DETERMINING WITHIN-PSU SAMPLE SIZES FOR THE CONSUMER EXPENDITURE SURVEY

Sylvia Johnson-Herring, Sharon Krieger, David Swanson
U.S. Bureau of Labor Statistics
2 Mass. Ave., NE, Room 3650, Washington, DC 20212

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

Key Words: variance, sample size, sample allocation, constrained least squares

Introduction

The Consumer Expenditure 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, and at that time many decisions need to be made, such as the number of geographic areas in which to collect data, and the number of households from which to collect data in each area. In this paper we describe an automated method of determining the number of households to sample in the selected geographic areas.

Background

The Consumer Expenditure 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 under contract with the BLS. One of the primary uses of the data is to provide expenditure weights for the Consumer Price Index (CPI).

The Consumer Expenditure Survey consists of two separate surveys, the Diary (CED) and Quarterly Interview (CEQ) surveys. The purpose of the CED is to obtain detailed expenditure data on small, frequently purchased items such as food and apparel. 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. This paper focuses on the CEQ.

Primary Sampling Units

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 sample of PSUs currently used in the survey consists of 105 geographic areas, of which 87 urban geographic areas[1] have also been selected by the BLS for the CPI. The 105 PSUs are classified into four size categories:

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

The 31 "A" PSUs are self-representing, and the 74 "B", "C", and "D" PSUs are non-self-representing. Examples of "A" PSUs are the Boston, Chicago, and San Francisco metropolitan areas. Examples of "B" PSUs are the Hartford, Connecticut; the Dayton, Ohio; and the Provo, Utah metropolitan areas. Examples of "C" PSUs are the Morristown-Jefferson City, Tennessee, and Mount Vernon, Illinois nonmetropolitan areas. An example of a "D" PSU is Caribou-Presque Isle, Maine, a rural area composed of parts of Aroostook County.

For some analyses the "B", "C", and "D" PSUs are grouped together by their region-size classes. There are 12 region-size classes which are created by cross-classifying the four regions of the country (Northeast, Midwest, South, and West) with the three size classes (B, C, D). When these region-size classes are used, they are treated just like the other self-representing PSUs. No "C" PSUs were selected in the Northeast region, so there are actually 11 region-size classes. Hence the CEQ can be thought of as having 42 self-

[1] The new official terms for "urban" and "rural" areas are "CBSA" and "Non-CBSA" areas, but in this memorandum we will use the terms "urban" and "rural" to denote them. "*A core-based statistical area (CBSA)* consists of the county or counties associated with at least one core of 10,000 or greater population, plus adjacent counties having a high degree of social and economic integration with the core(s) as measured by commuting ties." *Non-CBSA* areas are areas outside CBSAs. Non-CBSA areas "display a high degree of rurality."

representing geographic areas, 31 "A" PSUs plus 11 region-size classes for the smaller PSUs.

Within-PSU Sample Sizes

In the CEQ's current sample design, usable data are collected from 7,760 households in each calendar quarter of the year: 4,260 usable interviews in the "A" PSUs, and 3,500 usable interviews in the "B," "C," and "D" PSUs. In order to guarantee that enough data are collected to satisfy publication requirements for all 42 geographic areas, the sample of 7,760 households is allocated in a way that at least 120 usable interviews are obtained in each of the 38 geographic areas used in the CPI, with no minimum number of usable interviews required in the 4 "D" geographic areas.

Thus the sample allocation problem is to allocate 7,760 households to the 42 geographic areas in a way that the following constraints are satisfied:

- 4,260 usable interviews are collected in the 31 "A" PSUs
- 3,500 usable interviews are collected in the 11 "B," "C," and "D" geographic areas
- 120 or more usable interviews are collected in each of the 38 geographic areas used in the CPI

CEQ staff recently re-evaluated the minimum sample size requirement of 120 usable interviews to determine whether it is still an appropriate number. One of the results of the re-evaluation was the development of a new automated method of allocating the nationwide sample to individual geographic areas. The new method allowed repeated analyses to be conducted easily using different minimum sample size requirements. The method involved setting up the sample allocation problem as a mathematical optimization problem, and then using SAS software to solve the optimization problem.

"Target" versus "Required" Sample Sizes

In the past there were various interpretations of the word "required" in the phrase "minimum required sample size." At times the requirement that at least 120 usable interviews be obtained was interpreted as a "target" sample size, meaning that the expected number of usable interviews should be at least 120,

$$\mathrm{E}(x_i) \geq 120$$

while at other times it was interpreted as a "required" sample size, meaning that there should be a very high probability that at least 120 interviews be obtained,

$$\mathrm{P}\{x_i \geq 120\} \geq 0.95$$

where x_i is the number of usable interviews collected in geographic area = i.

For example, under the first interpretation ("target" sample size), we would have to visit 185 households in each quarter of the year in order to collect 120 usable interviews in the Boston metropolitan area, assuming that usable interviews are obtained at 65% of the residential addresses in the CEQ's sample.[2]

$$\mathrm{E}(x_i) = 185 \times 0.65 = 120$$

However, under the second interpretation ("required" sample size) we would have to visit 202 households in order to be 95% certain of getting at least 120 usable interviews, again assuming a 65% survey participation rate.

$$\mathrm{P}\{x_i \geq 120\} = \sum_{k=120}^{202} \binom{202}{k} 0.65^k (1-0.65)^{202-k} = 0.95$$

Table 1 shows the difference in the sample size that would be needed when using a "target" as opposed to a "required" minimum number of usable interviews. The number of selected addresses needed to achieve a "target" minimum sample size is approximately 10% less than that needed for a "required" sample size.

Table 1.
Minimum Sample Size
(assuming a 65% survey participation rate)

Number of sample households (n)	Expected number of usable interviews (=0.65n)	95% Confidence Interval
Target sample size (2-sided 95% confidence interval)		
62	40	[33, 47]
92	60	[51, 69]
123	80	[70, 90]
154	100	[88, 112]
185	120	[107, 133]
215	140	[126, 154]
Required sample size (1-sided 95% confidence interval)		
72	47	[40, +∞)
105	68	[60, +∞)
137	89	[80, +∞)
170	110	[100, +∞)
202	131	[120, +∞)
234	152	[140, +∞)

These estimates were produced using formulas from the binomial distribution for the mean and variance of the number of usable interviews,

$$\mu = E(x_i) = 0.65\,n$$

$$s^2 = V(x_i) = 0.65(1-0.65)\,n$$

[2] Approximately 15% of the residential addresses selected for the CEQ survey are ineligible for the survey, and 20% do not participate in the survey due to refusal or no one being home. This leaves 65% of the sample who participate in the survey.

and the normal distribution was used to approximate the binomial distribution to estimate a 95% confidence interval on the number of usable interviews:

1-sided confidence interval: $[\ \mu - 1.64\sigma\ ,\ +\infty\ )$

2-sided confidence interval: $[\ \mu - 1.96\sigma\ ,\ \mu + 1.96\sigma\ ]$

After some discussion, it was decided that "target" sample sizes would be satisfactory since the width of the 2-sided confidence intervals are relatively small.

Setting Up the Optimization Problem

As mentioned earlier, the current CEQ sample design calls for allocating 7,760 households to the 42 geographic areas in a way that the following constraints are satisfied:

- 4,260 usable interviews are collected in the 31 "A" PSUs
- 3,500 usable interviews are collected in the 11 "B", "C", and "D" geographic areas
- 120 or more usable interviews are collected in each of the 38 geographic areas used in the CPI

This can be written in mathematical terms as the following:

- $x_1 + x_2 + \cdots + x_{31} = 4{,}260$
- $x_{32} + x_{33} + \cdots + x_{42} = 3{,}500$
- $x_i \geq 120$ for i=1,2,…,38

where x_i is the number of usable interviews collected in geographic area=i.

The objective of the CEQ's sample design is to allocate the nationwide sample of households to individual geographic areas in a way that minimizes the standard error of the CEQ's published expenditure estimates at the national level. Allocating the sample proportional to the population of each geographic area comes very close to achieving that goal. Although allocating the sample proportional to population does not minimize the nationwide standard error, it is a very simple sample design that is known to achieve "near minimization." We chose to implement this method because of its simplicity and its "near optimal" properties.

For some of the geographic areas with small populations (e.g., Anchorage and Honolulu) the requirement that at least 120 usable interviews be collected during each calendar quarter of the year conflicts with the objective of allocating the sample proportional to population. For example, the Anchorage metropolitan area has approximately 0.09% of the U.S. population, and allocating the 7,760 usable interviews directly proportional to population would give Anchorage only enough addresses to obtain 7 usable interviews. This conflicts with the publication requirement that at least 120 usable interviews be collected in each geographic area.

Since the objective cannot be achieved, we decided to allocate the sample as close to population proportionality as possible. This involved setting up a least squares problem in which the squared difference between each geographic area's proportion of the population and its proportion of the sample is computed, and then the sum of those 42 squared differences is minimized.

Thus the sample allocation problem is to solve the following constrained least squares problem:

Given values of n, p_i , p, find values of n_i that

Minimize $$\sum_{i=1}^{42}\left(\frac{n_i}{n} - \frac{p_i}{p}\right)^2$$

Subject to
$$n_1 + n_2 + \cdots + n_{31} = 4{,}260$$
$$n_{32} + n_{33} + \cdots + n_{42} = 3{,}500$$
$$n_i \geq 120 \quad \text{for } i=1,2,\ldots,38$$
$$n_i \geq 0 \quad \text{for } i=39,\ldots,42$$

where

n_i = number of housing units assigned to geographic area=i

n = number of housing units nationwide (n = 7,760)

p_i = population of geographic area=i

p = population in all geographic areas($p = p_1 + p_2 + \cdots + p_{42}$)

Solving the Optimization Problem

The sample allocation problem described above can be seen to have both equality and inequality constraints. This creates a practical problem because optimization problems with equality constraints are usually solved with different techniques than those with inequality constraints.

Least squares problems with equality constraints are usually solved with linear algebra and linear regression theory, while problems with inequality constraints are usually solved with iterative search techniques.

Fortunately, the SAS procedure PROC NLP (**N**on**L**inear **P**rogramming) can handle both kinds of constraints simultaneously. An example using PROC NLP to solve the problem above is given at the end of this paper.

Estimating the CEQ's Standard Error

The CEQ's variance resulting from the sample allocation process described above was estimated using the following formula:

$$V(\bar{x}) = V\left(\sum_{i=1}^{42}\left(\frac{p_i}{p}\right)\bar{x}_i\right) = \sum_{i=1}^{42}\left(\frac{p_i}{p}\right)^2 V(\bar{x}_i) = \sum_{i=1}^{42}\left(\frac{p_i}{p}\right)^2 \frac{s^2}{n_i}$$

where

$\bar{x}_i$ = sample mean of geographic area = i

$\bar{x}$ = sample mean of the nation,

$$\bar{x} = \frac{\sum_{i=1}^{42} p_i \bar{x}_i}{\sum_{i=1}^{42} p_i} = \frac{\sum_{i=1}^{42} p_i \bar{x}_i}{p} = \sum_{i=1}^{42}\left(\frac{p_i}{p}\right)\bar{x}_i$$

s^2 = expenditure variance of a randomly selected household

The CEQ's variance under the proposed sample allocation method is estimated by substituting the values of n_i obtained from the optimization problem (the output of "PROC NLP") into the formula

$$V(\bar{x}) = \sum_{i=1}^{42}\left(\frac{p_i}{p}\right)^2 \frac{s^2}{n_i}.$$

Then the standard error is computed by taking the square root of the variance.

$$\text{SE} = \sqrt{\sum_{i=1}^{42}\left(\frac{p_i}{p}\right)^2 \frac{s^2}{n_i}}$$

This allows comparisons to be made to the current method of sample allocation. The value of σ does not have to be known because the change in standard error is the number in which we are interested, and when the ratio of two estimates of the standard error is taken (to compare the standard error of using, say, 80 as the minimum sample size compared to 120) the value of σ in the numerator and σ in the denominator cancel out with each other.

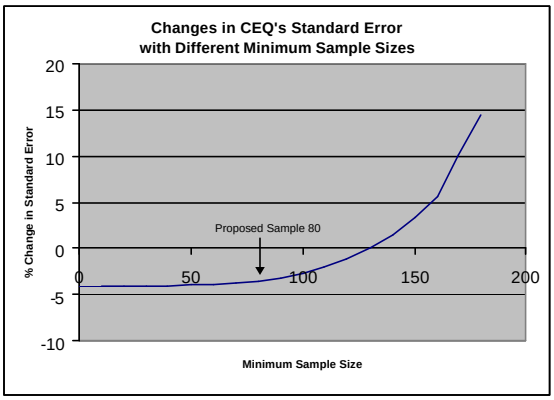
CEQ's Standard Error with Different Minimum Sample Size Requirements

After allocating the CEQ's nationwide sample to individual geographic areas using PROC NLP, the percent change in the CEQ's standard error was computed for various minimum target sample sizes. Standard errors were computed at the All Items, All U.S. level. The baseline used in the comparison was the current sample allocation. The current minimum target sample size is around 120, but for technical reasons it is not exactly equal to 120. The results of the comparisons are shown in Table 2.

Table 2.
The Effect of Minimum Target Sample Sizes on CEQ's Standard Error

Minimum Target Sample for each PSU	Change in CEQ's Standard Error (%)
0	–4.16
10	–4.16
20	–4.15
30	–4.10
40	–4.04
50	–3.96
60	–3.88
70	–3.74
80	**- 3.54**
90	–3.21
100	–2.72
110	–2.04
120	**- 1.14**
130	+0.06
140	+1.45
150	+3.28
160	+5.63
170	+10.07
180	+14.41

From this table it can be seen that the CEQ's standard error is minimized when the sample is allocated directly proportional to population, i.e., when 0 is the minimum number of usable interviews required in each geographic area. Reducing the target number of usable interviews from 120 to 0 would reduce CEQ's standard error by 4.16%. The other extreme is where the sample is divided equally among all the geographic areas, i.e., 180 usable interviews per area. Increasing the target number of usable interviews from 120 to 180 would increase CEQ's standard error by 14.41%.



Reducing the minimum target number of usable interviews from 120 to 80 per geographic area would reduce the CEQ's standard error by 3.54%. The

above graph of Table 2 shows that nearly all of the reduction in the CEQ's standard error is achieved by reducing the minimum target sample size to 80, and that little further reduction is achieved by reducing the minimum target sample size below 80.

A More Detailed Analysis

Four additional sample allocation methodologies were examined to study the effect that the minimum target sample size of 80 would have on the standard error of the CEQ and CPI estimates. They are:

Method 1. The number of households allocated to each geographic area is as close to population proportionality as possible, but with the added constraint that a minimum number of usable interviews be collected in each of the 38 geographic areas used in the CPI.

Method 2. The number of households allocated to each geographic area is as close to population proportionality as possible, but with the added constraints that (1) a minimum number of usable interviews be collected in each of the 38 geographic areas used in the CPI, (2) the number of households in each of the four "D" geographic areas is less than or equal to 100.

Method 3. The number of households allocated to each geographic area is as close to population proportionality as possible, but with the added constraints that (1) a minimum number of usable interviews be collected in each of the 38 geographic areas used in the CPI, and (2) the number of households in each of the four "D" geographic areas is exactly equal to 100.

Method 4. The number of households allocated to each geographic area is as close to population proportionality as possible, but with the added constraints that (1) a minimum number of usable interviews be collected in each of the 38 geographic areas used in the CPI, and (2) the total number of households in the four "D" geographic areas is equal to 400.

The percent changes in standard error relative to the current allocation were computed for the four proposed methods of sample allocation at the Total U.S. (CE population) and Urban only (CPI population) levels. The CEQ's actual 2000 sample allocation is used as the basis of comparison.

Table 3.
The Effect on CEQ's Standard Error when Minimum Target Sample Size is 80

Proposed Method	Total U.S. (CE population)	Urban only (CPI population)
Method 1	–5.56	–1.21
Method 2	–1.00	–4.00
Method 3	–0.95	–3.86
Method 4	**- 3.54**	**- 3.86**

Table 3 shows that the reduction in standard error for the CPI population closely matches the reduction in standard error for the CE population when using Method 4 to allocate the sample. The standard error is reduced for both surveys by approximately the same amount.

Table 4.
The Current and Proposed Sample Allocations for "A" PSUs in the West Region, and its Effect on CEQ's Standard Error, with Minimum Target Sample Size 80

PSU	Population	Current Sample Size	Proposed Sample Size	Change in SE (%)
A419 Los Angeles	8,863,164	231	290	–10.81
A420 Greater L.A.	5,668,365	152	187	–9.88
A422 SanFrancisco	6,253,311	158	206	–12.44
A423 Seattle	2,970,328	119	100	+9.08
A424 San Diego	2,498,016	104	85	+10.78
A425 Portland	1,793,476	130	80	+27.48
A426 Honolulu	836,231	112	80	+18.32
A427 Anchorage	226,338	125	80	+25.00
A429 Phoenix	2,238,480	132	80	+28.45
A433 Denver	1,980,140	121	80	+22.98
Total U.S.	240,218,238	7,760	7,760	–3.54

Table 4 shows the current and proposed sample sizes for "A" PSUs in the West region after using PROC NLP to allocate the sample using Method 4. The table also shows the effect of using this method of sample allocation on CEQ's standard error. PSUs with populations under 4 million will have their sample sizes reduced, while PSUs with populations over 4 million will have their sample sizes increased. This shows that reducing the minimum target number of usable interviews collected quarterly in each urban geographic area to 80 would reduce the standard error in the larger "A" PSUs and increase the standard error in the smaller "A" PSUs. Using Method 4 reduces the standard error for the nationwide sample by approximately 3.54%. The standard errors increase in some geographic areas and decrease in others, but the standard error for the nation as a whole decreases by 3.54%.

Conclusion

Finding a way to allocate the CEQ's nationwide sample of households to individual geographic areas is facilitated by automating the sample allocation process. Estimates of standard error can be computed from the output of PROC NLP and analyzed to find an allocation method which minimizes the nationwide standard error. Our research shows that allocating the CEQ's sample using Method 4 reduces the standard error of the published expenditure estimates at the national level by about 3.54%. This automated process to select an optimal sample allocation method can be effective in future sample redesigns.

Appendix
Automating the Sample Allocation Process

Below is the optimization problem for the sample allocation using Method 4, along with a SAS program that solves it.

Given values of n, p_i, p, find values of n_i that

Minimize $$\sum_{i=1}^{42}\left(\frac{n_i}{n}-\frac{p_i}{p}\right)^2$$

Subject to
$$n_1+n_2+\cdots+n_{38}=7{,}360$$
$$n_{39}+n_{40}+n_{41}+n_{42}=400$$
$$n_i \geq 80 \text{ for } i=1,2,\ldots,38$$
$$n_i \geq 0 \text{ for } i=39,\ldots,42$$

Where

- n_i = number of housing units assigned to geographic area=i
- n = number of housing units nationwide (n = 7,760)
- p_i = population of geographic area=i
- p = population in all geographic areas($p = p_1 + p_2 + \cdots + p_{42}$)

```
*******************************************
* COMPUTE THE SQUARED DIFFERENCE BETWEEN EACH *
* AREA'S PROPORTION OF THE POPULATION & ITS   *
* PROPORTION OF THE SAMPLE.                   *
*******************************************;

%MACRO MAC1;
SUM_POP = SUM(OF POP1-POP42);
%DO I=1 %TO 42;
    SQR&I = ((N&I/7760) - (POP&I/SUM_POP))**2;
%END;
%MEND MAC1;

***********************************************
* SOLVE A CONSTRAINED LEAST SQUARES PROBLEM TO  *
* FIND THE NUMBER OF HOUSING UNITS IN EACH PSU  *
* THAT MINIMIZES THE SUM OF SQUARED DIFFERENCES *
***********************************************;

PROC NLP DATA=POP DATA(KEEP=POP1-POP42) NOPRINT
     OUT=RESULTS(KEEP=N1-N42)

     /* CONVERGENCE  CRITERIA */
     GCONV=1E-15 FCONV2=1E-15 MAXITER=100000;

     /* DECISION  VARIABLES */
     DECVAR N1-N42;

     /* COMPUTE THE SQUARED DIFFERENCES */
     %MAC1;

     /* SUM THE SQUARED DIFFERENCES */
     F1=SUM(OF SQR1-SQR42);

     /* FUNCTION TO BE MINIMIZED */
     MIN F1;

     /* PROBLEM CONSTRAINTS */
     BOUNDS N1-N38>=80, N39-N42>=0;
     NLINCON F2=7360, F3=400;
     F2=SUM(OF N1-N38);
     F3=SUM(OF N39-N42);
RUN;
```