International Journal of Pure and Applied Mathematics

Volume 40 No. 3 2007, 399-410

OPTIMAL PAIRING FOR STRATUM COLLAPSE METHODS WITH INTERVIEWER-LEVEL MEASUREMENT ERROR SHARED ACROSS STRATA

Moon Jung Cho[1], John Eltinge[2 §], Eungchun Cho[3]

[1,2]U.S. Bureau of Labor Statistics
2 Massachusetts Avenue NE, Washington, DC, 20212, USA
[1]e-mail: Cho.Moon@bls.gov
[2]e-mail: Eltinge.John@bls.gov
[3]Division of Mathematics and Sciences
Kentucky State University
Frankfort, KY 40601, USA
e-mail: eung.cho@kysu.edu

Abstract: This paper introduces simple variance estimators based on a collapsed-stratum approach. The collapse procedure is intended to ensure that the newly paired primary sampling units do not share a common interviewer, and the modified standard replicate-based variance estimators are expected to take into account both sampling error and interviewer level measurement error. Expanded collapsing procedure is considered to incorporate similar population characteristics. Specific matching algorithms are developed and applied to data from the U.S. Consumer Expenditure Interview Survey.

AMS Subject Classification: 68Q20
Key Words: collapse method, consumer expenditure interview survey, interviewer effects, nonsampling error, stratified multistage sample design

1. Introduction: Variance Estimation for the U.S. Consumer Expenditure Interview Survey

The Consumer Expenditure (CE) Survey uses a stratified multistage probability

Received: August 25, 2007

[§]Correspondence author

sample of households which represents the total U.S. civilian noninstitutional population. For some general background on the CE Interview Survey and variance estimation therein, see Chapter 16 of the BLS Handbook [2] .

To select a representative sample of the population, the CE Survey divides the nation into many areas and then selects some of these areas, and the selected areas are called "Primary Sampling Units" (PSUs). The PSUs are groups of counties, or independent cities, and are selected with unequal probabilities. There are self-representing PSUs and non self-representing PSUs. The self-representing PSUs are the ones which are selected with probability=1 and are from metropolitan areas. Within each selected PSU, a given sample consumer unit (CU), roughly equivalent to a household, is randomly assigned to one of two modes of data collection: interview or diary. This paper will consider only data from the CE Interview Survey.

The set of sample PSUs used for the survey consists of 101 areas, and there are 105 PSUs in the CE interview data of the Year 2000. The CE Interview Survey collapsed 105 PSUs to form 80 variance PSUs and then assigned two variance PSUs to each variance stratum. Variance estimates are obtained by the balanced repeated replication (BRR) method. For some general background on the replicate-based variance estimations, see Wolter [21]. Under a standard design of BRR, two PSUs are selected with replacement from each stratum and then one PSU is selected from each stratum in a balanced manner to form a set of half samples. These half samples are used to compute the resulting variance estimator.

In the traditional sampling literature, justification for this approach uses the (approximate) independence of sample selection across strata and PSUs, and focuses only on the sampling error component of survey error. However, in the CE Interview Survey, we often need to have variance estimators that account for both sampling and measurement error. In addition, interviewers often collect data in more than one PSU. For some variables, the interviewer-level component of measurement error may be nontrivial. Consequently, one must consider the modification of standard replicate-based variance estimators that will account appropriately for the correlation across strata and PSUs induced by interviewer level measurement error. This paper considers some simple variance estimators based on a collapsed-stratum approach. The collapsing procedure is intended to ensure that the newly paired variance-PSUs do not share a common interviewer but have similar population characteristics. Specific matching algorithms are developed and applied to data from the CE Interview Survey. These algorithms arise from optimality criteria and use stratum and primary-unit level variables such as population size.

2. Optimal Pairing

2.1. Pairing of Objects

In this section, a general problem of matching n objects into $n/2$ pairs according to a given compatibility criterion is considered.

Let $X = \{a_1, \ldots, a_n\}$, and $\mathbf{c}$ be a real valued function on $X \times X$ with values between 0 and 1 measuring compatibility of pairs. $\mathbf{c}\,(a_i, a_j) = 0$ indicates a_i and a_j are absolutely incompatible, and $\mathbf{c}\,(a_i, a_j) = 1$ means a_i and a_j are perfectly compatible. For a simplified version of the CE application, each a_i is a PSU, containing several interviewers. The pairing algorithm will be constructed to minimize the extent to which the paired PSUs share the same interviewers.

We are interested in finding an optimal pairing, which, by definition, is a set P of disjoint $n/2$ pairings of n objects such that the sum $\sum\limits_{(a_i,a_j)\in P} \mathbf{c}\,(a_i, a_j)$ is maximal, i.e., the set which achieves a maximal total compatibility. Obviously, an exhaustive search for an optimal pairing is infeasible; the number of all possible ways of pairing n objects is $n!/\left(2^{n/2}(n/2)!\right) = \Pi_{i=1}^{n/2}(2i - 1)$. For example, for $n = 20$, the number of all possible pairings is $654, 729, 075$. We propose an efficient algorithm for finding a pairing. The proposed pairing is in practice, optimal in terms of the number of iterations it takes to achieve optimal pairing criteria, the number of remaining pairs after the final iteration, and the total processing time. We note that this algorithm found an optimal solution in all simple cases for which we could check exhaustively.

Algorithm:

1. Construct a matrix M with $m_{ij} = \mathbf{c}\,(a_i, a_j)$.

2. Construct a row vector S with entries equal to the column sums of M, i.e., $S_j = \sum\limits_i m_{ij}$.

3. Find i such that $S_i \leq S_j$ for all j.

4. Find j such that $m_{j,i} \geq m_{k,i}$ for all k, i.e., a_j is maximally compatible with a_i.

5. Pair off a_i with a_j.

6. Remove i-th and j-th rows and columns from M, i-th and j-th elements from S.

7. Repeat until M becomes empty.

At each iteration, the algorithm picks one element, say a_i, that is least compatible to the rest of the objects, and then a_i is paired with an element that is most compatible to a_i. The algorithm assumes that n is even. If n is odd, one object, say a_i, for which S_i is minimal, i.e., the one that is minimally compatible to others, is removed from $X = \{a_1, \dots, a_n\}$.

The optimal pairing problem, in general, does not have unique solution. An extreme example is when $X = \{a_1, \dots, a_n\}$ and

$$\begin{aligned} \mathbf{c}\,(a_i, a_j) &= 0, \text{ if } i = j\,, && (1) \\ &= 1, \text{ otherwise }, && (2) \end{aligned}$$

in which case any set of complete pairings is optimal.

If the pairing criterion is to pair the sets which are disjoint, then an optimal pairing is achieved when all paired sets are disjoint. For example, if $X = \{\{1,2\}, \{1,3\}, \{3,4\}, \{2,4\}\}$, then $\{\ [\{1,2\},\{3,4\}]\ ,\ [\{1,3\},\{2,4\}]\ \}$ is the optimal pairing.

If the pairing criterion is to pair the sets of similar sizes (cardinality) that share no or very few elements in common, then following function $\mathbf{c}$ that incorporates this criterion

$$\mathbf{c}\,(a_i, a_j) = \frac{|a_i \Delta a_j|}{2\max(|a_i|, |a_j|)}\,, \tag{3}$$

where $|a|$ is the number of elements in a, $a_i \Delta a_j = (a_i - a_j) \cup (a_j - a_i)$, the symmetric difference of a_i and a_j. For example, if $A = \{1,2\}$, $B = \{2,3\}$, $C = \{3,4\}$ and $D = \{4,5\}$ then $\mathbf{c}(A,A) = 0$, $\mathbf{c}(A,B) = \frac{1}{2}$, $\mathbf{c}(A,C) = 1$, $\mathbf{c}(A,D) = 1$, etc., and $\{[A,C], [B,D]\}$ is an optimal pairing.

In practice, there are several characteristics that determine the compatibility between objects. In those cases, the following $\mathbf{c}$ given as a weighted sum can be used

$$\mathbf{c}\,(a, b) = \sum \gamma_k\, \mathbf{c}_k\,(a, b)\ , \tag{4}$$

where $\mathbf{c}_k$ is a measure of compatibility based on k-th characteristics of objects, γ_k are nonnegative weights and $\sum \gamma_k = 1$.

2.2. Pairing in a Metric Space

The pairing algorithm can be extended to any abstract space equipped with a metric. The points are not necessarily numbers or vectors, but sets, figures, names, or objects with complex structures.

A metric space X is a set with a metric $\mathbf{d}$, a nonnegative valued function on $X \times X$ satisfying nontriviality, symmetry, and the triangle inequality, see [18].

Suppose T is a set of n points in a metric space X. $T = \{x_1, \ldots, x_n\}$. A function $\mathbf{c}$ measuring the compatibility between the elements can be defined in terms of $\mathbf{d}$. Roughly speaking, shorter distance between points would correspond to higher compatibility. Such a function $\mathbf{c}$ can be defined in many ways, here we propose a simple algebraic form

$$\begin{aligned} \mathbf{c}\,(a,b) &= 0\,, \quad \forall a = b \in X\,, & (5)\\ &= \frac{1}{1+\mathbf{d}\,(a,b)}\,, \quad \forall a \neq b \in X\ . & (6) \end{aligned}$$

That $\mathbf{c}\,(a,a) = 0$ indicates, obviously, we do not want to pair an object with itself.

For example, the set of four vertices v_i's of a regular tetrahedron (the standard 3-simplex) satisfies $\mathbf{d}\,(v_i, v_j) = \sqrt{2}$ thus $\mathbf{c}\,(v_i, v_j) = 1/(1+\sqrt{2})$ for $i \neq j$. And the three possible pairings $\{[v_1, v_2]\,, [v_3, v_4]\}$, $\{[v_1, v_3]\,, [v_2, v_4]\}$, and $\{[v_1, v_4]\,, [v_2, v_3]\}$, are all optimal. For another example, the set of four vertices v_i's of a square can be optimally paired by $\{[v_1, v_2]\,, [v_3, v_4]\}$ or by $\{[v_1, v_4]\,, [v_2, v_3]\}$.

Since a positive linear combination of metrics is also a metric, $\mathbf{d}\,(a,b) = \sum \gamma_k\, \mathbf{d}_k\,(a,b)$ is a new metric, where γ_k are positive and $\sum \gamma_k = 1$.

It can incorporate measurements of various aspects of objects.

3. Variance Estimators Based on Standard and Modified Pairings of Variance PSUs

Consider a survey variable Y_{hij} for CU j in design PSU i in design stratum h and define the population total

$$Y = \sum_{h=1}^{L} Y_h\,,$$

where $Y_h = \sum\limits_{i=1}^{N_h} Y_{hi}$, $Y_{hi} = \sum\limits_{j=1}^{M_{hi}} Y_{hij}$, L is the number of design strata, N_h is the number of design PSUs in design stratum h, and M_{hi} is the number of CUs in design PSU i in design stratum h. Given probability weights w_{hij} and a two-PSU-per-stratum design, a simple estimator of Y is $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$, where $\hat{Y}_h = \hat{Y}_{h1} + \hat{Y}_{h2}$ and $\hat{Y}_{hi} = \sum_{j \in S_{hi}} w_{hij} \hat{Y}_{hij}$. A data analysts often wishes to estimate the variance

$$V(\hat{y}) = E\left[\{\hat{y} - E(\hat{y})\}^2\right] ,$$

where the expectation operator involves integration with respect to both the sample design and the model associated with measurement errors including interviewer effects. In addition, a simple estimator of the variance of $\hat{Y}$ is

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}(\hat{Y}_h) , \tag{7}$$

where

$$\hat{V}(\hat{Y}_h) = \left(\hat{Y}_{h1} - \hat{Y}_{h2}\right)^2 \quad . \tag{8}$$

Under the assumption that sampling and interviewing are independent across strata and that design PSUs are selected with replacement, $\hat{V}(\hat{Y})$ is unbiased for the design variance of $\hat{Y}$.

Under the conditions described in Section 1, the estimator (7) may have a negative bias if a given interviewer collects data in both PSUs selected from a given stratum h. Consequently, one may wish to replace the variance estimator (7) with

$$\hat{V}^*(\hat{Y}) = \sum_{g=1}^{G} \hat{V}^*(\hat{Y}_g) , \tag{9}$$

where the $\{2 \times L\}$ design PSUs are partitioned into G groups called "variance strata"; variance stratum g contains $n_{(g)}$ design PSUs; these $n_{(g)}$ design PSUs are placed into two groups called "variance PSUs" $s_{(g1)}$ and $s_{(g2)}$ such that no interviewer collects data in both $s_{(g1)}$ and $s_{(g2)}$; $\hat{Y}_{(g)} = \hat{Y}_{(g1)} + \hat{Y}_{(g2)}$ and $\hat{V}^*(\hat{Y}_{(g)}) = \left(\hat{Y}_{(g1)} - \hat{Y}_{(g2)}\right)^2$. Under the assumption that the expectations of $\hat{Y}_{(g1)}$ and $\hat{Y}_{(g2)}$ are equal and additional regularity conditions, $\hat{V}^*(\hat{Y})$ will be approximately unbiased or conservative for the combined variance of $\hat{Y}$.

We applied optimal pairing algorithm to the CE Interview Survey data. Using the notation in Section 2.1, $X = \{s_{(11)}, s_{(12)}, s_{(21)}, s_{(22)}, \ldots, s_{(G1)}, s_{(G2)}\}$, and our optimality criterion was that no variance PSUs of the variance stratum shared the common interviewers.

For this application, we used the weighted function

$$\mathbf{c}\,(a,b) = \gamma_1 \mathbf{c}_1\,(a,b) + \gamma_2 \mathbf{c}_2\,(a,b)\ , \tag{10}$$

where $\gamma_1 \in [0,1]$, $\gamma_2 = 1 - \gamma_1$, $\mathbf{c}_1 = 0$ if any overwrapping, $\mathbf{c}_1 = 1$ if purely disjoint, and

$$\mathbf{c}_2(a,b) = 1 - \left[\frac{\max(|a|,|b|) - \min(|a|,|b|)}{\max(|a|,|b|)}\right]. \tag{11}$$

Moreover, we define the column sums

$$S_{1j} = \sum_{i=1}^{n} \mathbf{c}_1\,(a_i, a_j)\,, \quad S_{2j} = \sum_{i=1}^{n} \mathbf{c}_2\,(a_i, a_j)\,,$$

and

$$S_j \;\;=\;\; \gamma_1 S_{1j} + \gamma_2 S_{2j}\,.$$

Note that if $\gamma_1 = 1$, the algorithm will be driven entirely by the pattern of interviewer overlap across PSUs. Conversely, if $\gamma_1 = 0$, the algorithm will be driven entirely by the pattern of relative sizes of PSUs.

The remainder of this paper presents numerical results from the newly paired variance PSUs and variance strata with $\gamma_1 = 1$ in the CE application.

4. Applications to the U.S. Consumer Expenditure Survey

The CE Survey uses two modes of data collection: diary and interview. The principal reason for this use of multiple collection modes is that some expenditures (generally small or frequently purchased items) are believed to be more readily captured through a diary, while other items (generally purchases that are larger, less frequent, or more salient) are more readily captured through a periodic in-person interview. Expenditures are reported at a relatively fine level of aggregation known as the six digit Universal Classification Code (UCC) level, see [4].

We considered only the components of the mean monthly expenditure of the CE Interview Survey that contribute to current CE production estimates. In particular, we exclude interview data collected for UCCs that are published on the basis of diary data only. Consequently, the "Overall Mean" entries are based on data from the 432 UCCs for which publication is based on the

	Estimate	$se_{current}$	$se_{proposed}$
Overall	2106.49	20.460	48.605
Apparel*	36.65	0.894	1.951
Home Furnishing	103.48	2.666	3.518
Travel*	3.97	0.324	0.338
Utilities	227.54	1.944	3.634

Table 1: Monthly mean expenditure estimates

interview reports. In addition, the entries for "Apparel" are based on apparel UCCs that are published from interview data; similarly for home furnishings; travel and utilities.

The data used for this analysis were generated from the monthly expenditures files, and the CU characteristics and income files of the Phase 3 databases. For computing the mean monthly expenditure, we used the UCCs collected in the interview survey and used for publishing quarterly expenditures. We did not include pension and social security expenditures.

We computed the mean monthly expenditure for the selected subgroups of CE expenditure data. Some examples of subgroups are house furnishing, apparel, travel, and utility. We chose those subgroups based on the frequency of purchase and the saliency of the expenditure. Table 1 presents item categories, monthly mean expenditure estimates, standard error estimates using the current method and standard error estimates using the proposed method. Note that asterisk in the table indicates the published expenditure estimate is mainly from diary survey. We observed that for overall expenditure, apparel and utilities, standard error estimates using the proposed method are about twice as large as the ones using the current method.

A misspecification effect is one measure of the bias of a variance estimator computed from data collected through a complex sample design. For general background on misspecification effects, see Skinner [20]. Define a univariate misspecification effect,

$$\Delta = E(\hat{V}_{proposed}) \,/\, E(\hat{V}_{current})\,, \tag{12}$$

where $\hat{V}_{proposed}$ is a variance estimator computed through our new pairing of variance PSUs within variance strata using our optimality criterion, and $\hat{V}_{current}$ is a variance estimator currently used in production, and $E(\cdot)$ represents expectation evaluated with respect to both the sample design and the nonsampling

	$\hat{\Delta}$	$\hat{r}$
Overall	5.644	0.123
Apparel	4.764	0.100
Home Furnishing	1.740	0.020
Travel	1.090	0.002
Utilities	3.495	0.066

Table 2: Within-Interviewer correlation

error model. If $\hat{V}_{proposed}$ is unbiased for the true variance of our point estimator, and if the bias in $\hat{V}_{current}$ is attributable only to the clustering effects associated with our interviewers, then we may use Δ to approximate the effect of clustering in terms of the measure of homogeneity, see [13]:

$$(12) = 1 + (\bar{b} - 1)r\,, \tag{13}$$

where $\bar{b}$ is an average workload of interviewers, and r is a within-interviewer correlation. Table 2 presents item categories, univariate misspecification effect estimates and within-interviewer correlation estimates. We observed that for overall expenditure, apparel and utilities, within-interviewer correlation estimate is greater than or equal to 10%.

We compared correlation structure between variance estimators of two methods. We computed a correlation coefficient matrix, ρ, using following formula:

$$\rho = D^{-\frac{1}{2}} V D^{-\frac{1}{2}}\,, \tag{14}$$

where $D^{-\frac{1}{2}}$ is an inverse square root diagonal matrix of variance and V is a variance-covariance matrix. We then obtained confidence intervals using the Fisher z transformation and the Bonferroni inequality. Fisher showed the transformation applied to correlation coefficients produced quantities that were asymptotically normally distributed [15]

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \tanh^{-1} r\,,$$

with mean $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and variance $s = \frac{1}{n-3}$, where $n = df + 1$, and df is a value of degrees of freedom. The $(1-\alpha)\%$ Bonferroni confidence interval [10] of r is $r \pm t_{n-1,\frac{\alpha^*}{2}}\sqrt{s}$ with $\alpha^* = \frac{\alpha}{p}$ and p the number of correlation coefficients we were interested in. Table 3 presents the Bonferroni simultaneous 50% confidence

Table 3: Selected simultaneous 50% confidence intervals

Method	i	j	Lower	$\hat{\rho}_{i,j}$	Upper
Current	Overall	Apparel	-0.15	0.18	0.48
Proposed	Overall	Apparel	0.64	0.80	0.89
Current	Utilities	Apparel	-0.30	0.02	0.34
Proposed	Utilities	Apparel	0.35	0.61	0.78

intervals for correlations between mean estimators for selected expenditure categories. We observed the confidence intervals of correlation estimators from the two methods did not overlap for these cases.

5. Discussion

To implement the variance PSU matching algorithm in Section 2, it was necessary to conduct all possible pairwise comparisons among variance PSUs, and see whether two variance PSUs share common interviewers. Note that the number of interviewers within a variance PSU may differ from the number of interviewers within other variance PSU. See Cho and Eltinge [3] for the discussion on the use of commercial software and *Maple* program code to re-group PSUs into new variance strata and PSUs.

Depending on the algorithms, it is sometimes necessary to run a number of iterations to assign variance PSUs which do not share the interviewers to each variance stratum for all variance strata. The algorithm presented in Section 2.1 showed the separability (No-Common-Interviewers) through one iteration.

It is possible to incorporate various optimality criteria in the objective function matrix used to construct a given set of variance PSUs and strata. For example, one may match pairs on the basis of the number of interviewers assigned to a variance PSU, or on the basis of the number of interviewers that a given variance PSU shared with other variance PSUs, or on the basis of population characteristic variables in variance PSU pairing. When including more than one variables in the optimal criteria, we found that results were sensitive to how we weighed components. Therefore one issue is how much weight we want to assign on each variable considered in this case.

Acknowledgements

Eungchun Cho was partially supported by the NIH/NICHD EARDA grant HD 038435. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

References

[1] B. Bailar, L. Bailey, J. Stevens, Measures of interviewer bias and variance, *Journal of Marketing Research*, **XIV** (1977), 337-43.

[2] Bureau of Labor Statistics, *BLS Handbook of Methods*, U.S. Department of Labor, Washington, DC, available at: www.bls.gov/opub/hom/hom-toc.htm

[3] M.J. Cho, J.L. Eltinge, Optimal pairing for stratum collapse methods with interviewer-level measurement error shared across strata, In: *Proceedings of the American Statistical Association*, Section on Survey Research Methods (2005), 2873-2880.

[4] J.L. Eltinge, A. Sukasih, W. Weber, Feasibility of constructing combined estimators using consumer expenditure interview and diary Data, Manuscript, Office of Survey Methods Research, U.S. Bureau of Labor Statistics (2000).

[5] R.E. Fay, Theory and application of replicate weighting for variance calculations, *Proceedings of the American Statistical Association*, Section on Survey Research Methods (1989), 212-217.

[6] W.A. Fuller, Sampling with random stratum boundaries, *Journal of the Royal Statistical Society*, Series B, Methodological, **32** (1970), 209-226.

[7] M.H. Hansen, W.N. Hurwitz, W.G. Madow, *Sample Survey Methods and Theory*, John Wiley and Sons, New York (1953).

[8] H.O. Hartley, J.N.K. Rao, G. Kiefer, Variance estimation with one unit per stratum, *Journal of the American Statistical Association*, **64** (1969), 117-123.

[9] K.M. Heal, M.L. Hansen, K.M. Rickard, *Maple V - Learning Guide*, Springer-Verlag, New York (1996).

[10] R.A. Johnson, D.W. Wichern, *Applied Multivariate Statistical Analysis*, 4-th Ed., Prentice-Hall (1998).

[11] D.R. Judkins, Fay's method for variance estimation, *Journal of Official Statistics*, **6**, No. 3 (1990), 223-239.

[12] L. Kish, Studies of interviewer variance for attitudinal variables, *Journal of the American Statistical Association*, **57**, No. 297 (1962), 92-115.

[13] L. Kish, *Survey Sampling*, John Wiley and Sons, New York (1995).

[14] E.L. Korn, B.I. Graubard, Estimating variance components by using survey data, *Journal of the Royal Statistical Society*, Series B, **65**, Part 1 (2003), 175-190.

[15] D.F. Morrison, *Multivariate Statistical Methods*, Second Edition, McGraw-Hill Book Company, New York (1976).

[16] C. O'Muircheartaigh, P. Campanelli, A Multilevel exploration of the role of interviewers in survey non-response, *Journal of the Royal Statistical Society*, Series A, **162**, No. 3 (1999), 437-446.

[17] C. O'Muircheartaigh, P. Campanelli, The relative impact of interviewer effects and sample design effects on survey precision, *Journal of the Royal Statistical Society*, Series A, **161**, No. 1 (1998), 63-77.

[18] J. Munkres, *Topology*, Second Edition, Prentice Hall (1999).

[19] K. Rust, G. Kalton, Strategies for collapsing strata for variance estimation, *Journal of Official Statistics*, **3**, No. 1 (1987), 69-81.

[20] C.J. Skinner, D. Holt, T.M.F. Smith, Eds *Analysis of Complex Surveys*, John Wiley and Sons, New York (1989).

[21] K.M. Wolter, *Introduction to Variance Estimation*, Springer-Verlag, New York (1985).