SOME PRACTICAL CONSIDERATIONS IN CHOOSING A SURVEY'S SAMPLE DESIGN

Jeffrey Blaha, Sylvia Johnson-Herring, Sharon Krieger, Martine Ferguson
U.S. Bureau of Labor Statistics
2 Mass. Ave. NE, Room 3650, Washington, DC 20212
(Blaha.Jeffrey@bls.gov)

Any opinions expressed in this paper are those of the authors, and do not constitute policy of the Bureau of Labor Statistics.

Key Words: cluster sampling, variance, cost model, housing unit string

1. Introduction

The Consumer Expenditure (CE) Survey is a national household survey conducted by the Bureau of Labor Statistics (BLS) to find out how Americans spend their money. The survey's sample design is updated approximately every ten years. The current sample design consists of selecting a stratified random sample of geographic areas (called PSUs) around the United States, and then within each PSU selecting a systematic sample of housing units for participation in the survey. This design ensures that every segment of the population is included in the sample. However, the sample design also results in high survey costs, particularly in relation to the amount of travel required by data collectors. In this paper we summarize research conducted at BLS to determine the feasibility of sampling clusters of contiguous housing units in order to reduce survey costs while still keeping variances small.

A significant component of the research is to conduct a cost analysis in order to estimate potential savings in travel costs as a result of cluster sampling. At this time, analysis in this area is in the early phases of development and will only be discussed briefly in this paper.

2. Background

The CE Survey is a national household survey conducted by the BLS to find out how Americans spend their money. Data for the survey are collected by the Bureau of the Census (BOC) under contract with the BLS. One of the primary uses of the data is to provide expenditure weights for the Consumer Price Index.

The CE Survey consists of two separate surveys, the Quarterly Interview (CEQ) and Diary (CED) surveys. The purpose of the CEQ is to obtain detailed expenditure data on large items such as property, automobiles or major appliances; or expenses that occur on a regular basis, such as rent, utility bills, or insurance premiums. The purpose of the CED is to obtain detailed expenditure data on small, frequently purchased items such as food and apparel.

3. Current Sample Design

The selection of households for the survey begins with the definition and selection of primary sampling units (PSUs), which consist of counties (or parts thereof), groups of counties, or independent cities. The PSUs are classified into the following four size categories (Johnson-Herring, et. al., 2002):

- "A" PSUs, which are Metropolitan Statistical Areas (MSAs) with a population of 1.5 million or greater
- "B" PSUs, which are MSAs with a population less than 1.5 million
- "C" PSUs, which are nonmetropolitan areas used in the CPI
- "D" PSUs, which are nonmetropolitan areas not used in the CPI, often referred to as "rural" PSUs

The sampling frame (i.e., the list from which housing units are chosen) for this survey is generated from the BOC's decennial census file, which is augmented by a sample drawn from new construction permits. The population of interest is the total U.S. civilian population. Within this framework, the eligible population includes all civilian noninstitutional persons (for example, those living in homes, condominiums, or apartments) and all people residing in group quarters (such as those living in housing facilities for students and workers). Military personnel living on base are not included.

CE shares this sampling frame with other surveys conducted by the BOC. These surveys include the Current Population Survey (CPS), the National Crime Survey (NCS), the National Health Interview Survey (NHIS), the American Housing Survey (AHS), and the Survey of Income and Program Participation (SIPP). In the sample selection process, housing units are selected for a particular survey, and then those housing units are removed from the sampling frame prior to selecting samples for the subsequent surveys. Thus housing units cannot be visited for more than one survey.

Through research many of these surveys have determined that clustering households is beneficial in the sense that the reduction in survey costs outweighs the expected increase in variances that result from using a clustered sample design. Part of the research involved determining the optimal cluster size for the particular survey, which typically was a cluster size of 4 housing units.

For each survey, a certain number of samples is designated over the 10-year span for which the BOC's decennial census file serves as the sampling frame. The sample size for each of those surveys is allocated across the PSUs based on PSU size. Then the sampling frame is sorted geographically within the PSUs. Once sorted, strings of housing units equal in length to the product of the cluster size times the number of samples that will be designated in the upcoming 10-year span are chosen from the sampling frame. The samples are selected independently within each PSU.

For example, CPS uses 19 samples over a 10-year span (to accommodate its rotation system and the phasing in of new designs) with a cluster size of 4 housing units. As a result, strings of 76 (=19 × 4) housing units are selected, with groups of 4 housing units in the string being assigned to one of the particular samples (BOC and BLS, et. al., 2000). Table 1 illustrates this method.

Table 1.

CPS Sample Designation (SD) Assignment within a Selected Housing Unit (HU) String

HU	SD	HU	SD	HU	SD	HU	SD
1	A1	20	A5	39	A10	58	A15
2	A1	21	A6	40	A10	59	A15
3	A1	22	A6	41	A11	60	A15
4	A1	23	A6	42	A11	61	A16
5	A2	24	A6	43	A11	62	A16
6	A2	25	A7	44	A11	63	A16
7	A2	26	A7	45	A12	64	A16
8	A2	27	A7	46	A12	65	A17
9	A3	28	A7	47	A12	66	A17
10	A3	29	A8	48	A12	67	A17
11	A3	30	A8	49	A13	68	A17
12	A3	31	A8	50	A13	69	A18
13	A4	32	A8	51	A13	70	A18
14	A4	33	A9	52	A13	71	A18
15	A4	34	A9	53	A14	72	A18
16	A4	35	A9	54	A14	73	A19
17	A5	36	A9	55	A14	74	A19
18	A5	37	A10	56	A14	75	A19
19	A5	38	A10	57	A15	76	A19

The CEQ currently designates 24 samples over a 10-year span (for similar reasons as the CPS). Since CEQ currently uses systematic sampling, which can be thought of as systematic cluster sampling with a cluster size of 1. Therefore, strings of 24 (=24 × 1) housing units are selected from the sampling frame. Once a string is selected, it is removed from the sampling frame prior to selecting the samples for the other BOC surveys.

The CED sample selection process is identical to that of CEQ, except that 22 samples are designated over the 10-year span instead of 24 samples.

4. Sample Selection for CEQ Variance Analysis

To analyze the potential impact of cluster sampling on the CEQ's variance, we ran simulations using four years of data from the CEQ production database (1998-2001). The database for this time period contains 113,556 usable interviews, from which samples of 5,000 usable interviews were desired for the analysis. To obtain 5,000 usable interviews, a sample of 7,875 Consumer Units[1] (CUs) was selected from the database, and 36.5% of the selected CUs were randomly removed from the sample to simulate the nonresponse process, leaving 5,000 CUs with completed interviews.[2] The sample mean was then calculated from these participating CUs, with the variable of interest being the total dollar value of all expenditures reported by a CU.

This process was repeated 1,000 times, allowing estimates of the mean, bias, and variance to be made. Comparisons were then made between these values using systematic sampling (cluster size of 1) and those generated with systematic cluster sampling (cluster size greater than 1).

Prior to sample selection, the database was sorted similar to the way sampling frames are sorted in the current sample design. The sort variables are as follows (in order):

1. PSU
2. Urban/Rural Code

[1] A Consumer Unit is a group of people living together who share major expenditures, such as rent. In most cases a Consumer Unit is equivalent to a housing unit.

[2] 63.5% of the addresses visited by CEQ data collectors in 1998-2001 yielded usable interviews, or equivalently, 36.5% of the addresses did not yield usable interviews.

3. State
4. County
5. Stratum (CE Stratification Code)
6. Census tract
7. CU

The CUs within each cluster are geographically close to each other as a result of the order of the geographic sorting variables listed above.

All PSUs from the current CE sample design were included in the simulations, with the sample of 7,875 CUs being allocated to the PSUs proportional to each PSU's current sample size. Samples were drawn independently within each PSU.

5. CEQ Variance Analysis

Drawing 1,000 independent random samples from the data with 7,875 housing units in each sample, and then randomly removing 36.5% of the sample for nonresponse produced the results shown in Table 2.

Table 2.

CEQ: The Effect of Using Various Cluster Sizes

Cluster Size	Population mean, μ	Average of the 1,000 random samples, $\bar{x}$	Bias, $\bar{x}-\mu$	Standard Error, $SE(\bar{x}_i)$	Change in SE relative to a cluster size of 1
1	$8,580.65	$8,582.97	$ 2.32	$100.80	1.00
2	$8,580.65	$8,587.37	$ 6.73	118.90	1.18
3	$8,580.65	$8,584.80	$ 4.16	125.22	1.24
4	$8,580.65	$8,577.85	$-2.80	126.83	1.26
5	$8,580.65	$8,582.54	$ 1.90	138.32	1.37
6	$8,580.65	$8,581.37	$ 0.73	147.83	1.47
7	$8,580.65	$8,579.23	$-1.42	153.52	1.52
8	$8,580.65	$8,581.91	$ 1.26	149.53	1.48
9	$8,580.65	$8,589.54	$ 8.90	152.11	1.51
10	$8,580.65	$8,575.69	$-4.96	160.21	1.59

Here the population mean μ comes from all 113,556 CUs in the database, $\bar{x}=\frac{1}{1000}\sum_{i=1}^{1000}\bar{x}_i$, where $\bar{x}_i$ is the sample mean of the i^{th} random sample, and $SE(\bar{x}_i)=\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\bar{x}_i-\mu)^2}$.

Based on all 113,556 interviews in the database, the mean quarterly expenditure per CU was $8,580.65. Using the CEQ's current method of systematic sampling (cluster size of 1), the average value of the 1,000 different random samples was $8,582.97, indicating a bias of $2.32. The standard error of the sample mean (i.e., the standard deviation of the 1,000 different sample means) was $100.80.

The other rows of the table can be interpreted similarly, with each row representing the results generated from systematic cluster sampling using the indicated cluster sizes.

The last column in Table 2 compares the standard error of the sample mean for different cluster sizes to the standard error obtained from CEQ's current method of systematic sampling (cluster size of 1). These results show that as the cluster sizes increase, so do the standard errors. For example, going from a systematic sample (cluster size of 1) to a systematic cluster size of 2 increases the standard error by approximately 18%, and going to a systematic cluster size of 3 increases the standard error by approximately 24%.

The increased standard errors can most likely be attributed to the similarity of expenditure patterns of the housing units within each cluster. One may allow that this axiom holds true especially for large-ticket items or recurring major expenditures such as those collected in the CEQ. The data above do not provide strong support for using cluster sampling in the CEQ.

An additional factor in determining whether or not to sample clusters of housing units in a survey is the actual length of the survey. If the survey requires a considerable amount of time to complete (such as the CEQ) and cluster sampling is employed, a respondent may speak negatively about the survey's length to his or her neighbor. As a result, the neighbor may be more reluctant to respond when contacted by the CE data collector than if that person had no prior knowledge of the survey. This issue was raised by experienced data collectors for different surveys. The average length of a productive CEQ interview is approximately 90 minutes, which is longer than most, if not all, BOC surveys.

The length of the CEQ also has an impact on determining the optimal cluster size to employ in the sample design (if cluster sampling was implemented). Due to the length of the survey, a CE data collector would typically only visit 2 or 3 housing units in one day. As a result, there may be no benefit in terms of reducing travel costs by having clusters larger than 2 or 3 housing units.

6. CED Variance Analysis

Simulations similar to those run for the CEQ survey were run on the CED data so that comparisons could be made between the two surveys.

Four years of data from the CED production database (1998-2001) were used as the source database. This database contains 57,707 interviews. As with the CEQ analysis, obtaining samples of 5,000 usable interviews for the CED analysis was the goal. However, since in the CED a participation rate of 60% was assumed, a sample of 8,333 CUs was selected and 40% of the selected CUs were randomly removed from the sample, leaving 5,000 usable interviews for the analysis. All other aspects of the sampling process were identical to those in the CEQ analysis. Again, samples were analyzed for systematic clustering using different cluster sizes.

Drawing 1,000 independent random samples with 8,333 CUs and then randomly removing 40% of them produced the results shown in Table 3.

Table 3.

CED: The Effect of Using Various Cluster Sizes

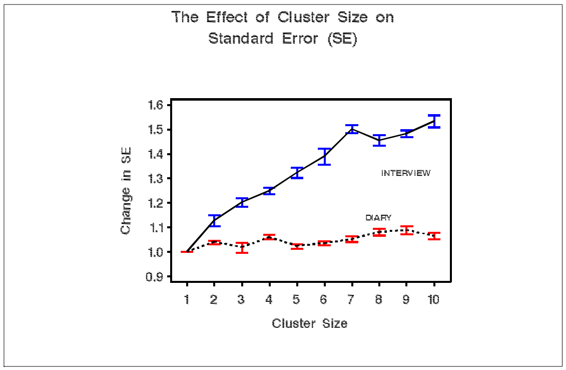
Cluster Size	Population mean, μ	Average of the 1,000 random samples, $\bar{x}$	Bias, $\bar{x}-\mu$	Standard Error, $SE(\bar{x}_i)$	Change in SE relative to a cluster size of 1
1	$572.73	$571.92	$–0.81	$20.88	1.00
2	572.73	572.85	0.12	21.81	1.04
3	572.73	573.65	0.92	22.17	1.06
4	572.73	572.21	–0.52	21.87	1.05
5	572.73	572.18	–0.55	21.29	1.02
6	572.73	573.60	0.87	21.25	1.02
7	572.73	574.35	1.62	22.55	1.08
8	572.73	572.97	0.24	22.77	1.09
9	572.73	573.37	0.64	21.92	1.05
10	572.73	572.86	0.13	22.86	1.09

Here the population mean μ comes from all 57,707 CUs in the database, $\bar{x}=\frac{1}{1000}\sum_{i=1}^{1000}\bar{x}_i$, where $\bar{x}_i$ is the sample mean of the i^{th} random sample, and $SE(\bar{x}_i)=\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\bar{x}_i-\mu)^2}$.

These results in Table 3 show that systematic clustering produces unbiased estimates of the mean expenditure for all cluster sizes. Also, the percent change in the standard error as the cluster size increases is small, especially for clusters of size 2, 3, and 4. For example, going from systematic sampling (cluster size of 1) to systematic clustered sampling with a cluster of size 2 increases the standard error by 4%, going to a cluster size of 3 increases the standard error by 6%, and going to a cluster size of 4 increases the standard error by 5%.

While housing units within clusters are expected to be homogenous and therefore have similar spending patterns, the domain of the CED allows the use of cluster sampling with only a slight increase in the variance. One may surmise that the expenditures reported for the more frequently purchased, less expensive items collected in the CED are more dependent upon the preferences of the household. Also, the domain of the CED is not driven by the more costly and recurring expenditures, such as rents and utilities.

The simulations for both CEQ and CED show that while the standard error increases as a function of cluster size, an unexpected decline occurs for CEQ cluster size of 8 and though less pronounced, for CED cluster sizes of 5 and 6. Repeated simulations of the clustered sampling methodology were run on the data to validate that this observation was not an anomaly of the data. The following graph shows the results of these repeated simulations.



The graphical representation of the effect of clustering on the standard error of the CED clearly supports the implementation of cluster sampling techniques. In addition, as opposed to the CEQ, the CED is on average a 30-minute survey, so issues related to the length of the survey do not apply to the CED.

7. Cost Analysis

As discussed in the Introduction, at this point analysis of potential cost savings is in the early stages of development. Researchers at the BOC have developed an effective model that simulates interviewer travel costs for different cluster sizes (Bienias, et. al., 1990). The model allows users to input a wide variety of parameters. The parameters include the following:

- Cluster size
- Number of clusters in a PSU
- Maximum number of trips a data collector can make
- Length of an interview
- Probability of obtaining a complete interview at any given contact
- Size of a PSU

- Varying population densities within a PSU (urban versus rural areas)
- Rate of travel

This model appears to be very applicable to surveys like the CEQ and CED. In fact, at the time of development of this model, the creators used it to estimate the potential travel cost savings that would result by changing from the current non-clustered sample design to a sample design with cluster sizes of 2 for the CED. The results concluded an estimated reduction in the range of 20-26%. CE researchers intend to adapt this model to known characteristics of the current CED and test various cluster sizes. No cost analysis will be done for the CEQ since cluster sampling is highly unlikely to be incorporated for that survey.

8. Conclusion

The CE Survey currently uses systematic cluster sampling but in its degenerate case (cluster size of 1). Research indicates that changing the CEQ sampling methodology from systematic sampling to systematic cluster sampling with a cluster size larger than 1 would significantly increase the standard error of the survey estimates. In addition, the length of the CEQ interview (average of 90 minutes) has an adverse impact on cluster sampling. As a result, using clustered sampling in the CEQ is not recommended.

While research indicated that the variance of the CEQ estimates is significantly increased when systematic cluster sampling is employed, the same is not true of the CED estimates. The expenditure estimates for the domain of the CED are seemingly not affected by the homogeneity of the households within the clusters. Since the increase in variance is minimal for cluster sizes of 2, 3 or 4 housing units, any of these could be implemented in the sample selection methodology. Analysis of other BOC surveys which at least partially use cluster sampling in their sample designs indicates that the cluster size normally employed is 4 housing units. In addition, the length of the CED interview (average of 30 minutes) does not negatively impact cluster sampling.

The focus of future research will be to develop an effective model to estimate potential savings in travel costs for various cluster sizes within the CED. Additional research will also include analysis of within-cluster correlations for Diary expenditures at disaggregate item levels.

9. References

Johnson-Herring, S., Krieger, S., and Swanson, D., “Determining Within-PSU Sample Sizes for the Consumer Expenditure Survey,” *Proceedings of the Section on Government Statistics*, American Statistical Association, 2002.

U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, “Technical Paper 63: Current Population Survey-Design and Methodology” (Chapter 3), Washington, DC: Government Printing Office, 2000.

Bienias, J., Sweet, E., and Alexander, C., “A Model for Simulating Interviewer Travel Costs for Different Cluster Sizes,” *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1990.