Comparison of The Random Group Variance and The Stratified Variance Estimators

Stephen M. Miller, Kenneth Robertson, and Shail Butani; Bureau of Labor Statistics
2 Mass. Ave.N.E., Postal Square Building Suite 4985, Washington DC 20212

1. Introduction

The office of Employment and Unemployment Statistics of the Bureau of Labor Statistics (BLS) frequently conducts quick response establishment surveys. These surveys are special one time surveys mandated by the Congress. Quick response surveys are normally National in scope, with a sample size of 7,500 to 10,000 establishments. Stratification for these surveys is by industry and size class. Quick response surveys are designed to measure characteristics such as the number of employees providing child care facilities, the number of employees having a drug testing program, and the number of employers who have job openings for specific occupations.

The characteristics estimated by quick response surveys are, in may cases, rare. Because of its ease of use the random group estimator is commonly employed. It is felt, from years of empirical experience, that for rare characteristics ($\rho \le 0.05$) the random group variance estimator overestimates the variance especially when the relative standard error of the estimate is large. The result of limiting a characteristic to one or two groups is a p-value which is high in these random groups and zero in the remaining random groups. In order to determine if the random group variance estimator is overestimating the variance, we decided to compare the random group variance estimator with the standard stratified variance estimator. This paper reviews the research done to compare these two variance estimators.

Section 2 details the theoretical research; Section 3 provides a description of the data and statistics used for our empirical investigation; Section 4 gives the empirical results, and Section 5 gives our conclusions and future research plans.

2. Theoretical Research

In this section we present a theoretical background for our investigation of the random group variance estimator and the standard stratified variance estimator. We will assume the following stratified sampling setup. Let there be L strata each indexed by h such that $h = 1,\ldots,L$. Within each strata let the population values be indexed as X_{hi} for $i = 1,\ldots,N_h$. We will let $N = \sum_{h=1}^{L} N_h$ denote the total population size across all strata. Within each strata let a simple random sample with replacement (SRSWR) be selected, where the sample units are arranged by random groups $k = 1,\ldots,K$ with m_h units falling in the $k-th$ random group in stratum h. Notice that for convenience we are assuming that the same number of sample units are allocated to each random group within a given strata. We also assume that the sample units are allocated to the groups at random. We are interested in estimating the total T by the estimator $\hat{T}$ as follows,

$$(1)\quad \hat{T} = \sum_{h=1}^{L}\sum_{k=1}^{K}\sum_{i=1}^{m_h} w_h x_{hki} \quad ; \quad T = \sum_{h=1}^{L}\sum_{i=1}^{N_h} X_{hi}$$

where $w_h = n_h^{-1} N_h$ is the sampling weight. We are interested in estimating the variance of $\hat{T}$ by the method of random groups, and also by the usual stratified variance estimator. In order to define these two estimators with a common notation, it is convenient to define

$$(2)\quad d_{hki} = w_h x_{hki} \ .$$

In the case of simple random sampling without replacement (SRSWOR) we can define $d^{*}_{hki} = \sqrt{1-f_h}\, w_h x_{hki}$ where $f_h = N_h^{-1} n_h$ is the sampling fraction. The random group variance estimator v_{RG}, and the usual stratified variance estimator v_{St} can be written as

$$(3)\quad v_{RG} = (K-1)^{-1} K \sum_{k=1}^{K} (d_{+k+} - d_{+\bullet+})^2$$

$$v_{St} = \sum_{h=1}^{L} (n_h - 1)^{-1} n_h \sum_{k=1}^{K}\sum_{i=1}^{m_h} (d_{hki} - d_{h\bullet\bullet})^2$$

where $n_h = K m_h$, and

$$(4)\quad d_{+k+} = \sum_{h=1}^{L}\sum_{i=1}^{m_h} d_{hki} \qquad d_{+\bullet+} = K^{-1}\sum_{k=1}^{K} d_{+k+}$$

$$d_{h\bullet\bullet} = n_h^{-1}\sum_{k=1}^{K}\sum_{i=1}^{m_h} d_{hki} \ .$$

We are interested in the joint behavior of the two variance estimators. We begin by defining a few population values. Let

$$(5)\quad \sigma_h^2 = N_h^{-1}\sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)^2$$

$$\overline{X}_h = N_h^{-1}\sum_{i=1}^{N_h} X_{hi}$$

$$\beta_h = \sigma_h^{-4} N_h^{-1}\sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)^4$$

Notice that β_h is the population kurtosis within stratum h

Under the SRSWR sampling scheme, each d_{hki} is distributed independently with mean $w_h \overline{X}_h$, variance $w_h^2 \sigma_h^2$, and kurtosis β_h. In addition, the d_{hki} are

identically distributed within each stratum. Under the SRSWR sampling scheme it follows that

$$(6) \qquad V\{\hat{T}\} = \sum_{h=1}^{L} n_h w_h^2 \sigma_h^2$$

We begin by presenting Result 1, which gives expressions for the means, as well as variances and covariances for the two variance estimators.

Result 1. Under SRSWR,

$$E\{v_{St}\} = E\{v_{RG}\} = V\{\hat{T}\}$$

$$V\{v_{St}\} = \sum_{h=1}^{L} n_h w_h^4 \sigma_h^4 \left[(\beta_h - 3) + \frac{2n_h}{n_h - 1} \right]$$

$$V\{v_{RG}\} = \sum_{h=1}^{L} n_h w_h^4 \sigma_h^4 (\beta_h - 3) + 2(K-1)^{-1} \left(\sum_{h=1}^{L} n_h w_h^2 \sigma_h^2 \right)^2$$

$$C\{v_{St}, v_{RG}\} = V\{v_{St}\}$$

$$Corr\{v_{St}, v_{RG}\} = \left[\frac{\sum_{h=1}^{L} n_h w_h^4 \sigma_h^4 \left[(\beta_h - 3) + 2n_h (n_h - 1)^{-1} \right]}{\sum_{h=1}^{L} n_h w_h^4 \sigma_h^4 (\beta_h - 3) + 2(K-1)^{-1} \left(\sum_{h=1}^{L} n_h w_h^2 \sigma_h^2 \right)^2} \right]^{\frac{1}{2}}$$

Proof. See Appendix.

The expressions for the means and variances can be found in other sources such as Wolter, but the result on covariance and correlation between the two estimators appears to be new. It is interesting to note that the covariance between the two estimators is the same as the variance of v_{st} . This fact will be exploited in Result 2 below. The correlation expression of Result 1 has some interesting interpretations. We begin by presenting a Table below which gives kurtosis values for three distributions.

Table 1. Kurtosis Values for Various Distributions (Wolter, 1985)

Distribution	Kurtosis β	Parameter
Normal	3	
Bernoulli	$\frac{3+[1-6p(1-p)]}{p(1-p)}$	ρ: Success
Poisson	$3+\lambda^{-1}$	λ:Expected Value

For a normally distributed population, the kurtosis is 3 which causes the first term in both the numerator and denominator of the correlation expression to disappear. In this case if the sample size is large then the correlation will be small and approach zero asymptotically. The situation is slightly different with the other two distributions. For both the Bernoulli and Poisson distributions the first term does not disappear from the numerator and denominator. In fact, if the populations are rare (p close to 0 for the Bernoulli, and λ close to zero for the Poisson) then the kurtosis is large and the first term in both the numerator and denominator can dominate the correlation expression which causes it to approach 1. Therefore, the two variance estimators should be more highly correlated when the populations are rare.

In the next result we express the random group variance estimator in terms of the usual stratified variance estimator plus an error.

Result 2. Under SRSWR, we can write

$$v_{RG} = v_{St} + error$$

where

$$E\{error\} = 0, \qquad C\{v_{St}, error\} = 0$$

$$V\{error\} = 2(K-1)^{-1} \left(\sum_{h=1}^{L} n_h w_h^2 \sigma_h^2 \right)^2 - 2 \sum_{h=1}^{L} n_h^2 w_h^4 \sigma_h^4 (n_h - 1)^{-1}$$

Proof. We can write $error = v_{RG} - v_{St}$, and Result 2 follows from Result 1.

Result 2 can be thought of as a *regression like result.* It says that if we were to regress the random group estimator on the usual stratified variance estimator that we would get an intercept of 0 and slope coefficient of 1. In addition the regression error is heteroscedastic, with a variance which increases with the variance of the estimated total.

In the next result we examine the approximate expectation of the ratio of the two variance estimators.

Result 3. If we approximate the ratio of variance estimates as

$$\frac{v_{RG}}{v_{St}} = 1 + \frac{1}{V\{\hat{T}\}} \left(v_{RG} - V\{\hat{T}\} \right) - \frac{1}{V\{\hat{T}\}} \left(v_{St} - V\{\hat{T}\} \right) - \frac{1}{V^2\{\hat{T}\}} \left(v_{RG} - V\{\hat{T}\} \right) \left(v_{St} - V\{\hat{T}\} \right) + \frac{1}{V^2\{\hat{T}\}} \left(v_{St} - V\{\hat{T}\} \right)^2 + remainder.$$

Then the expectation of the approximation, ignoring the remainder, is $E\left\{ \frac{v_{RG}}{v_{St}} \right\} \approx 1$

Proof. The result follows from applying Result 1.

The interesting thing about this result is the expectation of the second order terms in the approximation is zero. In Result 4 below, we examine the approximate variance of the ratio of the variances.

Result 4. If we approximate the ratio of variance estimates as

$$\frac{v_{RG}}{v_{St}} = 1 + \frac{1}{V\{\hat{T}\}}\left(v_{RG} - V\{\hat{T}\}\right) - \frac{1}{V\{\hat{T}\}}\left(v_{St} - V\{\hat{T}\}\right) + remainder$$

Then the variance of the approximation, ignoring the remainder, is

$$V\left\{\frac{v_{RG}}{v_{St}}\right\} \approx \frac{2}{K-1} - \frac{2\sum_{h=1}^{L} n_h^2 w_h^4 \sigma_h^4 (n_h - 1)^{-1}}{\left(\sum_{h=1}^{L} n_h w_h^2 \sigma_h^2\right)^2}$$

Proof. The result follows by applying Result 1.

We will have more to say about Result 4 after we present Result 5. In Result 5 below, we present some asymptotic theory about the behavior of the variance ratio.

Result 5. Assume an infinite sequence of sample designs and populations such that,

$$\frac{d_{+k+} - K^{-1}\mathrm{T}}{\sqrt{K^{-1}V\{\hat{\mathrm{T}}\}}} \xrightarrow{L} N(0,1), \text{ and } v_{St} - V\{\hat{\mathrm{T}}\} \xrightarrow{P} 0.$$

Then, $\frac{v_{RG}}{v_{St}} \xrightarrow{L} F_{\infty}^{K-1}$

where F_{∞}^{K-1} denotes an F random variable with numerator degrees-of-freedom $K-1$ and denominator degrees-of-freedom ∞.

Proof. Let

$$R_k = \frac{d_{+k+} - K^{-1}\mathrm{T}}{\sqrt{K^{-1}V\{\hat{\mathrm{T}}\}}}$$

then $v_{RG} = V\{\hat{\mathrm{T}}\}(K-1)^{-1}\sum_{k=1}^{K}(R_k - \bar{R})^2$.

By the first assumption,

$$\sum_{k=1}^{K}\left(R_k - \bar{R}\right)^2 \xrightarrow{L} \chi^2_{K-1},$$

so the result follows from the second assumption by Slutsky's theorem.

The assumptions of Result 5 are mild, and should be satisfied by most designs used in practice. Result 5 implies that under the asymptotic theory,

(7) $P\{v_{RG} \le v_{St}\} \to P\{\chi^2_{K-1} \le K-1\}$.

In Table 2 below, we give some of the probability values for various numbers of random groups.

Table 2. Asymptotic Values for Eq. (7)

Random Groups	$P\{v_{RG} \le v_{St}\}$
5	0.5940
6	0.5841
7	0.5768
8	0.5711
9	0.5665
10	0.5627
11	0.5595
12	0.5567
13	0.5543

Asymptotically, we would expect the random group variance estimator to be less than the usual stratified variance estimator on average, despite the fact that they are both unbiased. Result 5 also implies that

(8) $v_{RG} - v_{St} \dot{\sim} V\{\hat{T}\}(F_{\infty}^{K-1} - 1)$,

where $\dot{\sim}$ *stands for approximately distributed as* . That is, if we examine the difference between the two variance estimator, then asymptotically the distribution of the difference has a skewed distribution which is a multiple of a translated F-distribution. This relates back to Result 2 which gave expressions for the *error* , Result 5 tells us about the asymptotic behavior of the distribution of *error.*

Finally, notice that an F random variable has mean and variance given by

(9) $E\{F_{\infty}^{K-1}\} = 1 \qquad V\{F_{\infty}^{K-1}\} = 2(K-1)^{-1}$

We can compare these values to those of Result 3 and Result 4. Note that our approximate expectation and the mean of the asymptotic distribution are both 1, but expressions for the variances differ slightly. In particular, the approximate variance is smaller than that of the variance of the asymptotic distribution. Later we will examine which variance expression holds for our empirical data, and we will say something about how well the asymptotic theory works with rare populations.

3. ETJO Data

At the time we began this investigation, we had recently completed a special survey, a pilot study on employee Turnover and Job Openings (ETJO). The ETJO survey was conducted in response to a Congressional mandate to the Department of Labor to study the feasibility of measuring national labor shortage data and the associated cost. .

The ETJO survey employed a stratified (8 industries × 3 employment size classes) probability sampling plan. These 24 strata are called ETJO sampling cells. Because occupations vary by industry, a separate questionnaire was required for each industry. Therefore, the scope of the survey was limited to eight industries. The sample consisted of approximately 3300 units selected from a universe of about 1.5 million establishments. The universe covered private establishments with one or more employees during the first quarter of 1989 in the 50 States and the District of Columbia. Prior to sample selection, the total national frame was sorted within each ETJO cell

by 4-digit SIC/ Employment/ and State. This sort permitted further stratification in the sample selection process. Once this sort was completed, a random systematic start sample was selected within each ETJO sampling cell. The total sample was divided into three monthly panels. Each panel was surveyed 1 month, out of sample 2 months, and then back in the survey 1 month (ie. 1−2−1). Random Group assignments were made by systematic assignment after a random start. The reference data collection period for the survey was November 1990 to April 1991.

The data were collected by mail. There was one follow up mailing, after which nonrespondents were contacted using a Computer Assisted Telephone Interview (CATI) instrument. The overall usable response rate for the first and second collection periods, respectively, was 70 and 75 percent. Five data items were collected by detailed occupation: job separations, new hires, average wage of new hires, job openings, and the duration of job openings.

Estimates were produced for each of the five characteristics by occupation for each of the eight industries surveyed. Additionally, estimates were produced by occupation at the National level, or by ignoring industry and grouping by occupation only. The ETJO survey variance estimates were produced using a Random Group Variance (RGV) estimator. This estimator was calculated as follows

$$V_{RG}(\hat{X}_{jd}) = (K-1)^{-1} K \sum_{k=1}^{K} \left(X_k - \overline{X}_\bullet\right)^2$$

where $\hat{X}_{jd}$ is the estimate for characteristic X in industry j for occupation d. In addition K refers to the number of random groups, and X_k is the estimate based upon the k-th random group where, $X_k = \sum_{l=1}^{6}\sum_{c=1}^{3}\sum_{i} W_{ijclk} NX_{ijcdlk}$, and $\overline{X}_\bullet = K^{-1}\sum_{k=1}^{K} X_k$ where l is the month, c is the size class, i is the establishment, and NX_{ijcdlk} = the number reported for characteristic X by unit i for occupation d.

$$W_{ijclk} = \frac{1}{6} FW_{ijclk}\sqrt{1-f_{ijck}}\left(\frac{3}{1000}\right)$$

where FW_{ijclk} is the sampling weight, which is the inverse of the probability of selection adjusted for nonresponse and changes in the frame, and $(1-f_{ijck})$ is the finite population correction factor. Notice that a divisor of 6 was used to obtain an average over six months of data, and multiplied by (3/1000) to produce an estimate of monthly total in units of one thousand.

The standard stratified variance estimator was computed as follows

$$V_{St}(\hat{X}_{jd}) = \sum_{c=1}^{3}\left(\frac{n_{jc}}{n_{jc}-1}\right)(1-f_{jc})\sum_{i=1}^{n_{jc}}\left(X_{ijc} - \overline{X}_{jc}\right)^2$$

where n_{jc} is the number of establishments in industry j and employment size class c, and

$$\overline{X}_{jc=n_{jc}^{-1}} \sum_{i=1}^{n_{jc}} X_{ijc}\ ,\ X_{ijc} = W_{ijc} NX_{ijc}\ ,\ W_{ijc} = \frac{1}{6} FW_{ijc}\left(\frac{3}{1000}\right).$$

In our investigation we decided to use three of the characteristics from the survey. These characteristics were Separations, Job Openings, and New Hires. These statistics were produced for each SIC/Occupation group. Additionally, these statistics were produced using 6, 9, and 12 random groups. This gave us a total of 81 sets of statistics to review, with each set containing a varying number of statistics according to the number of occupations in the industry. In addition, the computer package PC CARP was used to compute the standard stratified variance estimates. The statistics for each set were grouped by size to produce frequency distributions. Since most of these differences were small, we decided to use the interval from zero to three in units of 0.2.

4. Empirical Results

In this section we describe our empirical research using the ETJO survey data.

We begin by presenting plots of variance ratios. The chart labelled "Theoretical vs. Empirical Distribution" was constructed using Separations data for 6, 9, and 12 random groups. This chart shows three histograms of ratios of variance estimates of total separations for the 8 SICs and various occupations. The variance ratios were defined to be v_{RG}/v_{st}. The histograms were normalized by the bin width and total number of observations, so that they are an estimate of the density of the variance ratio. We also plotted the density for the associated F random variable. This is the density of the asymptotic distribution implied by Result 5. It is interesting to note that the empirical histogram density estimator appears to be shifted more to the right than the theoretical density estimator. This indicates that the random group estimator tends to be larger relative to the standard stratified estimator than the theory would predict. Theoretically we would expect that both variance estimators would be unbiased, that the random group variance estimator would be more variable, and that, on average, the random group variance estimator would be smaller than the standard stratified variance estimator. In fact, the theory says that

$P\{v_{RG} \le v_{St}\} = P\{\chi_8^2 \le 8\} \doteq 0.567$.

The empirical histogram estimator shows that for 9 random groups 42 % of the distribution falls to the left of 1.0 which indicates that the distribution has moved substantially to the right.

The theoretical results can be made to more closely match the empirical results by allowing the noncentrality parameters to be greater than zero. In general, the

noncentrality parameters cause the distributions to move to the right, and pack more tightly around 1.0.

5. Conclusions

Our hypothesis at the beginning of this paper was that the random group variance estimator tends to overestimate the variance when the value to be estimated is a rare characteristic in the surveyed population. We have found empirical evidence that the random group variance estimator tends to overestimate the variance relative to the standard stratified variance estimator. We speculated that this could be caused by the sorting and systematic selection of sample units, as well as the systematic allotment of random group assignments. We performed small scale simulations using the sample as the sampling frame. Whether we used systematic sampling or random sampling in these simulations we still observed the random group variance estimator overestimating the variance relative to the standard stratified variance estimator.

Theoretical results show that for a normal population $E(V_{RG}) = E(V_{St})$. Our research also specified an *F* distribution for the ratio of these two estimators. This *F* distribution, for a normal population, shows that $P(V_{RG} < V_{St}) > 1/2$. It was also shown that for a population with large kurtosis, which can be caused by rare characteristics, that this distribution can change shape rather dramatically. The theoretical section showed that for a normal population the kurtosis has no affect on the correlation between these two variance estimators. However, as the population begins to depart from normality (or as the kurtosis becomes larger than 3) the kurtosis does influence the correlation between the estimators. When the kurtosis of the population is large this figure becomes the dominant one in the correlation equation and the correlation between these variance estimators begins to approach 1.

Our empirical findings revealed several interesting things. First, as expected, when the number of random groups is increased the variance of the ratio V_{RG}/V_{St} decreases. This reduction in the variance of the ratio causes a larger percentage of the distribution to be close to 1. This result is as described by previously established theory.

Second, as the number of random groups is increased, it appears that the portion of the distribution most heavily affected was those observations where $V_{RG}/V_{St} < 1$. That is, the observations less than 1 move towards 1 more rapidly than those observations greater than 1 as the number of random groups is increased. In cases where the characteristic to be measured is rare the random group variance estimator underestimates the variance less often than theory predicts. The observations where $V_{RG}/V_{St} > 1$ also move towards 1 as the number of random groups is increased. However, for the numbers of random groups tested this part of the distribution conforms more closely to the theoretical distributions than the part of the distribution less than 1.

Third, the empirical results show that the random group variance estimator does tend to overestimate the variance, in cases where the characteristic to be measured is rare, more often than the theory would suggest. This overestimation tends to manifest itself by clustering around the point where the ratio of $V_{RG}/V_{St} = 1$. As the number of random groups is increased, this clustering becomes more apparent. It is also apparent that most of these observations clustering around 1 are coming from the part of the distribution which theory suggests should be < 1. This means that in many cases, when estimating rare characteristics, the random group variance estimator is a better estimator than the theory would suggest for two reasons. First, in general practice we would rather have a conservative variance estimator than one that continually underestimates the variance. It appears that the random group variance estimator is conservative when the population characteristic to be estimated is rare. Second, the estimated variance is approximately equal to the 'true' variance much more often than theory would suggest.

Future research plans will include, but not be limited to, an investigation of this nature into the properties of the Jackknife variance estimator.

References

Fuller, W.A., Kennedy, W.J., Schnell, D., Sullivan, G., and Park, H.J. (1986). *PC CARP*. Statistical Laboratory, Iowa State University, Ames, Iowa.

Koop, J.C. (1971). On Splitting a Systematic Sample for Variance Estimation. *Annals of Mathematical Statistics* **42** 1084-1087.

Krewski, Daniel (1978). On the Stability of Some Replication Variance Estimators in the Linear Case. *Journal of Statistical Planning and Inference* **2** 45-51.

Valliant, Richard (1987). Some Prediction Properties of Balanced Half-sample Variance Estimators in Single-stage Sampling. *Journal of the Royal Statistical Society Ser. B* **49** 68-81.

Vallaint, Richard (1990). Comparison of Variance Estimators in Stratified Random and Systematic Sampling. *Journal of Official Statistics* **6** 115-131.

Wolter, Kirk (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* **79** 781-790.

Wolter, Kirk (1985). *Introduction to Variance Estimation*. Springer-Verlag New York.

Wu, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika* **78** 181-188.

Theoretical vs. Empirical Distribution