Simple derivations of variance formulas in two-stage simple random sampling

Paul Knottnerus

Discussion paper (201128)

Statistics Netherlands

The Hague/Heerlen, 2011

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201128 X-10

Simple derivations of variance formulas in two-stage simple random sampling

Paul Knottnerus

Summary: This paper gives alternative derivations for the standard variance formulas in two-stage sampling. The derivations are based on a direct use of the statistical properties of the sampling errors in the second stage. For the ease of exposition we examine the specific case that simple random sampling is used in both stages. These derivations might be useful for readers looking for more elementary approaches to two-stage sampling.

Keywords: clusters, finite populations, sampling errors, subsamples.

1. Introduction

Proofs of variance formulas in two-stage sampling often require some algebraic skills. Also for the situation where a simple random sample without replacement (SRS) is drawn in the first stage, proofs might be rather intricate. This might be the case especially when proofs for the SRS situation are based on the more general situation with unequal probabilities. For an overview of variance estimation methods in multistage sampling, see Chaudhuri and Stenger (2005), and the references given therein.

For the specific case of two-stage sampling with SRS samples in both stages (TSS) this paper presents alternative derivations for the variance formulas without using heavy algebra. The derivations in this paper are a further elaboration of those given by Knottnerus (2003, pages 151-152) for two-stage sampling with unequal probabilities. However, the *sampling autocorrelation coefficient* used there is not really necessary for the TSS case.

2. Alternative derivations of TSS variance formulas

Consider a population $U = \{1,...,N\}$ consisting of N clusters. Let M_i denote the size of cluster i, say U_i (i=1,…,N). Furthermore, let Y_{ij} denote the value of study variable y for the jth unit in U_i $(j = 1,...,M_i)$ and let $Y_i = \Sigma_{j \in U_i} Y_{ij}$ denote the corresponding cluster total of U_i. The cluster mean Y_i / M_i is denoted by $\overline{Y}_i$ and $Y = \Sigma_{i \in U} Y_i$ is the population total. In the first stage of the TSS design an SRS

sample *s* of size *n* is drawn from the *N* clusters in *U*. Recall that an SRS sample of *n* clusters can simply be obtained by taking the first *n* clusters of *U* after the *N* clusters are put in a random order. In the second stage an SRS subsample is drawn from each of the *n* clusters selected in *s*; it is assumed that these *n* subsamples are drawn independently. When U_i is selected in *s*, denote the corresponding SRS subsample from U_i by s_i and its size by m_i. The SRS estimator $\hat{Y}_i$ of Y_i is given by

$$\hat{Y}_i = \frac{M_i}{m_i} \sum_{j \in s_i} Y_{ij}.$$

The well-known formula for the variance of $\hat{Y}_i$, say s_i^2, is

$$s_i^2 \circ \mathrm{var}(\hat{Y}_i) = M_i^2 (1 - f_i) \frac{S_i^2}{m_i},$$

where $f_i = m_i / M_i$, and

$$S_i^2 = \frac{1}{M_i - 1} \sum_{j \in U_i} (Y_{ij} - \bar{Y}_i)^2.$$

Let $i_1, \ldots, i_n$ denote the rank numbers of the *n* clusters selected in *s*. Henceforth, the quantities $\hat{Y}_{i_k}$ and Y_{i_k} of the *n* observed clusters are briefly denoted by the lower case letters $\hat{y}_k$ and y_k, respectively $(k = 1, \ldots, n)$. Define the corresponding sampling errors d_k by $d_k = \hat{y}_k - y_k$. Denote the sample means of the $\hat{y}_k, y_k$ and d_k by $\bar{\hat{y}}_s, \bar{y}_s$ and $\bar{d}_s$, respectively. In addition, denote the sample variance of the $\hat{y}_k$ by $s_{\hat{y}}^2$. That is,

$$s_{\hat{y}}^2 = \frac{1}{n - 1} \sum_{k=1}^{n} (\hat{y}_k - \bar{\hat{y}}_s)^2 .$$

Similarly, denote the sample variances of the y_k and d_k by s_y^2 and s_d^2, and their sample covariance by s_{yd}. For the *n* subsamples lower case letters are used as well. So the subsample means $\bar{y}_{sk}$ and the subsample variances s_k^2 are written as

$$\bar{y}_{sk} = \frac{1}{m_k} \sum_{j=1}^{m_k} y_{kj},$$

$$s_k^2 = \frac{1}{m_k - 1} \sum_{j=1}^{m_k} (y_{kj} - \bar{y}_{sk})^2,$$

respectively $(k = 1, \ldots, n)$.

The classical unbiased TSS estimator $\hat{Y}_{TSS}$ for Y can now be written as

$$\hat{Y}_{TSS} = N\bar{\hat{y}}_s = \frac{N}{n}\sum_{k=1}^{n}\hat{y}_k .$$

Its variance is

$$\mathrm{var}(\hat{Y}_{TSS}) = N^2\{(\frac{1}{n} - \frac{1}{N})S_y^2 + \frac{\sigma_d^2}{n}\}, \qquad (1)$$

where

$$S_y^2 = \frac{1}{N-1}\sum_{i\in U}(Y_i - \frac{Y}{N})^2,$$
$$\sigma_d^2 = \frac{1}{N}\sum_{i\in U}\sigma_i^2.$$

Proof of (1). Recall that i_k (k =1,…,n) can be regarded as a random variable with $P(i_k = i) = 1/N$ (i =1,...,N). Then using $E(d_k) = 0$, it is seen that for k =1,…,n

$$\begin{aligned}\mathrm{var}(d_k) &= E(d_k^2) = E\{E(d_k^2|i_k)\}\\ &= \frac{1}{N}\sum_{i\in U}E(d_k^2|i_k = i) = \frac{1}{N}\sum_{i\in U}\sigma_i^2 = \sigma_d^2;\\ \mathrm{cov}(d_k, y_k) &= E\{E(d_k y_k|i_k)\} = \frac{1}{N}\sum_{i\in U}Y_i E(d_k|i_k = i) = 0.\end{aligned}$$

In the same way it can be shown that d_k is uncorrelated with y_l and d_l ($l \neq k$). Now using $\bar{\hat{y}}_s = \bar{y}_s + \bar{d}_s$, it is seen that

$$\begin{aligned}\mathrm{var}(\bar{\hat{y}}_s) &= \mathrm{var}(\bar{y}_s + \bar{d}_s) = \mathrm{var}(\bar{y}_s) + \mathrm{var}(\bar{d}_s)\\ &= (\frac{1}{n} - \frac{1}{N})S_y^2 + \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{var}(d_k)\\ &= (\frac{1}{n} - \frac{1}{N})S_y^2 + \frac{\sigma_d^2}{n},\end{aligned}$$

from which (1) follows. This concludes the proof.

An unbiased estimator of the variance in (1) is

$$\hat{\mathrm{var}}(\hat{Y}_{TSS}) = N^2\{(\frac{1}{n} - \frac{1}{N})s_{\hat{y}}^2 + \frac{1}{Nn}\sum_{k=1}^{n}\hat{\sigma}_k^2\}, \qquad (2)$$

where $\hat{\sigma}_k^2 = M_k^2(1 - f_k)s_k^2/m_k$.

Proof that (2) *is unbiased.* The key feature in a TSS design is that $d_1,...,d_n$ are mutually uncorrelated random variables with zero expectation and variance σ_d^2.

Therefore, similar to two-stage sampling with replacement in the first stage, $E(s_d^2) = \sigma_d^2$; recall $E(s_y^2) = S_y^2$. As we have seen, d_k is uncorrelated with $y_1,...,y_n$. Hence,

$$E(s_{\hat{y}}^2) = E(s_{y+d}^2) = E(s_y^2 + s_d^2 + 2s_{yd}) = S_y^2 + \sigma_d^2. \tag{3}$$

Furthermore, in analogy with $E(\bar{\hat{y}}_s) = Y/N$,

$$E(\frac{1}{n}\sum_{k=1}^{n}\hat{\sigma}_k^2) = \frac{1}{N}\sum_{i\in U}\sigma_i^2 = \sigma_d^2 . \tag{4}$$

Using (3) and (4), it is seen that $\hat{\text{var}}(\hat{Y}_{TSS})$ in (2) is unbiased for (1). This concludes the proof.

Finally, observe that $\text{var}(d_k) = \sigma_d^2$ also holds if *s* in the first stage were drawn with replacement (TSSR). Moreover, in a TSSR design $d_1,...,d_n$ are independent. In contrast, in a TSS design $d_1,...,d_n$ are uncorrelated but need not be independent. For instance, assume that *N=3*, *n*=2, $\sigma_1^2 = \sigma_2^2 = 0$ and $\sigma_3^2 > 0$. Then $d_1 \neq 0$ implies $P(d_2 = 0) = 1$. Hence, d_1 and d_2 are not independent. For further details, see Knottnerus (2003, pages 157-158). In addition, recall that in the above notation the TSSR formulas are $\hat{Y}_{TSSR} = N\bar{\hat{y}}_s$, and

$$\text{var}(\hat{Y}_{TSSR}) = N^2\frac{\sigma_y^2 + \sigma_d^2}{n}, \tag{5}$$

where $\sigma_y^2 = (N-1)S_y^2/N$. The standard unbiased variance estimator is $\hat{\text{var}}(\hat{Y}_{TSSR}) = N^2 s_{\hat{y}}^2/n$; see, among others, Cochran (1977, pages 306-307). It emerges from (1) and (5) that the second stage in both TSS and TSSR leads to the same increase of the variance by the amount $N^2\sigma_d^2/n$.

References

Chaudhuri, A. and Stenger, H. (2005). *Survey Sampling*: *Theory and Methods*, Chapman & Hall/CRC, New York.

Cochran, W.G. (1977). *Sampling Techniques*, John Wiley & Sons, New York.

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*, Springer-Verlag, New York.