Journal of Official Statistics, Vol. 21, No. 3, 2005, pp. 433–439

A Note on the Hartley-Rao Variance Estimator

Phillip S. Kott[1]

The Hartley-Rao variance formula was designed to estimate the randomization variance of a Horvitz-Thompson estimator given a systematic probability proportional to size sample from a randomly ordered large population. Using an underappreciated formulation of this variance estimator, one can see that the Hartley-Rao variance estimator is unbiased under a model with a particular error structure given *any* sample. Moreover, even with a more general error structure, this variance estimator remains nearly model unbiased for a large sample and relatively larger population under mild conditions. A discussion follows concerning an extension of Hartley-Rao variance estimation to linear calibration estimators.

Key words: Goodman-Kish design; Sampford design; finite population correction factor; relatively larger population; linear calibration estimator.

1. Introduction

Let $\overline{Y} = \Sigma_U y_i/N$ denote a population mean, where U is a population of N units. The Horvitz-Thompson estimator for $\overline{Y}$ is $t_{HT} = N^{-1}\Sigma_S y_i/\pi_i$, where S is a sample of fixed size n, and π_i is the selection probability of unit i. It is well-known that t_{HT} is randomization unbiased; that is, unbiased with respect to the random sampling mechanism.

Suppose the units in the sample were drawn using a Goodman-Kish design, i.e., systematically from a randomly ordered list with arbitrary probabilities of selection (see Brewer and Hanif 1983, p. 22). Hartley and Rao (1962) offer the following estimator for the randomization variance of t_{HT}:

$$v_{HR} = 2^{-1}N^{-2}[(n-1)]^{-1}\sum_{i\in S}\sum_{j\in S}\left(1-\pi_i-\pi_j+\sum_{k\in U}\pi_k^2/n\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 \qquad (1)$$

It is proven there that when the population is large compared to the sample, v_{HR} is nearly randomization unbiased. Formally, if $max\{\pi_i\} < c$, and c is $O(N^{-1})$, then the randomization bias of v_{HR} is $O(N^{-2})$.

Asok and Sukhatme (1976) effectively show that v_{HR} has the same property under the Sampford sampling design. Moreover, the authors prove that using the Sampford design has, if anything, slightly less randomization variance (ignoring $O(N^{-3})$ terms) than the Goodman-Kish design. A short description of the Sampford design follows: Select unit i with probability proportional to π_i on the first of n draws and probability proportional to

[1] Chief Research Statistician, Research and Development Division, National Agricultural Statistics Service, Fairfax, VA 22030, U.S.A. Email: pkott@nass.usda.gov

Acknowledgment: The author would like to thank several anonymous referees and a diligent associate editor whose comments improved the quality of this article.

$\pi_i/(1-\pi_i)$, with replacement, on the subsequent $n-1$ draws; if any unit is selected more than once, reject the sample and begin again. Brewer and Hanif (1983) show that the Sampford design (which they called the Rao-Sampford design) is equivalent to many popular unequal probability schemes when $n = 2$.

Following Cumberland and Royall (1981), observe that the right-hand side of Equation (1) can be put in this more convenient form:

$$v_{HR} = N^{-2}[n/(n-1)]\sum_{i\in S}\left(1-\pi_i-\sum_{k\in S}\pi_k/n+\sum_{k\in U}\pi_k^2/n\right) \times\left(y_i/\pi_i-\left[n^{-1}\sum_{j\in S}y_j/\pi_j\right]\right)^2 \tag{2}$$

Although not new, this formulation of the Hartley-Rao variance estimator is not as well-known as it should be. The expression $1-\pi_i-\Sigma_S\pi_k/n+\Sigma_U\pi_k^2/n$, which collapses to $1-n/N$ in Equation (2) when all the π_j are equal, can be viewed as a term-wise finite population correction factor.

Suppose the y_i satisfy the model

$$y_i = \beta x_i + x_i\varepsilon_i \tag{3}$$

where the x_i are proportional to the π_i, and the ε_i are uncorrelated random variables with mean zero given x_i, then t_{HT} is a model unbiased estimator for $\overline{Y}$ in the sense that $E_\varepsilon(t_{HT}-\overline{Y}) = 0$ (see, for example, Brewer 1963). Cumberland and Royall (1981) show that v_{HR} is a model unbiased estimator for the model variance of t_{HT} when the ε_i have a common variance no matter what the corresponding x_i. As we shall see, v_{HR} remains a nearly model unbiased variance estimator under certain conditions when the ε_i are uncorrelated but have unequal variances. The model-based properties of v_{HR} do not require either a Goodman-Kish or Sampford sampling design.

One restriction on the population necessary for any without replacement probability-proportional-to-x_i sampling scheme is that no x_i be larger than $\Sigma_U x_j/n$. We assume that the population under study has this property.

2. Cumberland and Royall's Model-Based Result

Let the model variance of ε_i in Equation (3) be v_i. Since n is fixed, $\pi_i = nx_i/(N\overline{X})$, where $N\overline{X} = \Sigma_U x_i$. The model variance of t_{HT} is

$$E_\varepsilon[(t_{HT}-\overline{Y})^2] = E_\varepsilon\left\{\left[N^{-1}\sum_{i\in S}(\beta x_i+x_i\varepsilon_i)/\pi_i - N^{-1}\sum_{i\in U}(\beta x_i+x_i\varepsilon_i)\right]^2\right\}$$

$$= N^{-2}E_\varepsilon\left\{\left[\sum_{i\in S}x_i\varepsilon_i/\pi_i-\sum_{i\in U}x_i\varepsilon_i\right]^2\right\}$$

$$= (\overline{X}/n)^2 E_\varepsilon \left\{ \left[\sum_{i\in S} \varepsilon_i - \sum_{i\in U} \pi_i \varepsilon_i \right]^2 \right\}$$

$$= (\overline{X}/n)^2 \left\{ E_\varepsilon \left[\left(\sum_{i\in S} \varepsilon_i \right)^2 \right] - 2E_\varepsilon \left[\sum_{i\in S} \varepsilon_i \sum_{i\in U} \pi_i \varepsilon_i \right] + E_\varepsilon \left[\sum_{i\in U} \pi_i \varepsilon_i \right]^2 \right\}$$

$$= (\overline{X}/n)^2 \left(\sum_{i\in S} v_i - 2\sum_{i\in S} \pi_i v_i + \sum_{i\in U} \pi_i^2 v_i \right) \tag{4}$$

while the model expectation of v_{HR} is

$$E_\varepsilon(v_{HR}) = E_\varepsilon \left\{ N^{-2}[n/(n-1)] \sum_{i\in S} \left(1 - \pi_i - \sum_{k\in S} \pi_k/n + \sum_{k\in U} \pi_k^2/n \right) \right.$$

$$\left. \times \left(\{\beta x_i + x_i \varepsilon_i\}/\pi_i - \left[n^{-1} \sum_{j\in S} \{\beta x_j + x_i \varepsilon_j\}/\pi_j \right] \right)^2 \right\}$$

$$= N^{-2}[n/(n-1)] \sum_{i\in S} \left(1 - \pi_i - \sum_{k\in S} \pi_k/n + \sum_{k\in U} \pi_k^2/n \right)$$

$$\times E_\varepsilon \left\{ \left(x_i \varepsilon_i/\pi_i - \left[n^{-1} \sum_{j\in S} x_j \varepsilon_j/\pi_j \right] \right)^2 \right\}$$

$$= (\overline{X}/n)^2 [n/(n-1)] \sum_{i\in S} \left(1 - \sum_{j\in S} \pi_j/n + \sum_{j\in U} \pi_j^2/n - \pi_i \right)$$

$$\times E_\varepsilon \left\{ \left(\varepsilon_i - \left[n^{-1} \sum_{j\in S} \varepsilon_j \right] \right)^2 \right\}$$

$$= (\overline{X}/n)^2 [n/(n-1)] \sum_{i\in S} \left(1 - \sum_{j\in S} \pi_j/n + \sum_{j\in U} \pi_j^2/n - \pi_i \right)$$

$$\times (v_i - 2v_i/n + v_0/n) \tag{5}$$

where $v_0 = \Sigma_S v_j/n$.

When all the v_i in the population are equal to v_0, the right-hand sides of Equations (4) and (5) are both

$$V_0 = (\overline{X}^2/n)\left(1 - 2\sum_{i\in S}\pi_i/n + \sum_{i\in U}\pi_i^2/n\right)v_0$$

Thus, when the model in Equation (3) holds, and the ε_i have a common variance, v_{HR} is an unbiased estimator of the model variance of t_{HT}. Note that this is true no matter how the sample is drawn.

Even when the v_i are not all equal, v_{HR} may be almost model unbiased in some sense. For example, when $max\{\pi_i\} < c$, where c is $O(N^{-1})$, the model bias of v_{HR} is $O(N^{-1})$ since both the model variance of t_{HT} (from Equation (4)) and the model expectation of v_{HR} (from Equation (5)) are

$$V_N = (\overline{X}/n)^2\sum_{i\in S}v_i + O(N^{-1})$$

Ignoring $O(N^{-1})$ terms, but not $O(n^{-1})$ terms, is equivalent to completely ignoring finite population correction.

3. Large-Sample Results

Isaki and Fuller (1982) formulate a more standard asymptotic theory where the sample size, rather than the population size, is assumed to be large. Of course, N must be larger than n in this framework, but $1/N$ need not be any smaller than $O(1/n)$; that is to say, n/N can converge to a positive constant as n grows arbitrarily large.

Isaki and Fuller provide sufficient conditions on the sampling design and population for $\Sigma_S x_j/n - \Sigma_U \pi_j x_j/n$ and $\Sigma_S x_j v_j/n - \Sigma_U \pi_j x_j v_j/n$ to be $O_P(n^{-1/2})$, while $\Sigma_U x_j/N = O(1)$, and $\Sigma_S x_j/n$, $\Sigma_S v_j/n$ and $\Sigma_S x_j v_j/n$ are $O_P(1)$. Under these conditions, it is not hard to see that the model variance of t_{HT} is itself $O_P(1/n)$. Moreover, the model variance of t_{HT} and the model expectation of v_{HR} are both

$$V_n = (\overline{X}/n)^2\sum_{i\in S}(1 - \pi_i)v_i + O_P(n^{-3/2})$$

since $-\Sigma_S \pi_i v_i/n + \Sigma_U \pi_i^2 v_i/n$ in Equation (4) and $-\Sigma_S \pi_j/n + \Sigma_U \pi_j^2/n$ in Equation (5) are $O_P(n^{-1/2})$. Recall that $\pi_i = nx_i/\Sigma_U x_k$.

Note that although we are concerned here with model bias, we are invoking the large-sample properties of the sample design. We are not, however, averaging over all possible samples in this context as randomization-based theory does. Moreover, we will not call v_{HR} "nearly model unbiased" unless its model bias - the difference between the model variance of t_{HT} and model expectation of v_{HR} - is $O_P(n^{-2})$.

Suppose now that n is large and N is relatively larger, but not so much so that finite population correction can be completely ignored. This often happens in practice; n is large enough (say $n \approx 50$) for conventional asymptotic properties to have relevance; N is larger still (say $N \approx 250$), but finite population correction can have an effect on variance estimation.

Following Kott (1990), we formalize the notion of a relatively large population (compared to a large sample) by assuming $1/N$ is $O(n^{-3/2})$ rather than $O(1/n)$. If

$max\{\pi_i = nx_i/(N\overline{X})\} < c$, where c is now $O(n/N) = O(n^{-1/2})$, then the sampling design and population are such that both $-\Sigma_S \pi_j/n + \Sigma_U \pi_j^2/n$ and $-\Sigma_S \pi_i v_i/n + \Sigma_U \pi_i^2 v_i/n$ are $O_P(n^{-1})$ under mild conditions we assume to hold. As a consequence, the model bias of v_{HR} is $O_P(n^{-2})$; that is to say, v_{HR} is nearly unbiased under the model in Equation (3) with uncorrelated errors when the sample is large and the population relatively larger.

4. Discussion

Although neither a Goodman-Kish nor a Sampford sampling design is necessary for v_{HR} to have desirable model-based properties, the use of one of these selection schemes coupled with mild restrictions on the population is sufficient. Moreover, under either of these two designs, v_{HR} has an $O(N^{-2})$ bias as an estimator for the randomization variance of t_{HT} when $max\{\pi_i\} < c$ and c is $O(N^{-1})$.

Like Cumberland and Royall (1981), we have focused on a formula that estimates both the randomization and model variance of the Horvitz-Thompson estimator well under Goodman-Kish or Sampford sampling. The simultaneous variance estimator, v_{HR} in Equation (2), differs from the formulation in Brewer and Donadio (2003) (combining its Equations (16) and (18)) that attempts to estimate the model expectation of the randomization variance of the Horvitz-Thompson when all the v_i are equal.

A word of caution is in order. Goodman-Kish and Sampford sampling are often used in practice within tightly-defined design strata. When the sample size within each stratum is small, the large-sample model-based results of the last section have little relevance. Moreover, the advantage of the Cumberland-Royall formulation v_{HR} in Equation (2) over Equation (1) is muted.

Recently, however, the National Agricultural Statistics Service (NASS) has been using unequal probability designs either without strata or within very large strata (see Kott and Bailey 2000). The estimators employed by NASS have linear calibration form. In the context of estimating $\overline{Y}$, such an estimator can be rendered as $t_{LC} = N^{-1}\Sigma_S w_i y_i$, where the calibration weight for unit i is $w_i = (1/\pi_i)[1 + (\Sigma_U \mathbf{x}_k - \Sigma_S \mathbf{x}_k/\pi_k)(\Sigma_S \mathbf{h}_k' \mathbf{x}_k/\pi_k)^{-1}\mathbf{h}_i']$, and $\mathbf{x}_k$ and $\mathbf{h}_k$ are row vectors of the same dimension.

Estevao and Särndal (2000) discuss conditions under which t_{LC} has good model - assuming $y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$ with $E(\varepsilon_i|\mathbf{x}_i) = 0$ - and randomization-based properties. Under those conditions, and given either a Goodman-Kish or Sampford sample, a reasonable estimator for the randomization mean squared error of t_{LC} is

$$v_{HR} = N^{-2}[n/(n-1)]\sum_{i\in S}\left(1 - \pi_i - \sum_{k\in S}\pi_k/n + \sum_{k\in U}\pi_k^2/n\right) \times \left(w_i e_i - \left[n^{-1}\sum_{j\in S} w_j e_j\right]\right)^2 \quad (6)$$

where $e_i = y_i - \mathbf{x}_i(\Sigma_S \mathbf{h}_k' \mathbf{x}_k/\pi_k)^{-1}\Sigma_S \mathbf{h}_k' y_k/\pi_k$. Following Särndal, Swensson, and Wretman (1989), this formulation effectively replaces y_i/π_i in Equation (2) by $w_i e_i$ for model-based reasons.

Using arguments like those made earlier, v_{HR} in Equation (6) is also a nearly unbiased estimator for the model variance of t_{LC} under mild conditions when the ε_i are uncorrelated, the sample size is large, and the population size is relatively larger.

When a component of the vector $\mathbf{h}_k$, or some linear combination of components, is constant across the k (formally, $\mathbf{h}_i\mathbf{g} = 1$ for some column vector $\mathbf{g}$), the $\Sigma_S w_j e_j$ term in Equation (6) is equal to zero. To see why, first note that $\Sigma_S w_i e_i = \Sigma_S e_i/\pi_i$ by the way e_i is defined. When $\mathbf{h}_i\mathbf{g} = \mathbf{g}'\mathbf{h}_i' = 1$, we have

$$\begin{aligned}\sum_{i\in S} e_i/\pi_i &= \sum_{i\in S} y_i/\pi_i - \sum_{i\in S}(\mathbf{x}_i/\pi_i)\left(\sum_{k\in S}\mathbf{h}_k'\mathbf{x}_k/\pi_k\right)^{-1}\sum_{k\in S}\mathbf{h}_k' y_k/\pi_k \\ &= \sum_S y_i/\pi_i - \sum_S\left(\mathbf{g}'\mathbf{h}_i'\mathbf{x}_i/\pi_i\right)\left(\sum_S \mathbf{h}_k'\mathbf{x}_k/\pi_k\right)^{-1}\sum_S \mathbf{h}_k' y_k/\pi_k \\ &= \sum_S y_i/\pi_i - \sum_S \mathbf{g}'\mathbf{h}_k' y_k/\pi_k = 0\end{aligned}$$

The finite population correction factors $1 - \pi_i - \Sigma_S\pi_k/n + \Sigma_U\pi_k^2/n$ in Equation (6) can be ignored in practice when $\max\{\pi_i | i \in S\}$ and $\Sigma_U\pi_k^2/n$ are ignorably small. Both the randomization and model-based properties of v_{HR} depend on n being large. From a model-based point of view, even under ideal conditions ($\mathrm{Var}(\varepsilon_i|\mathbf{x}_i) \propto \pi_i^2$), e_i is only equal to ε_i *asymptotically*. Similarly, from a randomization-based viewpoint, w_i only need converge to $1/\pi_i$ as the sample size grows arbitrarily large. See Kott (2005) for a discussion of the relative speed of the asymptotics.

5. References

Asok, C. and Sukhatme, B.V. (1976). On Sampford's Procedure of Unequal Probability Sampling Without Replacement. Journal of the American Statistical Association, 71, 912–918.

Brewer, K.R.W. (1963). Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process. The Australian Journal of Statistics, 5, 93–105.

Brewer, K.R.W. and Donadio, M.E. (2003). The High Entropy Variance of a Horvitz-Thompson Estimator. Survey Methodology, 189–196.

Brewer, K.R.W. and Hanif, M. (1983). Sampling With Unequal Probabilities. New York: Springer-Verlag.

Cumberland, W.G. and Royall, R.M. (1981). Prediction Models and Unequal Probability Sampling. Journal of the Royal Statistical Society, Series B, 43, 353–367.

Estevao, V.M. and Särndal, C.E. (2000). A Functional Form Approach to Calibration. Journal of Official Statistics, 16, 379–399.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling With Unequal Probabilities and Without Replacement. Annals of Mathematical Statistics, 33, 350–374.

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

Kott, P.S. (1990). Estimating the Conditional Variance of a Design Consistent Regression Estimator. Journal of Statistical Planning and Inference, 24, 287–296.

Kott, P.S. (2005). Randomization-Assisted Model-Based Survey Sampling. Journal of Statistical Planning and Inference, 129, 263–277.

Kott, P.S. and Bailey, J.T. (2000). The Theory and Practice of Maximal Brewer Selection. Proceedings of the 2nd International Conference on Establishment Surveys, Invited papers, 269–278.

Särndal, C.E., Swensson, B., and Wretman, J. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of a Finite Population Total. Biometrika, 76, 527–537.

Received January 2004
Revised October 2004