Selecting a Sample of Households for the Consumer Expenditure Survey

Susan L. King and
Sylvia A. Johnson-Herring

Susan L. King is a mathematical statistician in the Division of Price Statistical Methods, Branch of Consumer Expenditure Surveys, U.S. Bureau of Labor Statistics.

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

Introduction

The Consumer Expenditure Survey (CE) is a nationwide household survey designed by the U.S. Bureau of Labor Statistics (BLS) to find out how Americans spend their money. The CE consists of two separate surveys, the Diary and Quarterly Interview surveys. Each quarter of the year, approximately 3,200 households are visited in the Diary survey and approximately 15,000 households are visited in the Interview survey to collect information on the expenditures of American households. A question frequently asked by the survey respondents is "How was my household selected to be in this survey?" This article answers that question by looking at the CE's sample design and the selection process.

Survey description

The CE is an important economic survey. One of the primary uses of its data is to provide expenditure weights for the Consumer Price Index (CPI). The CPI affects millions of Americans by its use in cost-of-living wage adjustments for many workers, cost-of-living adjustments to Social Security payments, and inflation adjustments to Federal income-tax brackets. CE data also are used to compare expenditure patterns of various segments of the population, such as elderly versus nonelderly people. In addition, the data are being used by the Federal Government in a new experimental poverty index.

The purpose of the Diary survey is to obtain detailed expenditure data on small, frequently purchased items such as food and apparel. The purpose of the Interview survey is to obtain detailed expenditure data on large items such as property, automobiles, and major appliances; and on recurring expenses such as rent, utilities, and insurance premiums. Under contract with BLS, field representatives from the U.S. Census Bureau personally visit the households in the Diary and Interview surveys' samples to collect the data.

Each household in the Diary survey is asked to record all of the expenditures it makes during a 2-week period. Field representatives visit each household in the sample three times. On the first visit, the field representatives introduce themselves, explain the survey, and leave a diary in which the household members are asked to record all their expenditures for a 1-week period. On the second visit, the field representatives pick up the first week's diary, ask whether there are any questions, and leave another diary for the second week. On the third visit, the field representatives pick up the second week's

diary and thank the household for participating in the survey. After participating in the survey for 2 weeks, the household is dropped from the survey, and it is replaced by another household.

Each household in the Interview survey is interviewed every 3 months for 5 consecutive quarters. Trained field representatives ask household members about their expenditures over the previous 3 months. Responses are entered into a laptop computer. Each interview takes approximately 70 minutes to complete. Expenditure information obtained in the first interview is used only for "bounding" purposes, which address a common problem in which survey respondents tend to report expenditures to have been made more recently than they actually were made. Thus, expenditure information from the first interview is not used. Only expenditure information from the second through fifth interviews is used in the CE's published estimates. The households in the Interview survey are on a rotating schedule, with approximately one-fifth of the households in the sample being new to the survey each quarter.

Sample design

The selection of specific households to participate in the CE survey is carried out in multiple stages. The first stage of sampling is defining and selecting a random sample of geographic areas called "primary sampling units" (PSUs) from across the United States. In this stage, all of the counties in the United States are divided into small groups of counties (called PSUs), and a representative sample of them is selected to be in the survey. After the PSUs are defined and selected, the second stage of sampling involves determining the number of households to be visited in each PSU. The CE's budget allows for a certain number of households to be visited each year nationwide, and, in this stage, that number is allocated to the individual PSUs selected for the survey. The final stage of sampling is selecting specific households to be visited within the PSUs. Households are selected using a systematic selection procedure to ensure that each category of households is well represented in the survey. This is a brief summary of the CE's sample design. The rest of this article describes these steps in more detail.

Defining and selecting the PSUs

In the first stage of sampling, PSUs are defined and selected for the survey. PSUs are counties or groups of counties grouped together into geographic entities called "core-based statistical areas" (CBSAs) by the U.S. Office of Management and Budget. CBSAs were defined for use by Federal statistical agencies in collecting data and tabulating statistics.

There are two types of CBSAs, metropolitan and micropolitan. Metropolitan CBSAs consist of one or more counties with at least one urban area of 50,000 or more people, while micropolitan CBSAs consist of one or more counties centered around an urban area with 10,000–50,000 people. Both include the adjacent counties that have a high degree of social and economic integration with the area's core as measured by commuting ties. Areas outside CBSAs are called "non-CBSA" areas and are mostly rural.

After the PSUs are defined, they are categorized according to their population and region of the country. There are four regions of the country (Northeast, Midwest, South, and West), and four PSU "size classes":

- "A" PSUs, which are metropolitan CBSAs with a population over 2 million people
- "X" PSUs, which are metropolitan CBSAs with a population between 50,000 and 2 million people
- "Y" PSUs, which are micropolitan CBSAs
- "Z" PSUs, which are non-CBSA areas and are often referred to as "rural" PSUs

By definition, the "A" PSUs are "self-representing" and, therefore, have a 100 percent probability of selection in the survey. The "X," "Y," and "Z" PSUs are "non-self-representing." The non-self-representing PSUs are grouped together into groups of PSUs (called "strata") according to a 5-variable geographic model whose variables are latitude, longitude, latitude squared, longitude squared, and percent of the population in the PSU who live in an urban area. A typical stratum has approximately 10 PSUs, and all of the PSUs are in the same "region-size class." After the PSUs are grouped into strata, one PSU per stratum is randomly selected with probability proportional to its population. The PSU that is randomly selected represents the whole stratum.

For example, table 1 shows stratum X344, which is a group of eight "X"

Table 1. **The PSUs in stratum X344**

	PSU	Population
	Charlotte, NC-SC	1,114,808
	Charleston-North Charleston, SC	549,033
✓	**Greenville, SC**	**379,616**
	Fayetteville-Fort Bragg, NC	302,963
	Columbus, GA-AL	274,624
	Gastonia, NC	190,365
	Wheeling, WV-OH	153,172
	Warner Robbins, GA	134,433
	Total	**3,099,014**

PSUs in the South. According to the 2000 Census, their populations ranged from 134,433 to 1,114,808, for a total stratum population of 3,099,014 people. One PSU was randomly selected to represent the entire stratum. The PSU was Greenville, South Carolina. It had 12.25 percent of the stratum's population (0.1225=379,616/3,099,014); hence, it had a 12.25 percent chance of being selected, and a random number generator selected it.

PSU definitions for the current CE sample are based on information from the 2000 Census. Prior to 2005 (1996–2004), PSUs were defined based on information from the 1990 Census. The two sample designs are called the "2000 Census-based sample design" and the "1990 Census-based sample design," respectively. The original 2000 Census-based sample design consists of 102 PSUs, of which 86 urban PSUs are designated as "CPI areas." The CE and CPI share the sample design, with the exception of the "Z" PSUs. The CE survey covers the entire Nation ("A," "X," "Y," and "Z" PSUs), while the CPI survey covers only the urban portion of the Nation ("A," "X," "Y," but not "Z" PSUs.) See table 2 for the number of PSUs by region and size class in CE's original 2000 Census-based sample design.

Shortly after this sample design was implemented, newly imposed budget constraints forced the CE and CPI to eliminate 11 "X" PSUs from the sample, and to change the size class of 7 "A" PSUs to the "X" category. As a result, the sample of PSUs currently used by the CE has 91 PSUs, of which 75 urban PSUs also are used by the CPI. The CE began collecting data in the original 2000 Census-based sample design in 2005 and in the revised 2000 Census-based sample design in 2006. (See table 3 for a summary of the revised 2000 Census-based sample design.)

A map of the PSUs in the revised 2000 Census-based sample design is shown in figure 1.

Allocating the national sample of households to individual PSUs

Once the PSUs are selected, the number of households to be visited in each PSU must be determined. In the original 2000 Census-based sample design, CE's budget allowed for 7,700 household interviews per year in the Diary survey and 7,700 household interviews per quarter in the Interview survey (interviews 2–5 only) at the national level. In this stage of sampling, those 7,700 households are allocated (divided) among the 102 PSUs in the original 2000 Census-based sample design.

The first step in determining the number of households to visit in each PSU is to group the "X," Y," and "Z" PSUs by region and size class. Cross-classifying the four regions of the country (Northeast, Midwest, South, and West) with the three non-self-representing PSU size classes ("X," Y," and "Z") yields 12 region-size classes, which are treated like the 28 self-representing ("A") PSUs. This gives 40 self-representing geographic areas.

The objective of this stage of sampling is to allocate the 7,700 households to the 40 areas in a way that minimizes the standard error of CE's published expenditure estimates at the national level. This can be accomplished by allocating the households in a manner that is directly proportional to the population that each area represents; this allocation method is a standard statistical technique that comes very close to minimizing the standard error at the national level.

Without any modifications, proportional allocation would have given 7,034 households to the urban ("A," "X," and "Y") areas and 666 households to the rural ("Z") areas. However, research indicated that increasing the number of households in urban areas to 7,300 and decreasing the number of households in rural areas to 400 would have a significant impact on lowering CPI's standard error but would have only a minimal impact on CE's standard error. Since the CPI is the CE's primary customer, the allocation process was modified to allocate exactly 7,300 households to the 36 urban areas, and exactly 400 households to the four rural areas. Further, to guarantee that enough interviews are collected to satisfy CPI's publication requirements in each of the 36 urban areas, the sample of 7,300 households is allocated in a way that at least 80 interviews are obtained in each area. Operationally, the 7,700 households were allocated to the 40 areas by solving the following nonlinear programming problem:

Table 2. **Original 2000 Census-based sample design (102 PSUs)**

PSU size class	Region				
	Northeast	Midwest	South	West	**Total**
A	6	5	7	10	**28**
X	4	12	18	8	**42**
Y	2	4	6	4	**16**
Z	2	4	6	4	**16**
Total	**14**	**25**	**37**	**26**	**102**

Table 3. **Revised 2000 Census-based sample design (91 PSUs)**

PSU size class	Region				
	Northeast	Midwest	South	West	**Total**
A	5	4	6	6	**21**
X	4	10	16	8	**38**
Y	2	4	6	4	**16**
Z	2	4	6	4	**16**
Total	**13**	**22**	**34**	**22**	**91**

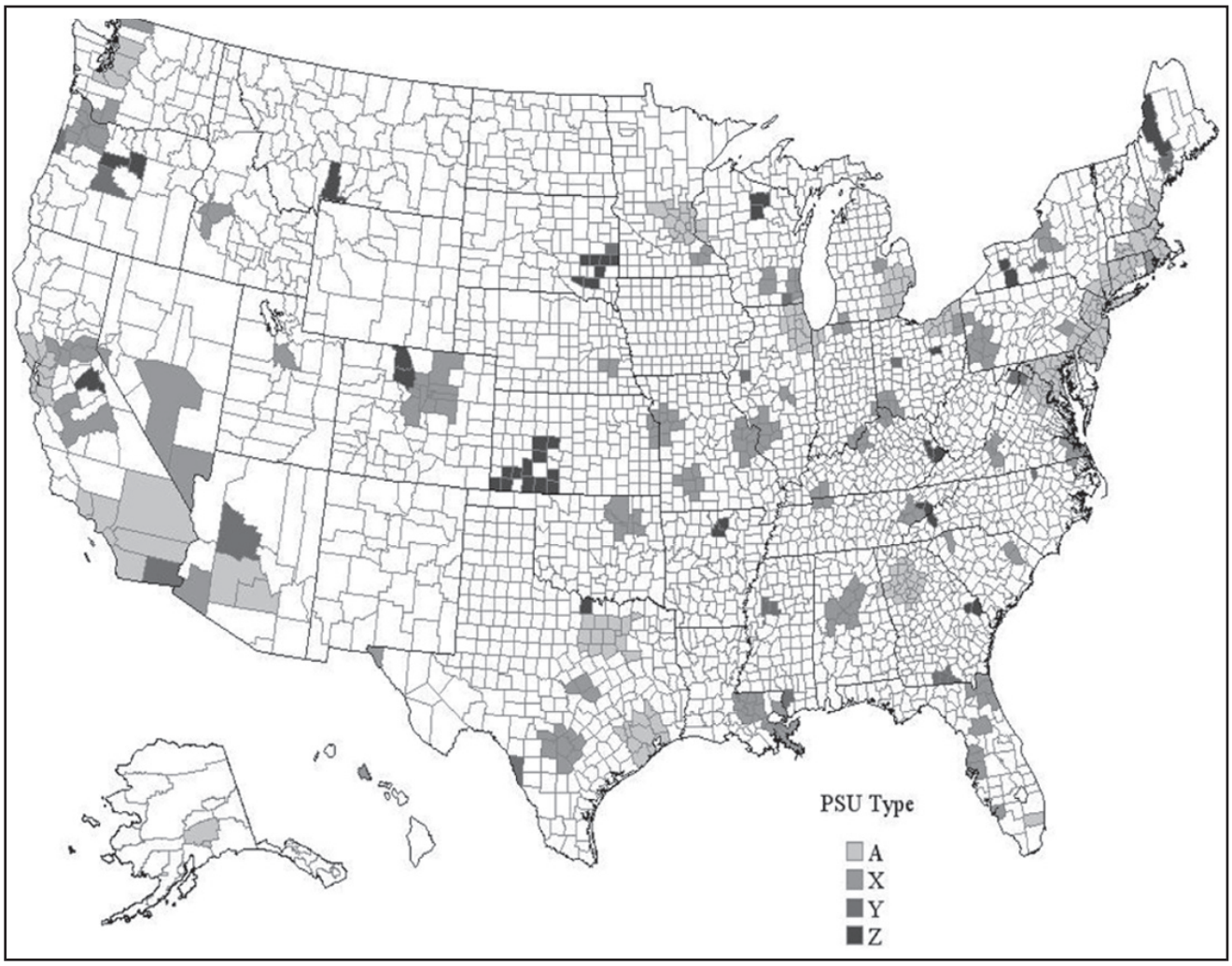

Figure 1. Spatial distribution of CE PSUs across the United States. The "A" PSUs correspond to the large population centers. The southern United States has a large number of "X" PSUs, and there are parts of the western United States without representation.

Given values of p_i and p, find values of x_i that…

minimize $\sum_{i=1}^{40}\left(\frac{x_i}{7{,}700}-\frac{p_i}{p}\right)^2$ (0)

subject to $\sum_{i=1}^{36} x_i = 7{,}300$ (1)

$\sum_{i=37}^{40} x_i = 400$ (2)

$x_i \geq 80 \qquad i = 1, 2, \ldots, 36$ (3)

$x_i \geq 0 \qquad i = 37, \ldots, 40$ (4)

where x_i = the number of households allocated to geographic 'area i'
p_i = the population represented by geographic 'area i'
$p = p_1 + p_2 + \ldots + p_{40}$

The output from this nonlinear program is an allocation of the 7,700 households to the individual geographic areas. The objective function (0) minimizes the sum of squared differences between each area's share of the national population and its share of the national sample of households. This allocates the sample of households as close to population proportionality as possible. Constraint (1) limits the sample of the 36 urban areas to 7,300 households. Constraint (2) limits the sample of the four rural areas to 400 households. Constraint (3) allocates at least 80 households to each urban area to ensure that the CPI's survey estimates are accurate enough to pub-

lish. Constraint (4) makes sure that the remaining areas are assigned nonnegative numbers of households.

After the 7,700 households are allocated to the 40 geographic areas, the households allocated to the 12 "X," "Y," and "Z" areas are suballocated to individual PSUs according to their proportion of the area's population.

Continuing the example from above, the nonlinear program allocated 1,342.32 out of 7,700 households to the "X" areas in the South. There are 18 "X" strata in the South, and stratum X344 has 6.20 percent of its population; hence, it was suballocated 6.20 percent of the sample. Thus, stratum X344 is given a target sample size of 83.22 interviewed households (83.22 = 1,342.32 x 0.0620).

Adjusting the PSU's target sample sizes for nonparticipation

Unfortunately, not all households selected for the survey participate in it. Some households cannot be contacted; some households are contacted but refuse to participate; and some households are ineligible for the survey. As a result of this "nonparticipation," the actual number of households designated for the survey must be larger than the target number of interviewed households. The designated number of households to be visited in each PSU is determined by adjusting the target sample size that was identified by the expected survey participation rate.

For example, the participation rate in stratum X344 is estimated to be 60 percent based on data from 1999–2001. Approximately 20 percent of the households are "out of scope" (the housing units are unoccupied, demolished, converted to nonresidential use, located on a military base, etc.), and 20 percent of the households are "in scope" but do not participate, leaving 60 percent of the households participating in the survey. Thus, the sample size inflation factor for stratum X344 is 1.66 (= 1/0.60), which means 166 households need to be selected for every 100 completed interviews that are wanted. Finally, the inflated target sample size is multiplied by 2 to account for the two surveys, Diary and Interview. This yields a "designated sample size" for each PSU. In stratum X344, the designated sample size is 276.29 households:

Designated Sample Size = (Target Sample Size) x (Nonparticipation Inflation Factor) x 2
= 83.22 x 1.66 x 2
= 276.29

This means that, each year, the U.S. Census Bureau selects 276.29 households in the Greenville, South Carolina, metropolitan area in order to collect data from 83.22 households per year in the Diary survey and 83.22 households per quarter in the Interview survey (interviews 2–5 only).

The revised sample allocation

As mentioned, shortly after the original 2000 Census-based sample design was implemented, newly imposed budget constraints caused the CE and CPI to eliminate 11 "X" PSUs from the sample and to change the size class of 7 "A" PSUs to the "X" category. When this change was implemented, a decision was made to keep the target sample sizes for the PSUs in the 2000 Census-based sample design and to drop the 642 households that had been allocated to the 11 eliminated PSUs. This effectively reduced the national target sample size from 7,700 to 7,058. Computations to reallocate the national sample were not carried out. Instead, the CE's original sample size was simply reduced by the sample sizes that were allocated to the 11 eliminated PSUs.

Selecting the households to visit

After determining the number of households to visit in each PSU, the final stage of sampling is selecting specific households to visit. The U.S. Census Bureau has a list of households across the Nation (called the "sampling frame"), and the specific households to visit are selected from that list.

The sampling frame is divided into four "segments": Unit, area, permit, and group quarters. The "unit" segment has about 80 percent of the households, and it consists of regular housing units with "city-style addresses" (street name, house number, apartment number, etc.). The "area" segment has about 10 percent of the households, and it consists of housing units that are physically located and listed by Census field personnel prior to sample selection. Most households in the "area" segment are in rural areas. The "permit" segment has about 9 percent of the households, and it consists of housing units that were constructed after April 1, 2000 (the date of the last census). Finally, the "group quarters" segment has about 1 percent of the households, and it consists of housing units in which the occupants share their living arrangements.

Within each PSU, a "systematic sample" of households is selected from each of the four segments. The households are sorted by variables that are correlated with their expenditures: Urban/rural; the market value of the home (for owners) or the rental value of the apartment (for renters); the number of people in the household; etc. This ensures that every kind of household is well represented in the survey. Although the specific variables used to sort the households differ slightly in each of the four segments, the procedures for selecting the sample are the same.

Once the list of households is sorted, a systematic sample of households is selected. The first household selected from the list is randomly selected using a random number generator to select one of the first k households on the list. Then, the remaining households are selected by taking every k^{th} household on the list after the first one. The number k is called the "sampling interval," and it is computed independently for each PSU by dividing the total number of households in the PSU by the number of households in the PSU that will be visited.

For example, in stratum X344 (Greenville, South Carolina), the sam-

pling frame has 176,654 households, and the CE draws a sample of 276.29 households per year in that area; hence, the sampling interval is $k = 639.38$:

k = PSU sampling interval

= (Number of Households in the PSU) / (Designated sample size)

= 176,654 / 276.29

= 639.38

This means that the first household selected for the sample is one of the first 639 households on the list. After the initial household is randomly selected, every 639th household on the list is selected for the sample as well. Thus, if the r^{th} household on the list is randomly selected ($1 \leq r \leq 639$), then the other households will be $r + 639$, $r + (2 \times 639)$, $r + (3 \times 639)$, etc. The selected households are assigned to the Diary and Interview surveys on an alternating basis.

Conclusion

This article describes the CE's selection of a representative sample of American households to participate in a survey about their expenditures. The first stage of sampling is defining geographic areas called "PSUs," which are small groups of counties. The PSUs are grouped into "strata," and one PSU is randomly selected from each stratum. Each randomly selected PSU represents itself plus the other nonselected PSUs. Then, the number of interviewed households targeted for the entire Nation is allocated to the individual PSUs, and those numbers are inflated to account for survey "nonparticipation." Finally, the specific households to be visited are selected from the complete list of households (called the "sampling frame") using a systematic selection procedure. The three-stage sampling process provides the CE with a well-balanced and representative sample of American households. ■