VARIANCE ESTIMATION FOR THE CURRENT EMPLOYMENT STATISTICS PROGRAM

Kirk Wolter, NORC & University of Chicago, Jun Shao, University of Wisconsin,
Larry Huff, Bureau of Labor Statistics
Kirk Wolter, NORC & University of Chicago, 55 E. Monroe, Chicago, IL 60603

1. INTRODUCTION

The Current Employment Statistics (CES) program supplies some of America's most vital leading-economic indicators. At the 1997 annual meeting in Anaheim, various authors reported on the status of an ongoing redesign of this important survey. Their papers discussed the rationale for the redesign, the new sampling design, the new estimation procedures for employment levels and trends, and special estimation issues for hours and earnings variables. In the current paper, we discuss variance estimation for all CES variables.

The overall variance estimation strategy involves primary use of the balanced half- samples (BHS) method for higher-level, aggregate statistics, and secondary use of generalized variance functions (GVF) for the more disaggregated statistics. Our BHS method addresses many CES design features, such as stratification, more than 2 primary units per stratum, clustering, birth and death sampling, and rotation sampling. We incorporate allowances for the imputation variance and for the finite population correction. To improve variance estimation for nonlinear statistics, we employ both the half sample estimator and its complement.

2. REVIEW OF THE METHOD OF BHS

In this section, we review the basics of the BHS method and describe a number of extensions that provide flexibility to handle a wide spectrum of sampling designs.

2.1 Basic Method

Suppose it is desired to estimate a population total, Y, from a statistical sampling design with two selected units per stratum, where the selected units comprise a simple random sample with replacement (srs wr). Let L denote the number of strata, N_h the number of units within the h-th stratum, N the size of the entire population, and n_h the sample size within the h-th stratum. Assuming complete response, the unbiased estimator of the population total is given by

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i \in s_h} W_{hi}\, y_{hi} ,$$

where s_h is the sample in stratum h and $W_{hi} = N_h / n_h$ is the base weight attached to the hi-th unit. The standard, unbiased estimator of the sampling variance is

$$v(\hat{Y}) = \sum_{h=1}^{L} \left(W_{h1}\, y_{h1} - W_{h2}\, y_{h2}\right)^2 .$$

The BHS method works in terms of half samples comprised of one unit from each stratum. Define indicator variables $\delta_{h1\alpha}$ and $\delta_{h2\alpha}$ that identify whether the first or second selected unit from the h-th stratum is in the α-th half sample. Then, the unbiased estimator of the population total derived from the α-th half sample

$$\hat{Y}_\alpha = \sum_{h=1}^{L} \sum_{i \in s_h} W_{hi\alpha}\, y_{hi\alpha} ,$$

where the half-sample weight, $W_{hi\alpha} = 2\, W_{hi}\, \delta_{hi\alpha}$, is either twice the full sample weight or zero. A set of $k \geq 2$ half samples are required in order to provide the following unbiased estimator of the sampling variance of $\hat{Y}$:

$$v_k(\hat{Y}) = \frac{1}{k} \sum_{\alpha=1}^{k} \left(\hat{Y}_\alpha - \hat{Y}\right)^2 .$$

It is desirable to employ a balanced set of half samples, where balance is achieved by defining half samples according to the rows of a Hadamard matrix (see Wolter (1985), Appendix A). Each column represents a different stratum and each row a different half sample. The symbol "+" in the αh-th cell of the matrix signifies that the first unit in the h-th stratum is in the α-th half sample, while "-" signifies that the second unit is in the half sample. To achieve *full, orthogonal balance* the order of the Hadamard matrix must exceed the number of sampling strata. To achieve *partial balance* the order of the Hadamard matrix can be smaller than the number of sampling strata, in which case the columns may be assigned to multiple strata. Given full, orthogonal balance, the BHS estimator of variance, v_k, is algebraically identical to the standard unbiased estimator, v.

In many survey applications, the last stratum is a certainty stratum, contributing zero sampling variability. To properly accommodate this situation, we must assign each and every unit in stratum L to each and every half sample. Thus, $\delta_{Li\alpha} \equiv 1$ and $W_{Li\alpha} = W_{Li}$ for $i = 1, \ldots, N_L$ and $\alpha = 1, \ldots, k$. The foregoing discussion of Hadamard matrices and full, orthogonal or partial balance applies only to the noncertainty strata.

In two-stage surveys with two primary sampling units (PSU) selected per stratum, the half samples should be defined based upon whole PSUs. The BHS estimator of variance, v_k, is defined as before, but in terms of the

$$\hat{Y}_\alpha = \sum_{h=1}^{L} \sum_{i \in s_h} \sum_{i \in s_{hi}} W_{hij\alpha}\, y_{hij} ,$$

half-sample estimator where the subscript j signifies the elementary unit selected within the hi-th PSU.

Let $\theta = \theta\,(Y, X, Z, \ldots)$ denote a general population parameter of interest, defined as a function of population totals. In our experience, virtually every parameter of interest in real sample surveys is of this general form. The standard estimator of θ is $\hat{\theta} = \theta(\hat{Y}, \hat{X}, \hat{Z}, \ldots)$ and the BHS estimator of its variance is defined by

$$v_k(\hat{\theta}) = \frac{1}{k} \sum_{\alpha=1}^{k} \left(\hat{\theta}_\alpha - \hat{\theta}\right)^2 ,$$

where the half-sample estimator is $\hat{\theta}_\alpha = \theta(\hat{Y}_\alpha, \hat{X}_\alpha, \hat{Z}_\alpha, \ldots)$.

2.2 Simple Methods for $n_h \geq 2$

We now consider the general case with two or more units selected per stratum. While it is possible in select cases to construct variance estimators based upon fully-orthogonal, fractional samples, such as for n_h a prime or a power of a prime, we believe there is merit in continuing the simple BHS method. There are two basic approaches. First, we may subdivide the units in each stratum into two random groups, and then apply the basic BHS method to the two groups. The subdivision may be done systematically or randomly. The estimated totals, $\hat{Y}$ and $\hat{Y}_\alpha$, and the BHS estimator of variance, v_k, are defined as in Section 2.1.

Second, we may subdivide the h-th stratum into G_h artificial strata, each of sample size 2, for $h = 1, \ldots, L$. Now there are $G = \sum G_h$ artificial strata overall, and the sample size in the h-th real stratum is $n_h = 2\,G_h$. Given this approach, the full-sample estimator and half-sample estimator are defined by

$$\hat{Y} = \sum_{h=1}^{L} \sum_{g=1}^{G_h} \sum_{i=1}^{2} W_{hgi}\, y_{hgi}$$

$$\hat{Y}_\alpha = \sum_{h=1}^{L} \sum_{g=1}^{G_h} \sum_{i=1}^{2} W_{hgi\alpha}\, y_{hgi} ,$$

respectively, where $W_{hgi} = N_h / n_h$, $W_{hgi\alpha} = 2\,W_{hgi}\,\delta_{hgi\alpha}$, $\delta_{hgi\alpha}$ is an indicator variable signifying whether the first or second selected unit from the hg-th artificial stratum is in the α-th half sample. We obtain the BHS estimator of variance, v_k, by applying a fully- or partially-balanced design to the G artificial strata.

Given $n_h \geq 2$ sample units in the h-th stratum, the standard, textbook estimator of variance is based upon n_h - 1 degrees of freedom (df), while the first and second simple, BHS estimators are based upon 1 and G_h df, respectively. On this basis, one might prefer the second simple estimator to the first. On the other hand, the first simple estimator is probably easier to apply in practice, and may be good enough as long as L is large.

2.3 Correction for Without Replacement Sampling

If sampling is without replacement within the h-th stratum, then we may find it desirable to incorporate in the variance calculations a finite population correction factor (fpc), $1 - f_h$, where f_h is the sampling fraction in the stratum. We define the corrected half-sample estimator of the population total by

$$\hat{Y}_\alpha^* = \sum_{h=1}^{L} \sum_{i \in s_h} W_{hi\alpha}^*\, y_{hi} \qquad (1)$$

where the corrected weight for the α-th half sample is $W_{hi\alpha} = \left\{1 + (1 - f_h)^{1/2}\,(2\,\delta_{hi\alpha} - 1)\right\} W_{hi}$. The BHS estimator of variance, v_k^*, is defined as before, but with $\hat{Y}_\alpha^*$ replacing $\hat{Y}_\alpha$.

Assuming full, orthogonal balance, we can show that v_k^* is algebraically equal to the standard estimator of variance for simple random sampling without replacement. Thus, we have successfully incorporated the fpc's, as desired.

2.4 Use of the Half Sample and Its Complement

Thus far, we have been discussing variance estimation based upon k half samples. For nonlinear estimators and for thin samples, we occasionally find it useful to employ both the half samples and their complements, or $2k$ half samples overall.

The *complement* of the α-th half sample is defined by the new indicator variables $\delta^c_{hi\alpha} = 1 - \delta_{hi\alpha}$. The corresponding estimator of the population total is

$$\hat{Y}^c_\alpha = \sum_{h1}^{L} \sum_{i \in s_h} W^c_{hi\alpha}\, y_{hi},$$

where the half-sample weights are now $W^c_{hi\alpha} = 2\, W_{hi}\, \delta^c_{hi\alpha}$.

One possible estimator of variance is a weighted average of the BHS estimator of variance based upon the k original half samples and of the BHS estimator based upon the k complementary half samples, defined by

$$\bar{v}_k(\hat{Y}) = \gamma\, v_k(\hat{Y}) + (1-\gamma)\, v^c_k(\hat{Y}).$$

A second possible estimator of variance is defined by

$$v^\dagger(\hat{Y}) = \frac{1}{\gamma^2 k} \sum_{\alpha=1}^{k} \left(\hat{Y}^\dagger_\alpha - \hat{Y}\right)^2$$

$$= \sum_{\alpha=1}^{k} \left(\hat{Y}_\alpha - \hat{Y}^c_\alpha\right)^2 / 4k,$$

where $\hat{Y}^\dagger_\alpha$ is an estimator based upon a weighted average of the half sample and its complement, defined by

$$\hat{Y}^\dagger_\alpha = \frac{1+\gamma}{2}\hat{Y}_\alpha + \frac{1-\gamma}{2}\hat{Y}^c_\alpha$$

$$= \sum_{h=1}^{L} \sum_{i \in s_h} W^\dagger_{hi\alpha}\, y_{hi}$$

and

$$W^\dagger_{hi\alpha} = W_{hi}\left\{(1+\gamma)\,\delta_{hi\alpha} + (1-\gamma)\left(1-\delta_{hi\alpha}\right)\right\}$$

While the mixing parameter, γ, may be chosen anywhere in the half-open interval (0, 1], we know of reason not to choose the value $\gamma = ½$. The estimators $v^\dagger_k$ and $\bar{v}_k$ are discussed by Wolter (1985) and Fay (1989).

2.5 Combining All the Features of the BHS Method

We now combine all of the features of the BHS method discussed in Section 2. This means we assume possibly multiple-stages of sampling, $n_h \geq 2$ primary sampling units per stratum, sampling without replacement, and use of both the half sample and its complement.

To estimate the population total, the combined, half-sample estimator is now

$$\hat{Y}^+_\alpha = \sum_{h=1}^{L} \sum_{i \in s_h} \sum_{j \in s_{hi}} W^+_{hij}\, y_{hij},$$

where the combined weight is

$$W_{hij} = W_{hij}\,[1 + (1-f_h)^{½}\{(1+\gamma)\delta_{hi\alpha} + (1-\gamma)(1-\delta_{hi\alpha}) - 1\}.$$

And to estimate a general parameter θ, the combined, half-sample estimator is $\hat{\theta}^+_\alpha = \theta\left(\hat{Y}^+_\alpha, \hat{X}^+_\alpha, \ldots\right)$. The BHS estimators of variance are

$$v^+_k(\hat{Y}) = \frac{1}{\gamma^2 k} \sum_{\alpha=1}^{k} \left(\hat{Y}^+_\alpha - \hat{Y}\right)^2$$

and

$$v^+_k(\hat{\theta}) = \frac{1}{\gamma^2 k} \sum_{\alpha=1}^{k} \left(\hat{\theta}^+_\alpha - \hat{\theta}\right)^2.$$

3. IMPUTATION VARIANCE

We now examine modifications to the BHS estimator of variance to accommodate the problem of nonresponse. To begin, we consider simple random sampling without replacement in one stratum, and for notational convenience, we temporarily suppress the stratum subscript. After developing methods for this simple problem, we shall reintroduce the full sampling design and correspondingly extend the methods.

We shall assume that r units respond and m units do not respond, such that $n = r + m$. And we assume a random response mechanism within stratum.

We are interested in estimating the stratum mean, $\bar{Y}$. Given complete response, one may consider the sample mean, $\bar{y}$, as the unbiased estimator of the stratum mean. We shall consider this estimator following imputation for missing values based upon a trend (or ratio) model.

Let x denote a lagged value of the y-variable. Suppose the x-variable is complete (observed) for all units in the sample, even for the m nonrespondents. Then the imputed data set consists of the values

$$\hat{y}_i = y_i \text{, if i responds}$$
$$= x_i \hat{\beta} \text{, if i does not respond,}$$

where $\hat{\beta} = \bar{y}_r / \bar{x}_r$ and $\bar{y}_r$ and $\bar{x}_r$ denote the sample means of the respondents. The estimator of the stratum mean, the sample mean of the imputed data set, is now given by

$$\bar{y}_I = \frac{1}{n} \sum_{i \in s} \hat{y}_i = (\bar{y}_r / \bar{x}_r)\bar{x} ,$$

where $\bar{x}$ denotes the overall sample mean of the x-variable, and s denotes the overall sample.

Using a conventional Taylor-series expansion, we find the sampling variance of $\bar{y}_I$ is given approximately by

$$\text{Var}\{\bar{y}_I\} \doteq \left(\frac{1}{r} - \frac{1}{N}\right) S_d^2 + 2\left(\frac{1}{n} - \frac{1}{N}\right) \beta S_{dx} + \left(\frac{1}{n} - \frac{1}{N}\right) \beta^2 S_x^2, \quad (2)$$

where $\beta = \bar{Y}/\bar{X}$ is the stratum-specific regression coefficient, $d_i = y_i - x_i \beta$ is the residual for the i-th unit, $S_d^2 = (N-1)^{-1} \sum d_i^2$, $S_x^2 = (N-1)^{-1} \sum (X_i - \bar{X})^2$, and $S_{dx} = (N-1)^{-1} \sum d_i x_i$.

Our aim is to employ the BHS approach to estimating the sampling variance. Towards that end, we establish two random groups -- s_1 and s_2, where $s = s_1 \cup s_2$ -- within the stratum and we employ separate imputation of the missing values within each of the groups. Note that separate imputation implies three sets of imputed values: one each for s, s_1 and s_2.

A provisional estimator of the sampling variance is given by

$$v(\bar{y}_I) = (\bar{y}_{I1} - \bar{y}_{I2})^2/4 ,$$

where $\bar{y}_{I1}$ and $\bar{y}_{I2}$ are the sample means of the two imputed data sets. Given the assumed random nonresponse mechanism, and given that the sizes of the two random groups (n_1 and n_2) are equal and that the numbers of respondents in the two random groups (r_1 and r_2) are equal, we find that the expectation of the provisional estimator of variance is

$$E\{v(\bar{y}_I)\} = \frac{1}{r} S_d^2 + 2\frac{1}{n} \beta S_{dx} + \frac{1}{n} \beta^2 S_x^2 . \quad (3)$$

Comparing Equations 2 and 3, we conclude that v is unbiased except for failure to apply the appropriate fpc's.

To correct for the upward bias due to without replacement sampling, we recommend the fpc-corrected estimator, $v^*(\bar{y}_I) = (1 - f)\, v(\bar{y}_I)$, where $f = r/N$ or n/N. The first choice of the sampling fraction f is the more conservative, leaving a residual upward bias. In some applications, the mean square, S_d^2, and cross product, S_{dx}, will be small relative to the mean square, S_x^2, of the lagged variable, in which case, the second sampling fraction may be preferred.

We close this section by returning to the full sampling design with L strata introduced in Section 2. The basic BHS estimator of variance essentially implements the preliminary estimator, v, within each stratum. Thus, in the presence of imputation for nonresponse, the basic BHS estimator of variance is upward biased. Similarly, the BHS estimator corrected for without replacement sampling essentially implements the fpc-corrected estimator, v^*, within each stratum. Thus, in the presence of imputation for nonresponse, we recommend the BHS estimator defined in Equation 1, defining the stratum sampling fractions as $f_h = r_h / N_h$ or n_h / N_h. And for applications that warrant a conservative approach, we advocate the smaller sampling fraction, $f_h = r_h / N_h$.

4. IMPLEMENTATION OF THE CES REDESIGN

The background and details of the CES Sample Redesign are detailed in Butani, Stamas and Brick (1997). The sample design is a stratified probability sample of Unemployment Insurance (UI) Accounts. These accounts may consist of one or many individual worksites in different SICs and MSAs within a single state. Although estimates are then produced using data collected for the individual worksites, this sampling mechanism provides the only operationally feasible method of capturing employment growth due to new worksites opening for existing businesses.

Sample stratification of UI accounts is based on SIC (11 major industry divisions), employment size class (8), and MSA of UI accounts. The MSA dimension is implicit, while the industry and size dimensions are explicit. The three largest size classes are collapsed into one size class for purposes of imputation. We decided to use these 6 imputation size classes for defining variance estimation strata since units within the largest (collapsed) size class are all sampled at or near certainty. Also, we decided not to recognize the MSA dimension of the stratification scheme in the variance calculations. These decisions result in 11x6 = 66 variance strata per state.

4.1 General Approach to Variance Estimation

Our general approach to variance estimation is to form two random groups within each variance stratum and to apply a BHS method to the two groups as described in Section 2.2. We will use 68 half samples per state to obtain a fully balanced set of replicates. The same half sample design will be repeated within each of the states to produce variance estimates for national estimates. The assignment of sample units to random groups is accomplished by first sorting all sample units within a state by industry/size/MSA/PRN (permanent random number). All sample units are then alternately assigned to random groups 1 and 2 from this sorted list. The random groups will thus reflect the same implicit stratification by MSA employed during sample selection.

A Hadamard matrix of order 68 will be used to define 68 replicates as described in Section 2.1. The rows of the matrix represent half sample replicates and the columns represent the 66 variance strata. The replicates for national estimates will be defined by repeating the columns for each state, thus creating a partially balanced design. However, the columns will be repeated in a circular fashion such that for a given industry different columns of the matrix will be used in different states. In our implementation, the first industry for the first state will use columns 1-6 of the matrix while the second industry for the first state will use columns 7-12 . The first state will use 66 of the 68 available columns (two of columns are not used). The first industry for the second state will use columns 7-12 of the matrix while the second industry for the second state will use columns 13-18. The last industry for the second state will use columns 1-6. Eighteen states will necessarily be forced to double up on the columns already used by the other states. This circular assignment of columns to states should increase the precision of the variance estimates for national estimates.

4.2 Sample Maintenance and Sampling Births

Once a year the sample will be re-selected for the purpose of maintaining near optimum sample allocations using the most current stratification variables available on the sample frame. The use of PRNs for sample selection provides an annual sample overlap of 85 to 90 percent. Once each quarter, except the first quarter when the entire sample is re-selected, the sample frame is updated for the purpose of identifying business (UI) births and deaths. Following a quarterly update, the deaths remain in the sample and random groups with collection variable values of 0. The births are sampled using the same stratification definitions as the initial sample, and assigned to random groups in the manner described earlier.

4.3 Incorporating Imputation Variance

Imputation will be implemented for each random group to incorporate the variance due to imputation in the BHS estimator. To this end we will also assign certainty units to the two random groups and use the fpc of $(1 - r_h/N_h)$ as described in Section 3. The results of our simulation study suggest that this will result in reasonable and conservative estimates of the variance. If all certainty units respond for any stratum, the fpc will result in a zero variance contribution for the stratum.

At least some of the variance strata, as defined above, will have relatively few numbers of sample units and even fewer responding sample units. Imputing for nonresponding units for each random group within these strata may pose problems if we are limited to the use of only one of two random groups within each stratum. Therefore, we decided to use the half sample and its complement – with weights as described in Section 2.5 – for the purpose of half-sample imputation, $\hat{y}_i$.

Incorporating all of these features leads to the estimator and weights as described in Section 2.5 with the conservative fpc of $(1 - r_h/N_h)$.

5. SIMULATION STUDY

The population used in the simulation is a real data set from the BLS. 1,000 stratified simple random samples were drawn without replacement from the population. The sampling fraction in each stratum (constructed by employment size and industry type) is similar to the real CES. Within each industry, there are some certainty strata consisting of large establishments.

For each unit in the population, there is an indicator value which equals 1 if the unit is a respondent, 2 if the unit is a nonrespondent but becomes a respondent in the next month, and 3 if the unit is a nonrespondent. All values of units in the population are observed, whether or not the indicator is 1, 2, or 3. These indicator values reflect the response pattern and are used in the simulation to generate respondents (and nonrespondents). With each sample, an independent stratified simple random sample of these indicator values was obtained and used as response indicator values for the obtained sample.

Two statistics were considered: the monthly total all employees and the monthly change.

Four different BHS estimators of variance were used. Method 1: fpc is used; separate imputation by half sample is not used. Method 2: fpc is used; separate imputation is used. Method 3: fpc is not used; separate imputation is not used. Method 4: fpc is not used; separate imputation is used.

Summary of the simulation results:

(1) Relative bias: Methods 1 and 3 both have a large negative relative biases. Method 2 has small bias but the bias is still negative. Method 4 has positive bias but the bias may be large.

(2) Stability: Variance estimators obtained without separate imputation have smaller variation than those obtained with re-imputation.

7. REFERENCES

Butani, S., Harter, R., and Wolter, K. (1997). "Estimation Procedures for the Bureau of Labor Statistics' Current Employment Statistics Program." Presented at the Joint Statistical Meetings, Anaheim.

Butani, S., Stamas, G., and Brick, M. (1997). "Sample Redesign for the Current Employment Statistics Survey." Presented at the Joint Statistical Meetings, Anaheim.

Fay, R.E. (1989), Theory and Application of Replicate Weighting for Variance Calculation, Proceedings of the Section on Survey Research Methods, American Statistical Association, 212-217.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. J. Amer. Statist. Assoc.

Werking, G. (1997). "Overview of the CES Redesign." Presented at the Joint Statistical Meetings, Anaheim.

Wolter, K. M. (1985), Introduction to Variance Estimation, New York: Springer-Verlag.

Table 1. Results of Simulation

	Estimator		Method 1		Method 2		Method 3		Method 4	
Month	RB	VAR	RB	CV	RB	CV	RB	CV	RB	CV
Monthly Total										
1	0.0	5.17×10^7	-47.8	32.5	-1.3	60.6	-33.4	39.1	24.3	71.0
2	0.0	8.98×10^7	-40.8	25.8	-10.6	37.8	-25.4	30.9	11.4	44.9
3	0.0	1.47×10^8	-31.4	29.3	-9.1	40.3	-14.5	34.7	12.6	47.3
4	0.0	2.35×10^8	-29.2	30.7	-9.8	20.9	-11.0	37.5	13.5	49.6
5	0.0	27.7×10^8	-21.6	36.4	-5.5	46.5	-1.1	45.4	19.0	57.6
6	0.0	2.70×10^8	-13.3	39.3	-6.9	41.4	10.0	49.0	17.9	51.6
Month-to-Month Trend										
1	0.3	5.17×10^7	-47.8	32.5	-1.3	60.6	-33.4	39.1	24.3	71.0
2	-0.5	5.94×10^7	-55.8	21.9	-10.9	40.5	-44.7	25.7	10.3	47.5
3	-1.3	7.48×10^7	-42.7	30.4	-5.7	46.3	-29.4	35.2	15.1	53.2
4	6.2	1.21×10^8	-45.2	32.3	-11.2	50.6	-29.9	40.7	13.1	63.0
5	3.6	1.13×10^8	-49.2	28.4	-8.8	59.2	-35.0	41.4	15.6	70.7
6	18.5	1.15×10^8	-49.5	28.4	-12.0	43.7	-36.2	33.3	9.9	51.6