let $\hat{Y} = \hat{Y}(y)$ denote an estimator of the population total, $\mathrm{Var}(\hat{Y}) = \sigma^2_{\hat{Y}}(y)$ the sampling variance of $\hat{Y}$, and $\mathrm{Var}(\hat{Y}) = \hat{\sigma}^2_{\hat{Y}}(y)$ an estimator of $\mathrm{Var}(\hat{Y})$ where y represents the characteristic attached to the individual units. In a specific example y would have subscripts

* Research on this problem began while the author was associated with Iowa State University.

suitable for the particular design in use. The functional notations $\sigma^2_{\hat{Y}}(y)$ and $\hat{\sigma}^2_{\hat{Y}}(y)$ are used to emphasize in certain places of the following discussion that the expressions are functions of the unit characteristics. Let ${}_iM$ be the size of the i-th stratum with $\sum_{i=1}^{L} {}_iM = M$. Small case letters will be used to denote the corresponding sample sizes. It is assumed throughout that the i index of stratification has not been used in drawing the sample; this allows the allocation of units to these strata after sampling.

Now, if

$$\begin{aligned} y' &= y \text{ if the unit is in the } i\text{-th stratum,} \\ &= 0 \text{ if the unit is not in the } i\text{-th stratum,} \end{aligned} \tag{2.1}$$

then ${}_i\hat{Y} = \hat{Y}(y')$, $\sigma^2_{\hat{Y}}(y')$ and $\hat{\sigma}^2_{\hat{Y}}(y')$ are respectively, an estimator of the stratum total, the sampling variance of ${}_i\hat{Y}$ and the sample variance of ${}_i\hat{Y}$.

Furthermore, if

$$\begin{aligned} c &= 1 \text{ if the unit is in the } i\text{-th stratum,} \\ &= 0 \text{ if the unit is not in the } i\text{-th stratum,} \end{aligned} \tag{2.2}$$

then ${}_i\hat{M} = \hat{Y}(c)$ is an estimator of the size of the i-th stratum. Thus ${}_i\hat{y} = {}_i\hat{Y}/\hat{M}$ is a ratio estimator of the stratum mean. The sampling variance of this estimator is approximately

$$\text{Var } ({}_i\hat{y}) = \sigma^2_{\hat{Y}}(w')/{}_iM^2,$$

where

$$\begin{aligned} w' &= y - {}_i\overline{Y} \text{ if unit is in the } i\text{-th stratum,} \\ &= 0 \text{ if unit is not in the } i\text{-th stratum,} \end{aligned} \tag{2.3}$$

and ${}_i\overline{Y}$ is the i-th stratum population mean. The zero substitution procedure has been used before in various contexts. For example, Hartley [3] used it for the analytic study of domains.

It is important to notice that this estimator is a ratio estimator and hence is usually biased. Similarly, the variance of ${}_i\hat{y}$ is also approximate. These approximations have been discussed previously; see, for example, Cochran [1]. They are often satisfactory but care must be taken if the sample size is small.

By considering each stratum individually Eq. (2.4) can be formed by post-stratified weighting as an estimator of the mean of the entire population.

$$\hat{y}_p = \sum_{i=1}^{L} \frac{{}_iM}{M} {}_i\hat{y}. \tag{2.4}$$

Furthermore, consideration of the algebra will show (see Section 5) that $\text{Var}(\hat{y}_p) = \sigma^2_{\hat{Y}}(w)/M^2$ where $w = y - {}_i\overline{Y}$ if the unit is in the i-th stratum, $i = 1, 2, \cdots, L$. So to obtain the variance of $\hat{y}_p$ in any type of sampling insert the variate w into the formula $\sigma^2_{\hat{Y}}(y)$ of the specified design and divide by M^2. This is the main result that allows us to write down the variance formulas for the post-stratification of many types of samples.

An approximate sample estimate $\text{var}(\bar{y}_p)$ of $\text{Var}(\bar{y}_p)$ is given by $\hat{\sigma}^2_{\bar{Y}}(w'')/M^2$ where $w''=y-{}_i\bar{y}$ if the unit is in the i-th stratum, $i=1, 2, \cdots, L$.

The estimator $\bar{y}_p$ is not defined if one or more of the m_i are zero. In this case, however, the estimator can be treated by replacing the missing stratum mean by some arbitrary constant A. That is, an exact definition of the post-stratified estimator would be

$$\bar{y}_p = \sum_{i=1}^{L} \frac{{}_iM}{M} {}_i\bar{y}', \tag{2.5}$$

where

$$\begin{aligned} {}_i\bar{y}' &= {}_i\bar{y} \qquad \text{if } {}_im > 0, \\ &= A \qquad \text{if } {}_im = 0. \end{aligned}$$

Now assuming that ${}_im$ has a binomial distribution with $P_i={}_iM/M$, then

$$\text{Prob}\,({}_im = \beta) = \binom{m}{\beta} P_i^{\beta} Q_i^{m-\beta} \qquad Q_i = 1 - P_i, \tag{2.6}$$

and

$$E({}_i\bar{y}') = Q_i^m A + (1 - Q_i^m)\,{}_i\bar{Y},$$

by ignoring the technical bias of the ratio estimator. Hence

$$E(\bar{y}_p) = \bar{Y} + \sum_{i=1}^{L} \frac{{}_iM}{M} Q_i^m (A - {}_i\bar{Y}). \tag{2.7}$$

This shows that $\bar{y}_p$ is not unbiased but is consistent. The best choice of A is the best estimate of ${}_i\bar{Y}$ taken in advance of sampling and must not be altered subsequently.

3. APPLICATIONS OF THE GENERAL RESULT

Imagine the population classified in two directions and let ${}_iM_t$ be the population size in the (i, t) cell, $i=1, 2, \cdots, L$, $t=1, 2, \cdots, N$. Also

$$\sum_{i=1}^{L} {}_iM_t = M_t, \qquad \sum_{t=1}^{N} {}_iM_t = {}_iM, \qquad \sum_{i=1}^{L}\sum_{t=1}^{N} {}_iM_t = M. \tag{3.1}$$

Furthermore, let ${}_iY_t$ denote the total of the ${}_iM_t$ values in the (i, t) cell with similar notation for their summation. Bars over the letters will be used to denote means and small case letters will be used to denote corresponding sample values.

As an illustration, consider the post-stratification of a simple random sample where m units have been drawn from M with equal probability and without replacement. Then if

$$\hat{\bar{Y}} = M\bar{y}, \tag{3.2}$$

where $\bar{y}$ is the sample mean, we have

$$\text{Var}\,(\hat{\bar{Y}}) = M^2\left(1 - \frac{m}{M}\right)\frac{1}{m}\cdot\frac{1}{M-1}\sum_{i=1}^{L}\sum_{j=1}^{{}_iM} ({}_iy_j - \bar{Y})^2, \tag{3.3}$$

where $\mathrm{Var}(\hat{Y})$ is written in a form which distinguishes the stratum to which a unit belongs. Now since the ratio estimator ${}_i\hat{\mathcal{Y}}$ reduces to the simple stratum mean,

$$\hat{\mathcal{Y}}_p = \sum_{i=1}^{L} \frac{{}_iM}{M}\, {}_i\bar{y}, \tag{3.4}$$

and replacing ${}_iy_j$ in Eq. (3.3) by ${}_iy_j - {}_i\overline{Y}$ (see Section 2) we obtain

$$\mathrm{Var}\,(\hat{\mathcal{Y}}_p) \doteq \left(1 - \frac{m}{M}\right) \frac{1}{m(M-1)} \sum_{i=1}^{L} \sum_{j=1}^{{}_iM} ({}_iy_j - {}_i\overline{Y})^2. \tag{3.5}$$

Note that in this simple case the substitution of ${}_iy_j - {}_i\overline{Y}$ in the term $\overline{Y}$ causes this term to disappear entirely.

Next consider a sample that has been drawn in a stratified manner. The t index will be used to denote this direction of stratification and hence Equations (3.6) and (3.7) denote the known formulas for sampling without replacement within strata.

$$\hat{\mathcal{Y}}_{st} = \sum_{t=1}^{N} \frac{M_t}{M}\, \bar{y}_t, \tag{3.6}$$

with

$$\mathrm{Var}\,(\hat{\mathcal{Y}}_{st}) = \sum_{t=1}^{N} \frac{M_t^2}{M^2}\left(1 - \frac{m_t}{M_t}\right) \frac{1}{m_t} \frac{1}{M_t - 1} \sum_{i=1}^{L} \sum_{j=1}^{{}_iM_t} ({}_iy_{tj} - \overline{Y}_t)^2. \tag{3.7}$$

Now regarding the i index as a second direction of stratification we obtain

$$\hat{\mathcal{Y}}_p = \sum_{i=1}^{L} \frac{{}_iM}{M}\, {}_i\hat{\mathcal{Y}}, \tag{3.8}$$

where ${}_i\hat{\mathcal{Y}} = {}_i\hat{Y}/{}_i\hat{M}$ with ${}_i\hat{Y}$ and ${}_i\hat{M}$ obtained as in Section 2.

Furthermore, substituting in Equation (3.7) for the ${}_iy_{tj}$'s (including those contained implicitly in $\overline{Y}_t$)

$$\mathrm{Var}\,(\hat{\mathcal{Y}}_p) \doteq \sum_{t=1}^{N} \frac{M_t^2}{M^2}\left(1 - \frac{m_t}{M_t}\right) \frac{1}{m_t} \frac{1}{M_t - 1} \cdot \sum_{i=1}^{L} \sum_{j=1}^{{}_iM_t} ({}_iy_{tj} - \overline{Y}_t - {}_i\overline{Y} + \overline{\overline{Y}}_t)^2, \tag{3.9}$$

where

$$\overline{\overline{Y}}_t = \sum_{i=1}^{L} \frac{{}_iM_t}{M_t}\, {}_i\overline{Y}.$$

A sample estimate $\mathrm{var}(\hat{\mathcal{Y}}_p)$ of (3.9) is given by the same method and has a similar form.

Next consider the multistage sampling scheme as follows: n primaries are selected from N with probability proportional to size $p_i = M_i/M$ and with replacement; then from each of the sampled primaries a fixed take of k secondaries is chosen with equal probability and without replacement. Now redefining the y variate in the usual manner in the known formulas for this sampling structure gives

$$\bar{y}_p = \sum_{i=1}^{L} \frac{{}_iM}{M} {}_i\bar{y}, \tag{3.10}$$

with

$$\text{Var}\,(\bar{y}_p) = \frac{1}{Mn} \sum_{i=1}^{N} \frac{M_i}{M} \left(1 - \frac{k}{M_i}\right) \frac{1}{M_i - 1} \sum_{i=1}^{L} \sum_{j=1}^{{}_iM_i} ({}_iy_{ij} - {}_i\overline{Y} - \overline{Y}_i + \overline{\overline{Y}}_i)^2$$
$$+ \frac{1}{n} \sum_{i=1}^{N} \frac{M_i}{M} (\overline{Y}_i - \overline{\overline{Y}}_i)^2. \tag{3.11}$$

Other multistage designs can be handled in the same manner.

4. CONCLUDING REMARKS

The procedure presented here can be used to post-stratify samples from any type design. It depends upon the fact that

$$\text{Var}\left(\frac{\bar{y}}{\bar{x}} X\right) \doteq \text{Var}\left(\bar{y} - \frac{Y}{X} \bar{x}\right). \tag{4.1}$$

If $\bar{y}$ (and $\bar{x}$) are linear in the observations and the sampling variance of $\bar{y}$ is a quadratic form in y_α, $\alpha = 1, 2, \cdots, N$, denoted Q; then the variance of $\bar{y} - R\bar{x}$ is the same quadratic form Q in $y_\alpha - R\bar{x}_\alpha$, $\alpha = 1, 2, \cdots, N$. Discussion of this variance derivation and approximation can be found in Cochran [1] and Hansen, *et al.* [2]. The use of the result in this paper where the concomitant variate is a count variate is a special case of the general form, Equation (4.1).

The post-stratified estimators are ratio estimators and as such they should be used with caution if small sample sizes appear in denominators. For large samples the technique yields results equivalent to those that could have been obtained by stratified sampling. For illustration, Table 6.1 of Cochran [1] contains data of a random sample of 49 of 196 United States cities. Both the 1920 and 1930 populations of the cities are given and estimation of the total 1930 city population is being considered. By grouping the sample on the basis of 1930 city size into the four strata, 0–60, 61–100, 101–200, 201+(population in thousands), it will be found that the proportional reduction in variance by post-stratification is 81 per cent. A ratio estimator utilizing the detailed 1920 city size yields a proportional reduction in variance of 96 per cent over simple random sampling.

As a second example, Hansen, *et al.* [2] Volume I, list in Table A-8 the number of stores per block for 87 randomly selected blocks in Buffalo, New York. If the sample is grouped into three strata, 0-9, 10-19, over 20 stores per block, the proportional reduction in variance is estimated to be 66 per cent.

The procedure is the same in any type of sampling. The over-all gain is, of course, largely dependent upon the dispersion of the stratum means, and differences in these means will usually result in gains. The procedure is easily generalized to utilize a concomitant variate.

REFERENCES

[1] Cochran, W. G., *Sampling Techniques*. New York: John Wiley and Sons, Inc., 1953.
[2] Hansen, M. H., Hurwitz, W. N., and Madow, W. G., *Sample Survey Methods and Theory, Volume II*. New York: John Wiley and Sons, Inc., 1953.
[3] Hartley, H. O., "Analytic Studies of Survey Data," *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, D. C., 1959. Pp. 146–54.
[4] Stephan, F. F., "The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate," *Annals of Mathematical Statistics*, 16 (1945), 50–61.