A Discussion of Three Papers on Variance Estimation

Phillip S. Kott
National Agricultural Statistical Service

Abstract

This note expands upon my comments at the 2005 FCSM Research Conference on the papers presented by Lenka Mach, Katherine (Jenny) Thompson, and Laura Ozcoshun.

1. Introduction

The three papers under discussion address variance estimation using some form (or forms) of replication. Lenka Mach discusses the bootstrap, Jenny Thompson discusses two versions of the jackknife, and Laura Ozcoshun discusses **b**alanced **r**epeated **r**eplication or BRR. All three focus on empirical analyses. By contrast, I will concentrate on what theory has to say about the methods the presenters employ. I will try to use the notation in the three papers when discussing each.

The estimator for which one needs a measure of precision is often complicated. It may be a nonlinear function of totals or incorporate complex adjustments for nonresponse (or undercoverage). Replication is a useful variance-estimation technique in such situations. In addition, when users with access to the survey data are not government statisticians, replication can be helpful in protecting respondent confidentiality (as alluded to by Mach) or in providing users with a simple way to measure the precision of complicated estimators they devise.

Let me state some well-known theoretical results. If finite population correction (fpc) can be ignored, then the bootstrap, BRR, and stratified jackknife variance estimators for an expansion estimator are all *exactly* unbiased. By contrast, the delete-a-group (dag) jackknife need not be. Furthermore, if fpc can be ignored, and the estimator is a smooth but *non*linear combination of expansion estimators, then (under mild conditions) all the replication variance estimators are only *nearly* unbiased.

2. A Few Comments on Mach

The analysis in this paper is complicated by different randomizations:

One for the original sample, which has two stages.
One for the bootstrap samples.
One for the simulations.

As Mach herself notes, the sample sizes in her empirical work are too small to draw firm conclusions.

Mach points out that for a linear estimator t of θ under a two-stage sample,

$E_B[(v_{BS}(t)] = v_{WR}(t)$,

where $v_{WR}(t)$ denotes the standard with-replacement-in-the-first-stage variance estimator for Var(t). *It is not necessary for the actual first-stage (school) sample to have been drawn with replacement for this equality to always hold true.*

Let t be an unbiased linear estimator under an unstratified two-stage sample using *without*-replacement sampling in the first stage of sampling. Then

$$E[v_{WR}(t)] = E_1E_2\{ \sum_{i\in F}(t_i/\pi_i)^2 - \sum_{i\neq j}(t_i/\pi_i)(t_j/\pi_j)/(n-1)\}$$

$$= E_1\{ \sum_{i\in F}(\theta_i/\pi_i)^2 - \sum_{i\neq j}(\theta_i/\pi_i)(\theta_j/\pi_j)/(n-1)\} \quad + \quad E_1\{ \sum_{i\in F} Var_2[t_i]/\pi_i^2)\},$$

where t_i is an estimator for θ_i, F is the first-stage sample of schools, and π_i is the selection probability for school i. Note that $E_1\{ \sum_{i\in F} Var_2[t_i]/\pi_i^2)\}$ is the first stage expectation of the second-stage variance. It is captured exactly by the with-replacement variance estimator.

Mach attempts to adjust for the finite population correction with $1 - n/N$, where n/N is the first-stage sampling fraction. I would have used

$$1 - \sum_S p_k^2/m$$

instead, where S denotes the second-stage sample, p_k the combined selection probability for student k ($[n/N][m_i/M_i]$), and m the number of students in the sample. This factor comes from assuming the variable of interest behaves like an independent and identically distributed random variable across students. In particular, it ignores the possibility that the student attitudes to the question asked are correlated within a school. For the opposite extreme, where all the students in a school are assumed to have identical attitudes but that attitude is independent across schools, a reasonable fpc factor would be $1 - n \sum M_i^2/M^2$, where the summations are over all the schools in the population.

3. A Few Comments on Thompson

I will concentrate mostly on the dag jackknife since I coined the term. Consider a stratified simple random sample. The standard sampling weight for $i \in S_h$ is

$$w_i = N_h/n_h.$$

For the dag jackknife, the sample is first systematically divided into K groups so that the numbers of units in each group from stratum h are as close to equal as possible. Let $S_{h(k)}$ denote the set of units in stratum h and NOT in group k. I have proposed determining jackknife replicate weights for $i \in S_{h(k)}$ like so:

$$w_{i(k)}^{Kott} = N_h/n_{h(k)},$$

where $n_{h(k)}$ is the size of $S_{h(k)}$. By contrast, Thompson uses

$$w_{i(k)}^{Thompson} = (K/[K-1])\, N_h/n_h.$$

We both set $w_{i(k)} = 0$ for $i \in S_{hk}$, where S_{hk} is the set of sampled unit in stratum h and group k, and then compute

$$v_{dag}(\sum_{i\in S} w_iy_i) = \sum_{k=1}^{K} ([K-1)/K] \,(\sum_{i\in S} w_{i(k)}y_i - \sum_{i\in S} w_iy_i)^2$$

To see how these approaches differ, we investigate the properties of the different approaches under the simple prediction model for $i \in U_h$: $y_i \sim (\mu_h, \sigma_h^2)$, where the y_i are independent across the i. A purely randomization analogue to the subsequent analysis can be developed, but not as easily.

The model variance of $t = \sum_S w_iy_i$ is $\sum^H (N_h^2/n_h)\sigma_h^2$, and

$$E_M[v_{dag}^{Kott}] = Var_M(t) + \{\text{function of the } \sigma_h \text{ for those h where } N_h/K \text{ is not an integer}\}.$$

The bias has a slight tendency to be upward and is trivial when $n_h > 5$. By contrast,

$$E_M[v_{dag}^{Thompson}] = Var_M(t) + \{\text{function of the } \mu_h \text{ for those h where } N_h/K \text{ is not an integer}\}^2.$$

This bias is upward and persists even when the σ_h are zero. Nevertheless, $v_{dag}^{Thompson}$ may not be so bad for Thompson's application because σ_h can be large compared to μ_h when modeling annual capital expenditures.

To keep things simple, we next consider a count adjustment for nonreponse. For $i \in S_h$, one has the weight adjustment

$$w_i = (N_h / n_h)(n_h / r_h) = N_h / r_h ,$$

where r_h is the number of respondents in h. I would compute the replicate weight

$$w_{i(k)}^{Kott} = (N_h / n_{h(k)})(n_{h(k)} / r_{h(k)}) = N_h / r_{h(k)}$$

for $i \in S_{h(k)}$. Thompson computes either the *short-cut* version:

$$w_{i(k)}^{Shortcut} = (K/[K-1])\, N_h / r_h ,$$

which is akin to $w_{i(k)}^{Thompson}$ and so may not be bad, and a *fully-replicated* version:

$$w_{i(k)}^{"Fully"} = (K/[K-1])(N_h / n_h)(n_{h(k)} / r_{h(k)}),$$

which to me is neither fish nor fowl and cannot be recommended.

The theory behind the ratio adjustment for nonresponse used in Thompson's paper is more difficult, but also unnecessary. If ratioing to payroll makes sense in the context of an ACES stratum, then it should be done using the population total rather than the sample total. Furthermore, given the weak relationship between the dependent variables (different types of capital expenditure) and the auxiliary (payroll), the Census Bureau should consider using a separate-*regression* rather than a separate-ratio estimator with ACES-type data. Kott (2003) describes a method that usually ensures the implicit sample weights in a separate regression estimator will all be positive.

Finally, if there is really a need to incorporate an fpc factor, then a version of the stratified jackknife (with a separate-ratio or regression estimator) should be used rather than the dag jackknife, since the dag can not easily be adjusted for high sampling fractions in a theoretically defensible manner. Moreover, the added error due to the response/nonresponse mechanism is captured by employing $(1 - r_h / N_n)$ rather than $(1 - n_h / N_n)$. Notice that this factor can also be used in certainty strata that exhibit nonresponse.

4. **A Few Comments on Ozcoshun**

In what follows, I will ignore unit-on-property adjustment and small-cell problems. The latter obviates the need for modified half sampling. Modified half sampling is a reasonable thing to do in certain contexts, but is beyond the scope of my remarks.

The "nonresponse adjustment" Ozcoshun describes leads to nearly unbiased estimators under the *quasi-random response model*, where every sampled unit in a cell is equally likely to respond. That is the type of model invoked in the Thompson's paper.

The "second-stage ratio adjustment" leads to a nearly unbiased estimator under the *prediction model*, where the variable of interest (y_i) is *i i d* within each cell whether or not they respond (the response/nonresponse mechanism is *ignorable*). Unlike with the response model, a separate prediction model is needed for each variable of interest.

If the quasi-random response model is correct, but the prediction model is not, the second-stage ratio adjustment can still decrease mean squared errors. Ozcoshun appears to think that the cell definitions for the two adjustment must coincide. That is not the case.

Only one of the models need hold for near unbiasedness is some sense. If the prediction model holds, then partial replication is reasonable. If the prediction model fails, but the quasi-random model holds, then only full replication is reasonable. If the prediction model holds, there may be an *upward* bias in the short-cut method. Similarly, for the

response model.

The added variability in the replicate weights from a replicated ratio adjustment can create an *upward* bias. The smaller the cells, the greater the potential for bias. This is not a problem with the shortcut method, which gets the weights *right*, but the residuals *wrong*.

The last several observations parallel findings Thompson's and my versions of the dag jackknife. Thompson's version of the dag is like Ozcoshun's shortcut version of BRR, while my version is like her version of full replication. This is because my version of the dag jackknife weights treats the sampling strata like model groups but Thompson's does not.

5. Concluding Remarks

Replication does not work by magic. It works because there is asymptotic theory supporting its use. I have simplified the analysis by stripping away some of the complexity from the applications and by assuming simple models. My results nonetheless provide useful insights into what replication does in certain contexts. Unfortunately, we live in a finite world, where asymptotic theory is not always the last word. That is why we need empirical studies like those described in these three useful papers.

Additional Reference

Kott, P.S. (2003). A Practical Use for Instrumental-Variable Calibration, *Journal of Official Statistics* **19**, 265-272.