EXACT CONFIDENCE INTERVALS FOR FINITE POPULATION PARAMETERS IN TWO-STAGE SAMPLING

Richard K. Burdick, Arizona State University
Robert L. Sielken, Jr., Texas A&M University[1]

INTRODUCTION

The linear least-squares prediction approach has recently been applied by Royall [1976] in two-stage sampling from finite populations. Royall develops alternative estimators and their variances for the finite population total and compares them under various situations. This paper considers a special case of the super-population model assumed by Royall and discusses a technique for the unbiased estimation of variance and construction of an exact confidence interval on the finite population total.

THE MODEL FOR TWO-STAGE SAMPLING

A finite population of K elements is separated into N clusters of size M_i where $\sum_{i=1}^{N} M_i = K$. Letting y_{ij} denote the value associated with the j^{th} element in cluster i, the model describing the super-population from which the K elements are assumed to have been selected is

$$y_{ij} = \mu + \eta_{ij} \tag{1}$$

where the η_{ij}'s are normal random variables with mean zero and

$$\begin{aligned} E(\eta_{ij}\,\eta_{k\ell}) &= \tau^2 + \sigma^2, && i = k,\ j = \ell, \\ &= \tau^2, && i = k,\ j \neq \ell, \\ &= 0, && i \neq k. \end{aligned} \tag{2}$$

This two-stage model has previously been used by Fuller [1973] to estimate parameters of the super-population. It is also a special case of the model used by Royall [1976] and Scott and Smith [1969] in which the variance of y_{ij} is constant for all i. Royall also uses this simplified model when comparing alternative estimators for the population total and when considering efficient sample designs.

The methodology used by Royall [1976] in estimating the finite population total involves selecting a random sample s of n clusters, and from the M_i elements in each of the sampled clusters selecting a random sample s_i of size m_i. The finite population total is then partitioned into the sum of sampled elements, the sum of non-sampled elements from sampled clusters, and the sum of non-sampled elements from non-sampled clusters. The sum of the non-sampled elements is then estimated using the combined knowledge of s and the assumed super-population model.

One of the estimators for the population total, $T = \sum_{i}^{N} \sum_{j}^{M_i} y_{ij}$, suggested by Royall is

$$\hat{T}_H = \sum_{i\varepsilon s} \sum_{j\varepsilon s_i} y_{ij} + \sum_{i\varepsilon s} (M_i - m_i)\bar{y}_i + \sum_{i\varepsilon s} (m_i\bar{y}_i/k) \sum_{i\notin s} M_i \tag{3}$$

where $\bar{y}_i = \sum_{j\varepsilon s_i} y_{ij}/m_i$, $i\varepsilon s$, (4)

and $k = \sum_{i\varepsilon s} m_i$.

This estimator will be used to illustrate a new technique which constructs an exact confidence interval on T. Under the super-population model assumed in (1), the variance of $(\hat{T}_H - T)$ is

$$V(\hat{T}_H - T) = \tau^2 \left\{ \sum_{i\notin s} M_i^2 + \frac{\left(\sum_{i\notin s} M_i\right)^2 \sum_{i\varepsilon s} m_i^2}{k^2} \right\} + \sigma^2 \left\{ \sum_{i\varepsilon s} \frac{M_i^2}{m_i} - \frac{\left(\sum_{i\varepsilon s} M_i\right)^2}{k} + K\left(\frac{K}{k} - 1\right) \right\}. \tag{5}$$

A comment should be made about the preceding results and those to follow in this section. When clusters are of unequal size, even though $M_1,\ldots,M_N$ are fixed and assumed known, in a strict probabilistic sense the $M_1,\ldots,M_n$ corresponding to the sampled clusters are really random variables whose realization depends upon which clusters are sampled. Hence, the arguments used above and those to follow are really conditional arguments for given values of $M_1,\ldots,M_n$. However, since the unbiasedness of $\hat{T}_H$ and the confidence level of the corresponding confidence interval will not depend upon the values of $M_1,\ldots,M_n$, these properties will also apply in an unconditional sense.

The problems of estimating a linear combination of variance components such as $V(\hat{T}_H - T)$ for the unbalanced case are well known and the interested reader is referred to Searle [1971] for a complete discussion. However, new results due to Burdick and Sielken [1977] can be used to construct an exact confidence interval on T. The method considers the random variable $U_i = c_1 y_i + c_{2i} d_i$ for $i\varepsilon s$, where $d_i = \sum_{j\varepsilon s_i} \ell_{ij} y_{ij}$, $\sum_{j\varepsilon s_i} \ell_{ij} = 0$,

$c_{3i} = \sum_{j\epsilon s_i} \ell_{ij}^2$, and the ℓ_{ij}'s, c_1, c_{2i}'s, and c_{3i}'s are constants. Under model (1), $V(U_i) = c_1^2\tau^2 + (c_{2i}^2\, c_{3i} + c_1^2/m_i)\,\sigma^2$. Thus, with

$$c_1^2 = \sum_{i\notin s} M_i^2 + \left(\sum_{i\notin s} M_i\right)^2 \sum_{i\epsilon s} m_i^2/k^2 \quad (6)$$

and

$$c_{2i}^2\, c_{3i} = \sum_{i\epsilon s} \frac{M_i^2}{m_i} - \frac{\left(\sum_{i\epsilon s} M_i\right)^2}{k} + K\left(\frac{K}{k} - 1\right) - c_1^2/m_i \quad (7)$$

the U_i's are independent identically distributed $N(c_1\mu,\ V(\hat{T}_H - T))$. Letting $a = \{i \mid c_{2i}^2\, c_{3i} \geq 0\}$ and b denote the number of elements in set a, then $\sum_{i\epsilon a} (U_i - \bar{U})^2/V(\hat{T}_H - T) \sim \chi^2_{(b-1)}$ where $\bar{U} = (1/b) \sum_{i\epsilon a} U_i$. An unbiased estimator for $V(\hat{T}_H - T)$ is therefore

$$v_H = \frac{1}{b-1} \sum_{i\epsilon a} (U_i - \bar{U})^2. \quad (8)$$

For the special case where all $m_i = m$, $b = n$.

Burdick [1976] has shown that when $M_i = M$ and $m_i = m$ for all i, $(\hat{T}_H - T)$ is independent of $(b-1)\, v_H/V(\hat{T}_H - T)$. Thus, since $(\hat{T}_H - T) \sim N(0, V(\hat{T}_H - T))$ and $(b-1)\, v_H/V(\hat{T}_H - T) \sim \chi^2_{(b-1)}$, it follows that in this case $(\hat{T}_H - T)/\sqrt{v_H}$ will have an exact t-distribution with (b-1) degrees of freedom and that an exact $100(1 - \delta)\%$ confidence interval on T is

$$[\hat{T}_H \pm t_{\delta/2;\ b-1} \sqrt{v_H}]. \quad (9)$$

It should be noted that (9) is an exact confidence interval for any choice of ℓ_{ij} as long as $\sum_{j\epsilon s_i} \ell_{ij} = 0$ and equations (6) and (7) are satisfied for all i. Since the length of the confidence interval is determined by the value of $\sqrt{v_H}$, it would seem to be important to minimize this quantity when selecting the ℓ_{ij}. However, since the distribution of v_H does not depend on ℓ_{ij}, any convenient set of ℓ_{ij} may be used.

For example, if m is even, let

$$\ell_{ij} = -1,\ j = 1,\ldots,\frac{m}{2},$$
$$= +1,\ j = \frac{m}{2} + 1,\ldots,\ m, \quad (10)$$

and, if m is odd, let

$$\ell_{ij} = -1,\ j = 1,\ldots,\frac{m-1}{2},$$
$$= 0,\ j = \frac{m+1}{2},$$
$$= +1,\ j = \frac{m+3}{2},\ldots,m, \quad (11)$$

for all i. As discussed by Burdick and Sielken [1977], these values represent a good choice with respect to the robustness of the confidence interval to model breakdown.

In the more general case where either all of the M_i's or all of the m_i's are not equal $(\hat{T}_H - T)$ is not necessarily independent of v_H and the confidence interval given by (9) is only approximate. Furthermore, when all of the m_i's are not equal it is possible that $b < n$. This implies that some of the observations used in calculating $\hat{T}_H$ would be ignored in the calculation of v_H. This weakness can sometimes be avoided by using a "pooling" procedure to estimate v_H as suggested by Burdick and Sielken [1977].

FOOTNOTES

[1]This work was done at Texas A&M and supported by a grant from the Army Research Office, contract number DAHC04-74-C0018.

REFERENCES

Burdick, R. K. [1976]. "A Super-Population Approach to Multi-Stage Sampling" Ph.D. dissertation, Institute of Statistics, Texas A&M University.

Burdick, R. K., and Sielken, R. L. [1977]. Exact Confidence Intervals for Linear Combinations of Variance Components in Nested Classifications. Article presently under review by J. Amer. Statist. Assn.

Fuller, W. A. [1973]. Regression Analysis for Sample Surveys, paper presented at the Vienna meeting of the International Institute of Survey Statisticians.

Royall, R. M. [1976]. The Linear Least-Squares Prediction Approach to Two-Stage Sampling. J. Amer. Statist. Assn. 71, 657-664.

Scott, Alastair, and Smith, T. M. F. [1969]. Estimation in Multi-Stage Surveys. J. Amer. Statist. Assn. 64, 830-840.

Searle, S. R. [1971]. Topics in Variance Components Estimation. Biometrics 27, 1-76.