Determining Area Sample Sizes for the Consumer Expenditure Survey

SYLVIA A. JOHNSON-HERRING
SHARON KRIEGER
DAVID SWANSON

Sylvia A. Johnson-Herring is a mathematical statistician in the Division of Price Statistical Methods, Consumer Expenditure Surveys, Bureau of Labor Statistics.

Sharon Krieger is a mathematical statistician in the Division of Price Statistical Methods, Consumer Expenditure Surveys, Bureau of Labor Statistics.

David Swanson is Branch Chief, Division of Price Statistical Methods, Consumer Expenditure Surveys, Bureau of Labor Statistics.

The Consumer Expenditure Survey (CE) is a national household survey conducted by the U.S. Bureau of Labor Statistics (BLS) to find out how Americans spend their money. The survey's sample design, based on the decennial census, is updated approximately every 10 years. 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. This article describes a new method for making these decisions, one that has been incorporated in the sample design to be introduced in 2005.

Background

The CE is used to produce the most accurate estimate of consumer expenditures possible at the national level. The U.S. Consumer Price Index (CPI) program relies on CE data to produce inflation estimates. The most comprehensive CPI is based on the expenditure patterns of consumers in urban and metropolitan areas and is denoted CPI-U. The CPI-U population represents about 87 percent of the total U.S. population. The CE is designed to balance the goals of the CE and CPI programs. These goals compete with each other when BLS allocates the CE's nationwide sample of households to geographic areas covered by the two programs.

The number of households in the CE's national sample is determined by the survey's data collection budget. Allocating this fixed number of households to individual geographic areas must be done in a way that satisfies the competing goals of the CE and CPI programs as much as possible. The CE program's goal is to allocate the sample households to the selected geographic areas in proportion to their share of the U.S. population, whereas the CPI program's goal is to allocate sample households to the selected urban areas in proportion to their share of the Nation's urban population. The CPI program further strives to include a minimum number of households in each selected urban area to ensure the statistical quality of its published price indexes for those areas.

This article describes a new automated method of allocating the CE's nationwide sample of households in a way that balances competing goals and constraints. The CE actually consists of two surveys, the Diary and Interview surveys, but this article focuses on the Interview survey.

Geographic areas in the CE sample

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 design currently used in the survey, based on the 1990 census, consists of 105 PSUs, classified into 4 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 urban areas
- 18 “D” PSUs, which are nonmetropolitan rural areas. The “D” PSUs are used in the CE program but not in the CPI program.

These 105 PSUs are grouped according to the geographic areas they represent. A populous PSU constitutes its own geographic area, which is called a “self-representing” geographic area. The 31 A PSUs are self-representing geographic areas, and they are in the sample with certainty. The 74 B, C, and D PSUs are “non-self-representing” PSUs. They were randomly selected to represent all of the less populous PSUs in the Nation. The 74 non-self-representing PSUs are grouped into 11 geographic areas called region-size classes, which are formed by cross-classifying the 4 regions of the country (Northeast, Midwest, South, and West) with the 3 size classes (B,C, and D) as shown in the shaded area of the table below. There are only 11 region-size classes for the areas that are not self-representing because no C PSUs were selected in the Northeast.

These 11 region-size classes are treated just like the 31 A PSUs and are also referred to as self-representing geographic areas. Hence, the CE can be thought of as having 42 self-representing geographic areas: 31 A PSUs plus 11 region-size classes for the smaller PSUs. Because the 4 D region-size classes are used by the CE only, there are only 38 self-representing geographic areas used by the CPI.

The sample allocation problem

In the CE’s current sample design, usable interviews are collected from 7,760 households[1] in each calendar quarter of the year: 4,260 households in the A PSUs, and 3,500 households in the B, C, and D PSUs. To guarantee that enough data are collected to satisfy CPI’s publication requirements, the sample of 7,760 households is allocated so that at least 120 usable interviews are obtained in each of the 38 geographic areas used by the CPI, with no minimum number of usable interviews required in the 4 D geographic areas.

Thus, the problem is to allocate the 7,760 households in the CE’s national sample to the 42 geographic areas in a way that satisfies the following constraints:

- The 31 A PSUs are allotted 4,260 households.
- The 11 B, C, and D region-size classes are allotted 3,500 households.
- Each of the 38 geographic areas used in the CPI is allotted 120 or more households.

Table 1. **PSU region-size classes**

Region	Size				
	A	B	C	D	Total
Northeast	6	8	–	4	18
Midwest	8	10	4	4	26
South	7	22	4	8	41
West	10	6	2	2	20
Total	31	46	10	18	105

BLS staff recently reevaluated the minimum sample size requirement of 120 usable interviews to determine whether it is still an appropriate number. One of the results of the reevaluation was the development of a new automated method of allocating the nationwide sample of households to geographic areas. The new method allowed repeated analyses to be conducted quickly and easily using different minimum sample size requirements. The method involved setting up the sample allocation problem as a mathematical optimization problem and using SAS statistical software to solve it.

Target versus required sample size

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:

$$E(x_i) \ge 120.$$

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,

$$P\{x_i \ge 120\} \ge 0.95$$

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

For example, under the first interpretation (target sample size), data collectors would have to visit 185 households in each quarter of the year to collect 120 usable interviews in the Boston metropolitan area, assuming that usable interviews are obtained at 65 percent of the residential addresses in the CE’s sample.[2]

$$E(x_i) = 185 \times 0.65 = 120$$

[1] In 2000 the average number of usable interviews collected per quarter in the CE Interview Survey was 7,760.

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

However, under the second interpretation (required sample size), data collectors would have to visit 202 households to be 95-percent certain of getting at least 120 usable interviews, again assuming a 65-percent survey participation rate.

$$P\{ x_i \geq 120\} =$$

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

Table 2 shows the difference in the sample size that would be needed for a target versus a required minimum number of usable interviews. The number of selected addresses needed to achieve a target minimum sample size is approximately 10 percent less than that needed for a required sample size.

The estimates in table 2 were produced using formulas from the binomial distribution for the mean and variance of the number of usable interviews,

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

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

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

One-sided confidence interval:

$$[\, \mu - 1.64\sigma\, ,\, +\infty\,)$$

Two-sided confidence interval:

$$[\, \mu - 1.96\sigma\, ,\, \mu + 1.96\sigma\,]$$

After some discussion, staff decided that target sample sizes would be satisfactory. Because the widths of the two-sided confidence intervals are relatively small, it is unlikely that any sample sizes achieved will be greatly below the target level.

Setting up the optimization problem

The CE's current sample design calls for allocating 7,760 households to the 42 geographic areas in a way that satisfies the three constraints mentioned previously.

These constraints can be written in mathematical terms as follows:

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

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

Again, the objective of the CE's sample design is to allocate the nationwide sample of households to geographic areas in a way that minimizes the standard error of the expenditures estimate at the national level. Allocating the sample in proportion to the population that each geographic area represents comes very close to achieving that goal. Although this allocation does not minimize the nationwide standard error, it is a very simple sample design that is known to achieve near minimization. Staff chose to implement this method because of its simplicity and its near optimal properties.

Table 2. **Sample size needed to obtain a target versus a required minimum number of usable interviews for the Consumer Expenditure Survey**

Number of sample households (*n*)	Expected number of usable interviews assuming a 65-percent survey participation rate (=0.65*n*)	95-percent confidence interval
Target sample size (two-sided 95-percent 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 (one-sided 95-percent confidence interval)		
72	47	[40, +∞)
105	68	[60, +∞)
137	89	[80, +∞)
170	110	[100, +∞)
202	131	[120, +∞)
234	152	[140, +∞)

Based on research and evaluation, staff modified the sample allocation problem described above. More of the CE's sample households were allocated to the urban portion of the Nation (of interest to the CPI), and fewer households were allocated to rural areas. This change results in a slight oversampling of the urban areas: The CPI-U population represents about 87 percent of the total U.S. population, but it is given 95 percent of the CE's sample. An analysis showed that limiting the rural sample to 400 households would have a minimal effect on the nationwide standard error of the CE's expenditure estimates. Thus, the revised optimization problem allocates exactly 400 households to the 4 rural geographic areas, leaving 7,360 households to be allocated to the 38 urban geographic areas.

For some of the geographic areas with small populations—for example, Anchorage and Honolulu—the requirement that at least 120 usable interviews be collected during each calendar quarter conflicts with the objective of allocating the sample in proportion to the population. For example, the Anchorage metropolitan area has approximately 0.09 percent of the U.S. population, and allocating the 7,760 usable interviews proportionally would give Anchorage only enough addresses to obtain 7 usable interviews—not 120.

Because an exact proportional allocation cannot be achieved within the given constraints, BLS staff decided to allocate the sample as proportionally as possible. This involved setting up a least-squares problem to square the difference between each geographic area's proportion of the population and its proportion of the sample and then minimize the sum of those 42 squared differences.

Thus, the optimization task is to solve the following constrained least-squares problem:

Given values of n, p_i ,*and* 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
$$n_1+n_2+\cdots+n_{38}=7,360$$
$$n_{39}+n_{40}+n_{41}+n_{42}=400$$
$$n_i\geq 120 \text{ for } i=1,2,...,38$$
$$n_i\geq 0 \text{ for } i=39,\ldots,42$$

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 ($p = p_1 + p_2 + \cdots + p_{42}$)

Solving the optimization problem

The optimization 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 for nonlinear programming (PROC NLP) can handle both kinds of constraints simultaneously. An example using this SAS® procedure to solve the problem above is given at the end of this paper.

Table 3. **The effect of changes in minimum target sample size on the standard error for the Consumer Expenditure Survey**

Minimum target sample for each primary sampling unit	Percent change in standard error (from SE for a minimum target sample size of 120)
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	+.06
140	+1.45
150	+3.28
160	+5.63
170	+10.07
180	+14.41

Estimating the standard error

The variance of the estimate of consumer expenditures resulting from the sample allocation process described above was estimated using the following formula:

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

where

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

$\bar{x}$ = sample mean of the Nation

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

σ^2 = expenditure variance of a randomly selected household

The variance of the estimate of consumer expenditures 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

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

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

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

This formula allows comparisons to be made with 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 of interest; 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 instead of 120), the σ in the numerator and the σ in the denominator cancel each other.

Standard error with different minimum sample size requirements

After allocating the CE's nationwide sample to individual geographic areas using PROC NLP, staff computed the percentage change in standard error for various minimum target sample sizes. 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

Chart 1. **Changes in the Consumer Expenditure Survey's standard error with minimum sample size**

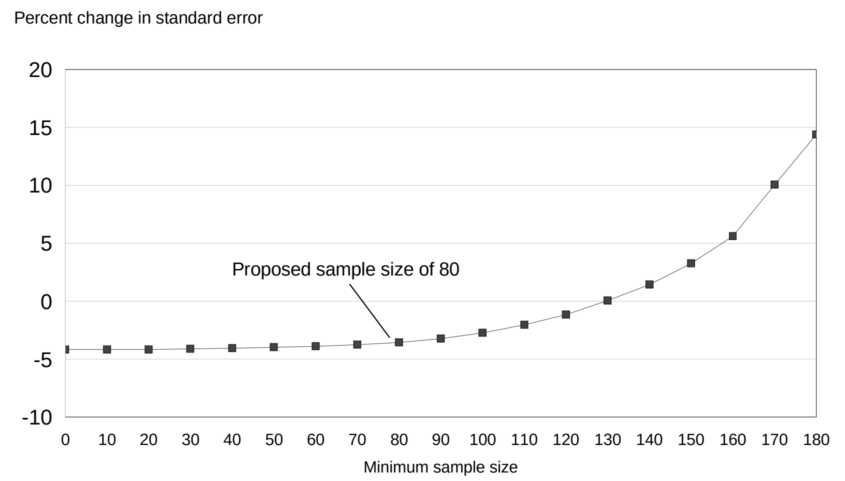

Table 4. **The effect of changing sample allocations on the standard error for the Consumer Expenditure Survey: Primary sampling units in the West**

Primary sampling unit	Population	Current sample size	Proposed sample size	Percent change in standard error
A419 Los Angeles	8,863,164	231	290	-10.81
A420 Greater Los Angeles	5,668,365	152	187	-9.88
A422 San Francisco	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

NOTE: Minimum target sample size is 80.

of the comparisons are shown above in table 3.

Standard error is minimized when the sample is allocated directly in proportion to population—that is, when 0 is the minimum number of usable interviews required in each geographic area (table 3). Reducing the target number of usable interviews from 120 to 0 would reduce the standard error by 4.16 percent. Standard error is maximized when the sample is divided equally among all geographic areas—180 usable interviews per geographic area. Increasing the target number of usable interviews from 120 to 180 would increase the standard error by 14.41 percent.

Reducing the minimum target number of usable interviews from 120 to 80 per geographic area would reduce the standard error by 3.54 percent. Nearly all the reduction in standard error is achieved by reducing the minimum target sample size to 80, and little further reduction is achieved by reducing the minimum target sample size below 80 (chart 1). Therefore, staff decided to reduce the minimum target sample size from 120 to 80 usable interviews per geographic area.

Other effects of the proposed allocation

A minimum target sample size of 80 usable interviews per geographic area reduces the national standard error by 3.54 percent and reduces the standard error in the urban portion of the Nation by 3.86 percent. After some discussion, staff decided that a minimum target sample size of 80 would be satisfactory for both surveys because the overall standard error would be reduced and publication criteria met for both the CE and CPI programs.

Table 4 shows current and proposed sample sizes for A PSUs in the West after applying the proposed sample allocation method. The PSUs with populations larger than 4 million will have their sample sizes increased, while the PSUs with populations less than 4 million will have their sample sizes decreased. This change will decrease the standard error in the larger A PSUs and increase the standard error in the smaller A PSUs, but the standard error for the Nation as a whole will be reduced.

BLS staff tested other methods to find one that satisfied the goals of both the CE and CPI programs. Some of the other methods tested had a positive effect on reducing the standard error for CE, but not for CPI, and vice versa. The chosen method reduced CE and CPI standard errors by about the same amount, 3.54 percent and 3.86 percent, respectively.

Conclusion

A new sample design for the CE will be introduced in 2005. Based on analysis of the current design, the new method of sample allocation could reduce the standard error of the estimate of consumer expenditures at the national level by from 3 percent to 4 percent.

The CE and CPI programs' competing goals and constraints complicated the process of allocating households to geographic areas in constructing the CE's national sample. CE program staff wanted to allocate the sample in a way that minimized the national variance, while CPI program staff wanted to minimize the variance of the urban portion of the Nation and also limit the variance of individual sampled areas. Setting up a mathematical optimization problem and then solving a constrained least-squares problem led to a solution that satisfied the requirements of both the CE and the CPI programs.

Writing the problem as a formal mathematical optimization problem had several advantages:

- It required the objectives and constraints to be stated clearly and explicitly.
- It helped document the allocation process.
- It allowed several different allocation methods to be tested quickly and easily.
- It led to an optimal solution to the problem.

This approach offers clear benefits for allocating the CE's nationwide sample of households to individual geographic areas while satisfying the CE and CPI programs' competing goals. ■

APPENDIX: Automating the Sample Allocation Process

Below is the optimization problem for the sample allocation, along with a SAS® program (PROC NLP) that solves it.

Given values of n, p_i , and 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 HOUSEHOLDS 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;
```