Jackknife Variance Estimation for Two Samples after Imputation under Two-Phase Sampling

Jong-Min Kim and Jon E. Anderson

Division of Science and Mathematics
University of Minnesota at Morris
Morris, MN, 56267, USA

Abstract

We propose a jackknife variance estimator for the population average from two, two-phase samples after imputation. The jackknife method has long been used to estimate and reduce bias, but has now become a valuable tool for variance estimation. We apply two different sampling methods, (simple random sampling and stratified random sampling) to derive jackknife variance estimators for the two-sample case after imputation under two-phase sampling.

1. Introduction

The jackknife method was designed to estimate the bias of an estimator, but now has become a valuable tool for variance estimation since the work of Tukey (1958). In an infinite population context, Tukey (1958) suggested that each jackknife-replicate estimate might be regarded as an independent and identically distributed random variable, which in turn suggests a very simple variance estimator. In the finite population sampling context, each jackknife replicate deletes one unit and modifies the weights of others. Although the one-sample jackknife method was investigated and widely used by many authors, the research on two-sample jackknife estimators is quite limited. Arvesen (1969) was the first to propose such a jackknife estimator as follows: let $X_1, ..., X_{n_1}$ and $X_1, ..., X_{n_2}$ be two independent samples of sizes n_1 and n_2 from the population. Suppose, in simple random samples of size n_1 and n_2, r_1 and r_2 units respond, and m_1 and m_2 do not respond. The observed x values in population k are used as donors to create imputed values for the missing data. For example in sample k, a simple random sample of m_k values from the r_k responders could be used as imputed values for the nonresponders. That is, the imputed value x_{ik}^* for the i^{th} observation in sample k would be one of the observed x values for one of the responders x_{jk} in the set of responders s_{rk}. The imputed estimator of the population mean of X using a sample k is given by

$$\overline{x}_{kI} = \frac{1}{n_k}\left(r_k\overline{x}_{r_k} + m_k\overline{x}_{m_k}\right), \tag{1}$$

where $\bar{x}_{m_k}$ is the mean of the imputed values.

2. Variance Estimation Under Two-Phase Sampling

Two-phase sampling or double sampling is often employed when it is relatively cheap to take a large preliminary sample in which an auxiliary variable x, correlated with a characteristic of interest y alone is measured. The first-phase sample gives a good estimate, $\overline{x}^{'}$, of the population mean,$\overline{X}$, while the second-phase subsample in which y is measured is employed to estimate the population mean $\overline{Y}$ through ratio estimation using $\overline{x}^{'}$ and $\overline{y}$.

2.1. Ratio Estimator in Simple Random Sampling

We now extend the two-phase sampling ideas to two samples. Suppose two, first-phase simple random samples $s_1^{'}$ of size $n_1^{'}$ and $s_2^{'}$ of size $n_2^{'}$ are taken without replacement from a population of N elements. Auxiliary attributes x_{1i}, x_{2i}, are observed for all elements $i \in s_1^{'}$, $i \in s_2^{'}$. Simple random subsamples s_1 of size n_1 and s_2 of size n_2 are taken without replacement from $s_1^{'}$ and $s_2^{'}$. Ratio estimators of $\overline{Y}$, are $\overline{y}_{r1} = (\overline{y}_1/\overline{x}_1)\overline{x}_1^{'} = \hat{R}_1\overline{x}_1^{'}$, $\overline{y}_{r2} = (\overline{y}_2/\overline{x}_2)\overline{x}_2^{'} = \hat{R}_2\overline{x}_2^{'}$, where $\overline{x}_1^{'}$ and $\overline{x}_2^{'}$ are the means for the first-phase samples $s_1^{'}$ and $s_2^{'}$ and $(\overline{x}_1, \overline{y}_1)$ and $(\overline{x}_2, \overline{y}_2)$ are the means for the second-phase samples s_1 and s_2. These estimators are design-consistent for $\overline{Y}$ as $n_1 \to \infty$ and $n_2 \to \infty$.

In single-phase sampling, jackknife variance estimation is often used when $\overline{X}$, is known. Under a model-based

framework, Royall and Eberhardt (1975) have shown that the jackknife variance estimator is asymptotically equivalent to a robust variance estimator. As a result, it performs well both unconditionally and conditionally given the ancillary statistic $\overline{x}$. The jackknife method for single-phase sampling is not readily applicable to two-phase sampling since y_i is not observed if $i \in s_k^{'} - s_k$ in samples $k = 1, 2$. We now obtain a jackknife variance estimator for $\overline{y}_{rk}$ by recalculating $\overline{y}_{rk}$ with the jth element removed for each $j \in s_k^{'}$ then using the variance of these $n_k^{'}$ jackknife values, $\overline{y}_{rk}(-j)$. Clearly, deleting unit j will affect $\overline{x}_1$ and $\overline{x}_2$ only if $j \in s_k$ and not if $j \in s_k^{'} - s_k$, for sample $k = 1, 2$ while it will affect $\overline{x}_k^{'}$ for all $j \in s_k^{'}$. Thus, we define

$$\overline{y}_{rk}(-j) = [\overline{y}_k(-j)/\overline{x}_k(-j)]\overline{x}_k^{'}(-j),$$

for all $j \in s_k^{'}$, where

$$\overline{x}_k(-j) = \begin{cases} \frac{n_k\overline{x}_k - x_j}{n_k - 1}, & \text{if } j \in s_k, \\ \overline{x}_k, & \text{if } j \in s_k^{'} - s_k \end{cases} \tag{2}$$

$$\overline{y}_k(-j) = \begin{cases} \frac{n_k\overline{y}_k - y_j}{n_k - 1}, & \text{if } j \in s_k, \\ \overline{y}_k, & \text{if } j \in s_k^{'} - s_k \end{cases} \tag{3}$$

and $\overline{x}_k^{'}(-j) = (n_k^{'}\overline{x}_k^{'} - x_j)/(n_k^{'} - 1)$ for all $j \in s_k^{'}$. Now apply the usual jackknife method to $\overline{y}_{rk}(-j)$ to get

$$v_{J_{rk}} = \frac{n_k^{'} - 1}{n_k^{'}} \sum_{j \in s_k^{'}} [\overline{y}_{rk}(-j) - \overline{y}_{rk}]^2 . \tag{4}$$

A linearized version of $v_{J_{rk}}$ for large n_k is obtained by noting that

$$\overline{y}_{rk}(-j) - \overline{y}_{rk} = \begin{cases} -\hat{R}_k \left(\frac{x_j - \overline{x}_k^{'}}{n_k^{'} - 1} \right) - \frac{\overline{x}_k^{'}(-j)}{\overline{x}_k(-j)} \left(\frac{x_j - \hat{R}_k x_j}{n_k - 1} \right), \\ \qquad\qquad \text{if } j \in s_k, \\ -\hat{R}_k \left(\frac{x_j - \overline{x}_k^{'}}{n_k^{'} - 1} \right), \quad \text{if } j \in s_k^{'} - s_k. \end{cases} \tag{5}$$

The jackknife variance estimator is a weighted average of two estimators, given by

$$\begin{aligned} v_{Jr} &= \frac{n_1^{'} v_{J_{r1}} + n_2^{'} v_{J_{r2}}}{n_1^{'} + n_2^{'}} \\ &= \frac{n_1^{'} - 1}{n_1^{'} + n_2^{'}} \sum_{j \in s_1^{'}} [\overline{y}_{r1}(-j) - \overline{y}_{r1}]^2 \\ &+ \frac{n_2^{'} - 1}{n_1^{'} + n_2^{'}} \sum_{l \in s_2^{'}} [\overline{y}_{r2}(-l) - \overline{y}_{r2}]^2 . \end{aligned} \tag{6}$$

Let's consider Rao and Shao (1992) adjusted jackknife variance estimator using data imputation framework. In one of our samples, $k = 1, 2$ we define our estimator of $\overline{Y}$ as $\overline{y}_{kI} = (1/n_k^{'}) \sum_{i \in s_k^{'}} y_i^*$. If the observation is part of the second phase sample in sample k , s_k, $y_i^* = y_i$, because it is observed. If the value of y is not directly observed because it is part of $s_k^{'} - s_k$ the value is obtained through ratio imputation as $y_i^* = (\overline{y}_k/\overline{x}_k)x_i$.

To use convenient one-phase sample variance formulae, Rao and Sitter (1995) proposed the following device to facilitate the computations. They defined

$$\hat{z}_{ki}(-j) = y_{ki}^* + \left\{ \frac{\overline{y}_k(-j)}{\overline{x}_k(-j)} x_{ki} - \frac{\overline{y}_k}{\overline{x}_k} x_{ki} \right\},$$

for sample $k = 1, 2$. Under this formulation, $\hat{z}_{ki}(-j) = y_{ki}^* = (\overline{y}_k/\overline{x}_k)x_{ki}$ for $j \in s_k^{'} - s_k$ in sample $k = 1, 2$, and $\hat{z}_{ki}(-j) = (\overline{y}_k(-j)/\overline{x}_k(-j))x_{ki}$ for $j \in s_k$ in sample $k = 1, 2$. We also define the adjusted estimator,

$$\overline{y}_{kI}^a(-j) = \frac{1}{n_k^{'} - 1} \sum_{i=1}^{n_k^{'}} \hat{z}_{ki}(-j),$$

and this helps define the jackknife variance estimator for sample k,

$$v_{J_{rk}} = \frac{n_k^{'} - 1}{n_k^{'}} \sum_{j \in s_k^{'}} [\overline{y}_{kI}^a(-j) - \overline{y}_{kI}]^2 , \tag{7}$$

where $\overline{y}_{kI} = \overline{y}_{kr}$ under ratio imputation. The jackknife variance estimator based on adjusted imputed estimators $\overline{y}_{kI}$, $k = 1, 2$ is a weighted average of two estimators, given by

$$\begin{aligned} v_{Jr}^a &= \frac{n_1^{'} v_{J_{r1}} + n_2^{'} v_{J_{r2}}}{n_1^{'} + n_2^{'}} \\ &= \frac{n_1^{'} - 1}{n_1^{'} + n_2^{'}} \sum_{j \in s_1^{'}} [\overline{y}_{1I}^a(-j) - \overline{y}_{1I}]^2 \\ &+ \frac{n_2^{'} - 1}{n_1^{'} + n_2^{'}} \sum_{l \in s_2^{'}} [\overline{y}_{2I}^a(-l) - \overline{y}_{2I}]^2 . \end{aligned} \tag{8}$$

2.2. Ratio Estimator in Stratified Random Sampling

Suppose the population of N units consists of L strata such that the h-th stratum consists of N_h units and $\sum_{h=1}^{L} N_h = N$. Suppose that an auxiliary variable, x, closely related to an item y is observed on all sample units, $s_{hk}^{'}$ in sample $k = 1, 2$ for stratum h. Ratio imputation uses $y_{hki}^* = (\overline{y}_{hk}/\overline{x}_{hk})x_{hki}$ for $i \in s_{hk}^{'} - s_{hk}$ where $\overline{y}_{hk}$ and $\overline{x}_{hk}$ are the means of y and x for the respondents s_{hk} in stratum h. Ratio imputation can be motivated by the fact that y_{hi}^*

is the best predictor of unobserved y_{hki} under the following "ratio" superpopulation model ξ:

$$E_\xi(y_{hki}) = \beta_h x_{hki}, \quad V_\xi(y_{hki}) = \sigma_h^2 x_{hki}, \\ \text{cov}_\xi(y_{hki}, y_{hkj}) = 0, \qquad \text{for } i \neq j \tag{9}$$

provided that the model also holds for the respondents s_{hk}; that is, if selection bias is absent. Note that response probabilities can depend on the x_{hki}'s. Särndal (1992) has named the above equation an "imputation" model. Under ratio imputation,

$$\overline{y}_{kI} = \sum_h^L W_h \overline{y}_{hkI} = \sum_h^L W_h (\overline{y}_{hk}/\overline{x}_{hk}) \overline{x}_{hk}^{'}, \tag{10}$$

where $\overline{x}_{hk}^{'}$ is the x mean for the full sample $s_{hk}^{'}$ from stratum h. Under the model (9), $\overline{y}_{kI}$ is design-model unbiased for $\overline{Y}$, provided that the model also holds for the respondents. Also, under a uniform response mechanism within each stratum h, the estimator (10) has the same properties as the standard two-phase sampling separate ratio estimator. This follows by noting that conditionally, given n_{hk}, s_{hk} is a simple random sample of fixed size n_{hk} drawn from $s_{hk}^{'}$. Rao (1996) mentioned that the estimator (10) is approximately design unbiased under uniform response in each stratum, proved that n_{hk} is large for each h. Note that strata act as imputation classes in the present context. It is readily seen that $y_{hki}^*(-hkj) = [\overline{y}_{hk}(-hkj)/\overline{x}_{hk}(-hkj)] x_{hki}^{'}$, under ratio imputation when hkjth respondent is deleted, where

$$\overline{y}_{hk}(-hkj) = [n_{hk}\overline{y}_{hk} - y_{hj}] / (n_{hk} - 1) \tag{11}$$

and

$$\overline{x}_{hk}(-hkj) = [n_{hk}\overline{x}_{hk} - x_{hj}] / (n_{hk} - 1), \tag{12}$$

for $k = 1, 2$.

To use convenient one-phase sample variance formulae, Rao and Sitter (1995) proposed the following device to facilitate the computations. We define

$$\hat{z}_{hki}(-hkj) = y_{hki}^* + \left\{ \frac{\overline{y}_{hk}(-hkj)}{\overline{x}_{hk}(-hkj)} x_{hki} - \frac{\overline{y}_{hk}}{\overline{x}_{hk}} x_{hki} \right\},$$

for sample $k = 1, 2$, stratum h. Under this formulation, $\hat{z}_{hki}(-hkj) = y_{hki}^* = (\overline{y}_{hk}/\overline{x}_{hk}) x_{hki}$ for $hkj \in s_{hk}^{'} - s_{hk}$ in sample $k = 1, 2$, and $\hat{z}_{hki}(-hkj) = (\overline{y}_{hk}(-hkj)/\overline{x}_{hk}(-hkj)) x_{hki}$ for $hkj \in s_{hk}$ in sample $k = 1, 2$, stratum h. We also define the adjusted estimator,

$$\overline{y}_{hkI}^a(-hkj) = \frac{1}{n_{hk}^{'} - 1} \sum_{i=1}^{n_{hk}^{'}} \hat{z}_{hki}(-hkj),$$

and this helps define the jackknife variance estimator for sample k.

Using these values, the jackknife variance estimator is given by

$$v_{J_r}(\overline{y}_{kI}) = \sum_{h=1}^{L} \frac{n_{hk}^{'} - 1}{n_{hk}^{'}} \sum_{j=1}^{n_{hk}^{'}} [\overline{y}_{kI}^a(-hkj) - \overline{y}_{kI}]^2 . \tag{13}$$

For $k = 1, 2$, we now obtain a linearized version of the jackknife variance estimator. This variance estimator is useful with computer programs that use the linearization method of variance estimation. Noting that $\overline{y}_{kI}^a(-hkj) - \overline{y}_{kI} = W_h [\overline{y}_{hkI}^a(-hkj) - \overline{y}_{hkI}]$, where $\overline{y}_{hkI}^a(-hkj)$ is the adjusted imputed estimator of the hth stratum mean $\overline{Y}_h$ when hkjth sample unit is deleted, we get

$$v_{J_r}(\overline{y}_{kI}) = \textstyle\sum_{h=1}^{L} W_h^2 v_{J_r}(\overline{y}_{hkI}) \\ = \sum_{h=1}^{L} W_h^2 \left[\frac{n_{hk}^{'} - 1}{n_{hk}^{'}} \sum_{j=1}^{n_{hk}^{'}} (\overline{y}_{hkI}^a(-hkj) - \overline{y}_{hkI})^2 \right].$$

The jackknife variance estimator is a weighted average of two estimators, given by

$$v_{Jrs} = \frac{n_1^{'} v_{J_r}(\overline{y}_{1I}) + n_2^{'} v_{J_r}(\overline{y}_{2I})}{n_1^{'} + n_2^{'}} \\ = \frac{n_1^{'}}{n_1^{'} + n_2^{'}} \left(\sum_{h=1}^{L} W_h^2 \left[\frac{n_{h1}^{'} - 1}{n_{h1}^{'}} \sum_{j=1}^{n_{h1}^{'}} (\overline{y}_{h1I}^a(-h1j) - \overline{y}_{h1I})^2 \right] \right) \\ + \frac{n_2^{'}}{n_1^{'} + n_2^{'}} \left(\sum_{h=1}^{L} W_h^2 \left[\frac{n_{h2}^{'} - 1}{n_{h2}^{'}} \sum_{j=1}^{n_{h2}^{'}} (\overline{y}_{h2I}^a(-h2j) - \overline{y}_{h2I})^2 \right] \right)$$

where $n_1^{'} = \sum_{h=1}^{2} n_{h1}^{'}$ and $n_2^{'} = \sum_{h=1}^{2} n_{h2}^{'}$.

3. Simulation Study

For our simulation of the Jackknife variance estimator in simple Random Sampling, we created a population of size $N = 10000$ by setting $Y = 0.8 \times X + \epsilon$, where $sd(\epsilon) = \sqrt{x}$. The population is displayed in Figure 3 To study Stratified Random Sampling, we used the same population but we divided the population into three strata: $X < 90$, $90 \leq X \leq 110$, $110 < X$. So stratum 1 size is $N_1 = 1633$ such that $W_1 = \frac{N_1}{N} = 0.1633$, stratum 2 size is $N_2 = 6805$ such that $W_2 = \frac{N_2}{N} = 0.6805$, and stratum 3 size is $N_3 = 1562$ such that $W_3 = \frac{N_3}{N} = 0.1562$.

In the figures that follow, we note that there is little difference in the Jackknife variance estimators for the cases studied here, but there are notable differences in the variation of these estimates. Figure 2 shows the reduction in

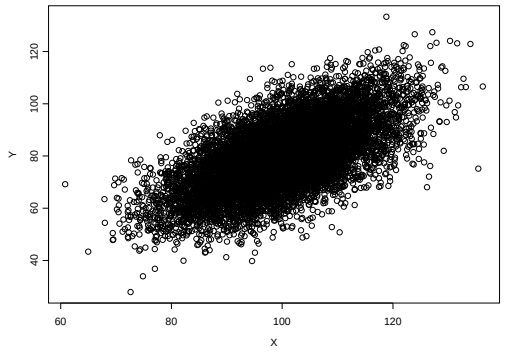

Figure 1. *Simulation Study Design: Simple Random Sampling*

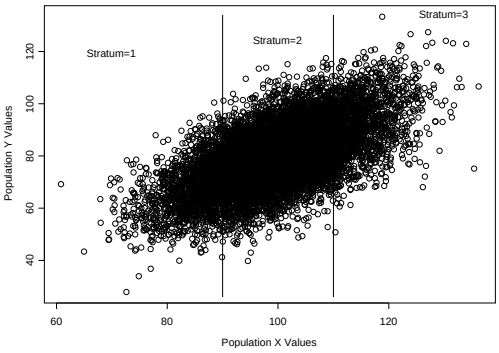

Figure 3. *Simulation Study Design: Stratified Random Sampling*

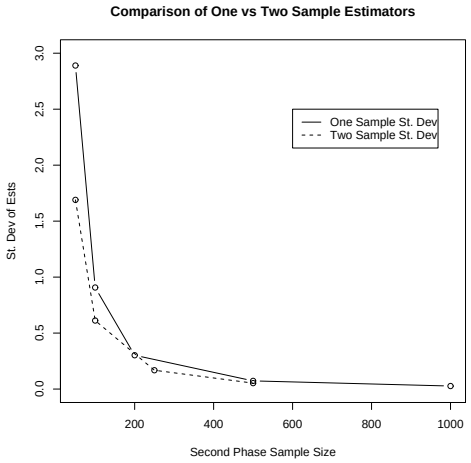

Figure 2. *Simple Random Sampling*

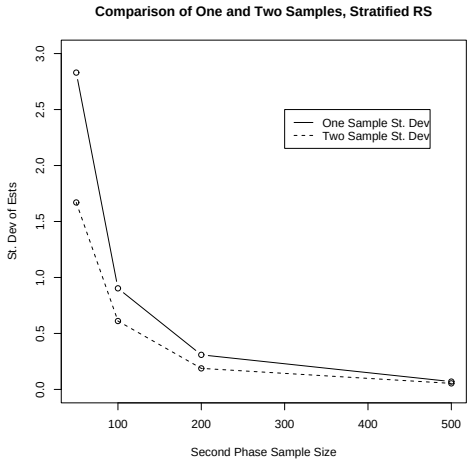

Figure 4. *Stratified Random Sampling*

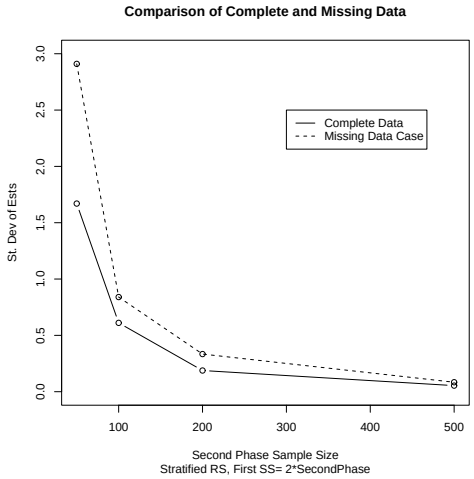

Figure 5. *Stratified Random Sampling*

variance obtained by increasing the second phase sample size, and how the difference in variance between the one and two sample approaches decreases as the overall study size increases. Figure 4 shows similar behavior for the Jackknife variance estimator for stratified sampling. Both of these figures use a first-phase sample size of 2 times the second phase size. Figure 5 illustrates the increase in variance caused by nonresponse. We note that the increase in variance of our Jackknife variance estimates is only substantial for very small samples, but shows that survey practitioners can gain protection from this instability by using a sufficiently large second phase sample size.

4. Conclusions

Figure 2, 4, and 5 show that, in case of small population size, the Jackknife variance estimator for two samples has less standard deviation (variation) than Jackknife variance estimator for one sample. In the future, we intend to study the Jackknife variance estimator for two samples in Stratified Multistage Sampling.

References

Arvesen, J.N. (1969). Jackknifing U-statistics. *Ann. Math. Statist.*, **40**, 2076-2100.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, **91**, 434, 499-506.

Rao, J. N. K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, **79**, 4, 811-822.

Rao, J. N. K., and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 2, 453-460.

Royall, R.M. and Eberhardt, K. R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, C, **37**,1, 43-52.

S*ä*rndal, C.E. (1992). Methods for Estimating the Precision of Survey Estimates when Imputation has been used. *Journal of the American Statistical Association*, **81**, 366-374.

Tukey, J. (1958). Bias and Confidence in not quite large samples *Ann. Math. Statist.*, **29**, 614.