AN EMPIRICAL INVESTIGATION OF A ONE-THIRD REPLICATION SCHEME FOR VARIANCE ESTIMATION FROM THE CURRENT POPULATION SURVEY

Brian Clarridge and Charles Palit
University of Wisconsin-Madison

Introduction

In this paper we provide an empirical demonstration of the relative efficiency of different one-third sample replication schemes for estimating sampling error from the 1973 CPS. Specifically we show the computing times and empirical estimates of the efficiency of the schemes for calculating the variance of means and correlations.

A common practice among users of CPS data at this time is to compute point estimates from the CPS and to ignore sampling error estimation. A substantial part of the motivation for this practice is the difficulty and expense involved in computing sample error estimates with some of the more complicated schemes, or with the 243 replication scheme of the Bureau of the Census.

By showing the relative efficiency of different sized one-third replication schemes, we hope to provide users of CPS data with information which will enable them to better assess the suitability and practicality of using such schemes and their variants. Here, we use replications of sizes 81, 27, and 9 since they are sequentially the next smallest, completely balanced and centered, one-third replication schemes after the 243 scheme used by the Census.[1] Naturally, a characteristic of these replication schemes is that they are less expensive to use than the 243 scheme because they use less computer time. What this paper provides is some information on the loss of stability in the variance estimate associated with reducing the number of replications.

Background

Almost from the beginning of the collection of data for the Current Population Survey by the Bureau of the Census, there has been concern over the reliability of the parameter estimates produced from these data. The interest from within the Bureau of the Census has been dual, "The two objectives presently considered in estimating variances of the major statistics of interest for CPS are--

1. to measure the precision of the survey findings.
2. To evaluate the impact of each of the stages of sampling and estimation on the overall precision of the survey." (Bureau of the Census, 1978, p. 90)

Over time, the methods used to provide the variance estimates for these two purposes have ranged from relatively complex direct methods to less complicated indirect methods (Hansen, Hurwitz, and Madow, 1953; Keyfitz, 1957; Bureau of the Census, 1963; McCarthy, 1966; Banks and Shapiro, 1971; Kish and Frankel, 1974; and Bureau of the Census, 1978).

Two of the most popular methods in use are the Keyfitz method and the replication method. Each of these has undergone design transformations which have either increased the precision of the variance estimate produced or simplified its application. Currently, the Bureau of the Census uses a modified Keyfitz method because it is believed to have greater precision than the replication method as well as a simplicity of application (Bureau of the Census, 1978, pp. 91-94).

Unfortunately, many non-Census CPS users have found both the Keyfitz method and the replication method with the 243 replicate design to be excessively costly in terms of computer time. The resulting situation is that the research done by social scientists generally contains parameter estimates without the corresponding estimates of error. In an effort to help alleviate this situation, we examine here the possibility of reducing the number of replicates used in the calculation of variance estimates from the CPS.

Procedure

Using March, 1973 CPS data we selected 9 variables more or less representative of those most often used in analyses. We calculated the variance of their estimated means and correlation coefficients, using one-third sample replication schemes involving 9, 27, and 81 replicates. We then compared the precision of the estimates obtained from these different-sized replications as well as the amount of computer time consumed in their estimation. Ultimately, we estimated the relative efficiency of the different-sized replication schemes so that prospective users might better assess their merits.

We describe the procedure used in two parts. In the first part, we describe how we divided the sample into one-third subsamples and how we determined the number of strata required for each sized replication scheme. In the second part, we examine the results obtained with the different-sized replication schemes and we compare them.

Number and Creation of Strata

The data used for our analysis were from the Annual Demographic File of the March, 1973 CPS. There were 47,770 household records and 136,122 person records in the sample. We wanted to divide the person records into one-third subsamples with the appropriate number of strata so that orthogonal replicates could be obtained using the method that Gurney and Jewett provided in their 1975 paper. Since one-third subsamples conform well to the sampling design of the CPS, we thought we could maintain nationally representative distributions of cases in each subsample. We used the random cluster codes and the state codes attached to each record to do this.

The Bureau of the Census aided our efforts by providing us a list of random cluster codes which were divided into three groups. There were 402 codes for the 191 Self Representing PSUs, 157 codes for the 132 Non-Self Representing PSUs with one PSU per stratum, and 192 codes for the 176 Non-Self Representing PSUs with two PSUs per stratum. The "strata" referred to here are those used by the Bureau of the

Census internally in the 110 strata scheme discussed in the paper by Gurney and Jewett. Since we had no information about the method used by the Bureau to create the one-third samples in their large replication scheme, we had to devise our own method of doing this for our smaller 81, 27, and 9 replication schemes.

As a first step in creating one-third samples with the appropriate number of strata, we had to know how many were needed for each scheme. The paper by Gurney and Jewett (1975) provides the following formula for this calculation,

$$N \leq (p^n-1)/(p-1)$$

where N = number of strata,
p = number of partitions into which the sample is divided,
n = an integer such that p^n is the number of replicates.

In our case, p = 3 and n = 4, 3, and 2 corresponding to the appropriate replication schemes. This constrains the number of strata to maximums of 40, 13, and 4. Since the stability of the estimates obtained would be decreased if fewer strata were used, we used the maximum in each case. Therefore, the task was to design a selection procedure to put the 499 PSUs into three one-third subsamples of 40, 13, and 4 strata respectively.

We were interested in obtaining a good mix of SR PSUs and NSR PSUs in each subsample. To achieve this, we treated the SR and NSR portions of the sample separately (see Figure 1). That is, we divided the SR PSUs into three one-third samples; then we divided the NSR PSUs into three one-third samples; and later we combined the two of them into three full one-third samples each containing representative SR and NSR PSUs. For the 40 strata design, we created 23 strata with the SR PSUs and 17 strata with the NSR PSUs. This was done to approximate the population distribution in the United States between SR and NSR PSU areas. The selection procedure for the specific SR and NSR subsamples differed somewhat because of the form in which the random cluster codes were listed.

For the SR PSUs, we stratified them with respect to size and geographic area and then we took a systematic one-third sample of them. This gave us three geographically-heterogeneous groups of PSUs, each representing roughly the same proportion of the population. There were 64 PSUs in two of these groups and 63 in the other. From here it was convenient to collapse the PSUs in each one-third sample into 23 strata by recollecting the PSUs into 22 geographic groupings and then creating the 23 stratum by splitting the largest of these into two parts. These became the 23 SR strata in the 81 replicate scheme.

The NSR PSUs presented a somewhat different situation in regard to their separation into three one-third samples. In the first place, the NSR group had 88 sets of paired PSUs along with 132 single NSR PSUs. Since it was assumed that the paired NSR PSUs were matched on their population characteristics before pairing, we split these 88 pairs into two groups and let them be two-thirds of the foundation on which the subsamples were built. Random clusters were selected for the third subsample on the basis of similarities of geographic characteristics and size to the two previously paired clusters. Not all NSR PSUs matched up well under this plan, but we did the best we could to match them given our limited knowledge of their creation. Once we had three one-third samples, we again stratified by geographic area and we combined a few adjacent geographical areas until we reduced the strata down to the 17 required for the 81 replication scheme.

Subsequently, both the SR strata and the NSR strata were collapsed to form 13 and 4 strata for the 27 and 9 replication schemes respectively. This was again based on some decisions we made with respect to geography and the number of cases represented by each strata.[2]

Once the strata were formed, we created the appropriate orthogonal designs for the 81, 27, and 9 replication schemes. We followed techniques explained previously (Gurney and Jewett, 1975) and created the generators for the replication matrices. These generators are displayed in Table 1.

Comparison of Results

Now that we had created both the strata and the orthogonal replicates for the different schemes, we proceeded to the calculation of the estimates. Each replication set was used together with the formula:

$$S_R^2 = \sum_k^R \left(x_k' - x'\right)^2 \Big/ R(L-1)$$

where x_k' = parameter estimate from the k^{th} replication
x' = parameter estimate from whole sample
R = # replications
L = # subsamples (3)

to produce variance estimates for means and correlations at the national level for the variables age of head, family size, family income, age, sex, race, education, income, and weeks worked.

These variance estimates provide the data which we use to evaluate the results of each replication scheme. For the 9-replication and the 27-replication scheme, we constructed a measure of the distance of each variance estimate from the corresponding estimate derived from the 81-replication scheme. We call this measure Δ_R .

$$\Delta_R = \left| \frac{S_R^2 - S_{81}^2}{S_{81}^2} \right|$$

where R = number of replications in the replication scheme
S_R^2 = the variance estimate for some parameter based on R replications.

Table 2 shows the values of Δ_R for each of the nine variables and for R = 9 and 27.

Figure 1

GENERAL DESIGN FOR THE CREATION OF STRATA IN THE 40 STRATA SCHEME

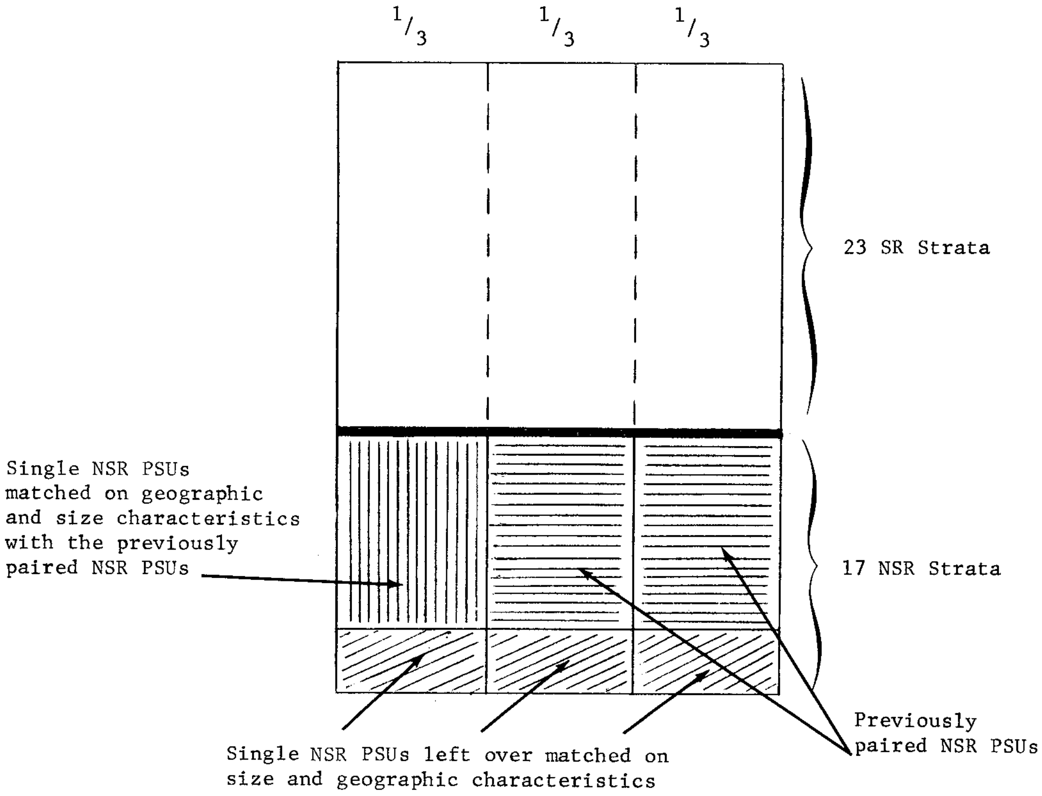
1/3
1/3
1/3
23 SR Strata
17 NSR Strata
Single NSR PSUs matched on geographic and size characteristics with the previously paired NSR PSUs
Previously paired NSR PSUs
Single NSR PSUs left over matched on size and geographic characteristics

TABLE 1

FIRST COLUMN OF MATRICES USED TO PRODUCE ORTHOGONAL REPLICATES OF SIZES 9, 27, AND 81

Replication Scheme	First Column
9	10122021
27	110020211221022200101211201
81	111201211212020221102011001222021002 0002222102122121010112201022002111012 0010001

TABLE 2

Δ_R VALUES FOR MEANS IN THE 9 AND 27 REPLICATE SCHEMES

Variable	Δ_9	Δ_{27}
Age of head	.385	.168
Family size	.001	.003
Family income	.080	.118
Age	.036	.071
Sex	1.000	.000
Race	.337	.000
Education	.212	.013
Income	.141	.101
Weeks worked	.059	.022
Average $\bar{\Delta}$	.250	.055

TABLE 3

DISTRIBUTION OF Δ_R VALUES FOR CORRELATION COEFFICIENTS IN THE 9 AND 27 REPLICATION SCHEMES

Value of Δ_R	Freq. D^n for Δ_9	Freq. D^n for Δ_{27}
0.0 - .10	17	32
.1 - .2	7	2
.2 - .3	5	1
.3 - .4	2	0
.4 - .5	1	1
.5 - .6	1	-
.6 - .7	1	-
.7 - .8	-	-
.8 - .9	-	-
.9+	2	-
Average $\bar{\Delta}_R$	.283	.053

As expected, $\bar{\Delta}_9$, the average of the Δ_9's for means, is larger than the corresponding $\bar{\Delta}_{27}$. $\bar{\Delta}_9 = .25$ while $\bar{\Delta}_{27} = .06$. We take this to mean that the effect of tripling the number of replications used in the variance estimation is to reduce the relative instability of the variance estimate by a factor of about four. On the other hand, the price paid for this was to increase the direct computational time by a factor of three.[3]

Table 3 shows the frequency distribution of Δ_9, and Δ_{27} for the Δ's associated with the variance estimates of 36 correlation coefficients estimated from pairs of the nine selected variables.

The average $\bar{\Delta}_R$ here are .28 and .05 which would seem to indicate that the stability of the variance estimate is converging at about the same rate for the correlation coefficients as for the means.

It may not be surprising that the relative stability of the variance estimates increases with an increase in the size of replication scheme. Our concern is with the rate at which the stability increases as we increase the number of replications. These results show a dramatic increase in stability as we move from 9 to 27 and a somewhat smaller increase as we move from 27 replicates to 81. This leads us to speculate that the increase in stability if we move from 81 replications to 243 replications may be even smaller. Should this be true, then we would think that the 27-replication scheme would be a reasonable choice for variance estimation for most researchers using the CPS sample on a national level.

FOOTNOTES

[1] For the same number of strata it is possible to get a balanced set of orthogonal replications fewer in number than provided for in the 1975 paper by Gurney and Jewett. However, such a set will not be centered.

[2] Details of this procedure will be provided upon request. It should be noted here that not all the random cluster codes supplied by the Bureau of the Census for March, 1973, had matches on our copy of the tape. Likewise, there were a few codes on the tape which were not on the list provided. There were about ten of each type of mismatch.

[3] Computation time on an IBM-370/135 was 118 minutes for 9 replicates, 361 minutes for 27 replicates and 1208 minutes for 81 replicates. While these timing figures represent the total machine time for each set of replications, it should be noted that the 27 replicate run and the 81 replicate run were broken down into smaller jobs to improve efficiency. During all runs done for this paper, our IBM machine was devoted solely to the performance of the task assigned by us with no other time-sharing constraints.

ACKNOWLEDGMENTS

The authors are indebted to Charles Jones, Gary Shapiro, and Tina Olsen of the Bureau of the Census and Gene Sharp of the University of Wisconsin for materials and computational assistance.

BIBLIOGRAPHY

Banks, M. J., and Shapiro, G. M. "Variances of the Current Population Survey, including within - and between - PSU components and the effect of the different stages of estimation." Proceedings of the Social Statistics Section, American Statistical Association, 1971, pp. 40-49.

Gurney, M. and Jewett, R. S. "Constructing orthogonal replications for variance estimation." Journal of the American Statistical Association, 70, (December), 1975.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. Sample Survey Methods and Theory, Vols. I and II. New York: John Wiley and Sons, Inc., 1953.

Keyfitz, N. "Estimates of sampling variance where two units are selected from each stratum." Journal of the American Statistical Association, 52, (December), 1957, pp. 503-510.

Kish, L. and Frankel, M. R. "Inference from complex samples." Journal of the Royal Statistical Society, No. 1, 1974, pp. 1-37.

McCarthy, P. J. "Replication: an approach to the analysis of data from complex surveys." National Center for Health Statistics, Series 2, No. 14, 1966.

U. S. Bureau of the Census. "The Current Population Survey—a report on methodology." Technical Paper No. 7, 1963.

__________________________. "The Current Population Survey—design and methodology." Technical Paper No. 40, 1978.