**How Was My Household Selected For The Consumer Expenditure Survey?**

**The Design and Selection of the Survey's Sample**

Susan L. King and Sylvia A. Johnson-Herring

## Introduction

The Consumer Expenditure Survey (CE) is a nationwide household survey conducted by the U.S. Bureau of Labor Statistics (BLS) to find out how Americans spend their money. The CE actually consists of two separate surveys, the Diary (CED) and Quarterly Interview (CEQ) surveys. Approximately 3,200 households are visited each quarter of the year in the CED, and approximately 15,000 households are visited each quarter of the year in the CEQ 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 method of selecting households for the survey.

## 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 are also used to compare expenditure patterns of various segments of the population, such as elderly versus non-elderly people, and it is currently being used by the federal government in a new experimental poverty index.

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, and major appliances; and on expenses that occur on a regular basis such as rent, utilities, and insurance premiums. Data for the CED and CEQ are collected by the Bureau of the Census under contract with the BLS, and are collected by personal visits to the households in the surveys' samples.

Each household in the CED is asked to keep a record of all 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 representative introduces herself, explains the survey, and leaves a diary in which the household members are asked to record all their expenditures for a one-week period. On the second visit the field representative picks up the first week's diary, asks whether there are any questions, and leaves another diary for the second week. On the third visit the field representative picks up the second week's diary and thanks the household for participating in the survey. After participating in the survey for two weeks the household is dropped from the survey, and it is replaced by another household.

Each household in the CEQ is interviewed every three months for five consecutive quarters. Interviews are conducted in person by trained field representatives who ask members of the household about their expenditures over the previous three months. The field representative enters the responses 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 addresses a common problem in which survey respondents tend to report expenditures to have been made more recently than they were actually made. As a result, expenditure information from the first interview is not used in the CE's published estimates. Only expenditure information from the second through fifth interviews is used in the published estimates. The households in the CEQ 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 CE survey has a multistage stratified sample design. The first stage of sampling is the definition and selection of a random sample of geographic areas called "primary sampling units" (PSUs) from across the United States. Once the PSUs are defined and selected, the second stage of sampling is determining the number of households to visit in each PSU. The final stage of sampling is the selection of specific households to visit within the PSUs.

**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 under 2 million people
- "Y" PSUs, which are micropolitan CBSAs
- "Z" PSUs, which are non-CBSA areas, and are often referred to as "rural" PSUs

The "A" PSUs are "self-representing" and are included in the survey with certainty. 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 independent variables are latitude, longitude, latitude squared, longitude squared, and the percent of people in the PSU who live in an urban area. A typical stratum has approximately ten 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" 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 lucky PSU was Greenville, South Carolina. It had 12.25% of the stratum's population (0.1225 = 379,616 / 3,099,014), hence it had a 12.25% chance of being selected, and a random number generator selected it.

**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** |

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.

**Table 2. Original 2000 Census-based Sample Design**
(102 PSUs)

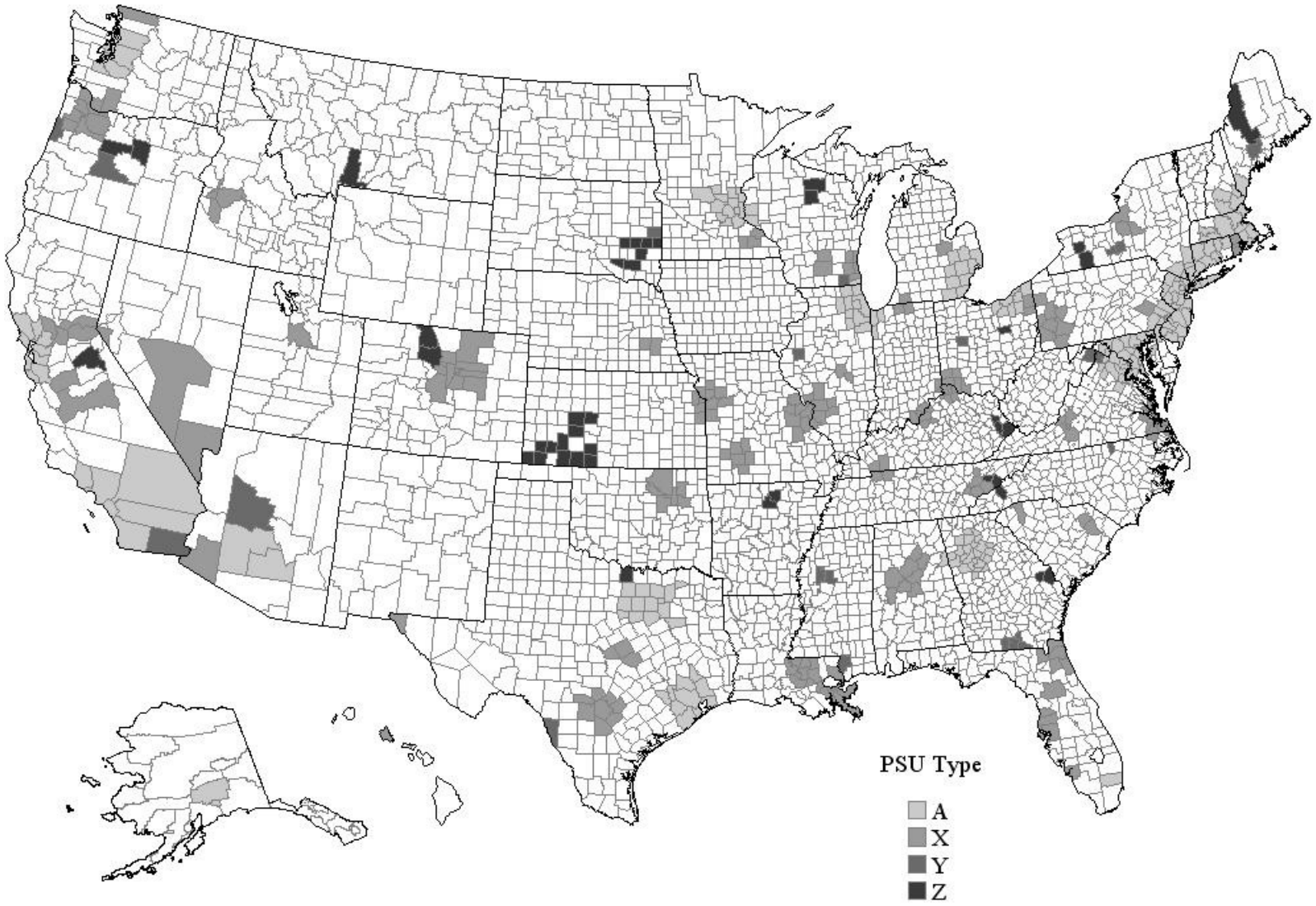
| PSU Size Class | Region | | | | |
|---|---|---|---|---|---|
| | Northeast | Midwest | South | West | **Total** |
| A | 6 | 5 | 7 | 10 | **28** |
| X | 4 | 12 | 18 | 8 | **42** |
| Y | 2 | 4 | 6 | 4 | **16** |
| Z | 2 | 4 | 6 | 4 | **16** |
| **Total** | **14** | **25** | **37** | **26** | **102** |

Shortly after this sample design was implemented, newly imposed budget constraints forced the CE and CPI to eliminate eleven "X" PSUs from the sample, and to change the size class of seven "A" PSUs to the "X" category. As a result of these changes, the sample of PSUs currently used by the CE has 91 PSUs, of which 75 urban PSUs are also used by CPI. 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.

**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** |

A spatial distribution of the PSUs in the revised 2000 Census-based sample design is shown in Figure 1.


PSU Type
A
X
Y
Z

**Allocating the National Sample of Households to Individual PSUs**

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

The first step in determining the number of HHs 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 HHs 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 HHs directly proportional to the population each area represents. This is a simple and 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 HHs to the urban ("A," "X," and "Y") areas and 666 HHs to the rural ("Z") areas. However, research indicated that increasing the number of HHs in urban areas to 7,300 and decreasing the number of HHs 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 biggest customer, the allocation process was modified to allocate exactly 7,300 HHs to the 36 urban areas, and exactly 400 HHs 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 HHs is allocated in a way that at least 80 interviews are obtained in each area. Operationally, the 7,700 HHs were allocated to the 40 areas by solving the following nonlinear programming problem:

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

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

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

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

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

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

where $x_i$ = the number of HHs 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 HHs 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 HHs. This allocates the sample of HHs as close to population proportionality as possible. The first constraint (1) limits the sample of the 36 urban areas to 7,300 HHs. The second constraint (2) limits the sample of the four rural areas to 400 HHs. The third constraint (3) allocates at least 80 HHs to each urban area to make sure the CPI's survey estimates are accurate enough to publish. The fourth constraint (4) makes sure the remaining areas are assigned nonnegative numbers of HHs.

After the 7,700 HHs are allocated to the 40 geographic areas, the HHs allocated to the 12 "X," "Y," and "Z" areas are sub-allocated 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 HHs to the "X" areas in the South. There are 18 "X" strata in the South, and stratum X344 has 6.20% of its population, hence it was sub-allocated 6.20% of the sample. Thus stratum X344 is given a target sample size of 83.22 interviewed HHs (83.22 = 1,342.32 × 0.0620).

**Adjusting the PSU s' Target Sample Sizes for Non-Participation**

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

For example, the participation rate in stratum X344 is estimated to be 60% based on data from 1999-2001. Approximately 20% of the HHs are "out-of-scope" (the housing units are unoccupied, demolished, converted to non-residential use, located on military a base, etc.), and 20% of the HHs are "in-scope" but do not participate in the survey, leaving 60% of the HHs participating in the survey. Thus the sample size inflation factor for stratum X344 is 1.66 (=1/0.60), which means 166 HHs 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, CED and CEQ. This yields a "designated sample size" for each PSU. In stratum X344 the designated sample size is 276.29 HHs:

$$\begin{aligned}\text{Designated Sample Size} &= (\text{Target Sample Size}) \times (\text{Non-participation Inflation Factor}) \times 2 \\ &= 83.22 \times 1.66 \times 2 \\ &= 276.29\end{aligned}$$

This means each year the Bureau of the Census needs to select 276.29 HHs in the Greenville, South Carolina metropolitan area in order to collect data from 83.22 interviewed HHs per year in the CED and 83.22 interviewed HHs per quarter in the CEQ (interviews 2-5 only).

**The Revised Sample Allocation**

As mentioned earlier, shortly after the original 2000 Census-based sample design was implemented, newly imposed budget constraints caused the CE and CPI to eliminate eleven "X" PSUs from the sample, and to change the size class of seven "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 simply drop the 642 HHs that had been allocated to the eleven eliminated PSUs. This effectively reduced the national target sample size from 7,700 to 7,058. Computations to re-allocate 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 eleven eliminated PSUs.

**Selecting the HHs to Visit**

After the number of HHs to visit in each PSU is determined, the final stage of sampling is the selection of specific HHs to visit. The Bureau of the Census has a list of HHs across the nation (called the "sampling frame"), and the specific HHs 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 HHs, 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 HHs, and it consists of housing units that are physically located and listed by Census field personnel prior to sample selection. Most HHs in the "area" segment are in rural areas. The "permit" segment has about 9 percent of the HHs, 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 HHs, and it consists of housing units in which the occupants share their living arrangements.

Within each PSU, a "systematic sample" of HHs is selected from each of the four segments. The HHs 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 HH; etc. This is done to ensure that every kind of HH is well represented in the survey. Although the specific variables used to sort the HHs differ slightly in each of the four segments, the procedures for selecting the sample are the same.

Once the list of HHs is sorted, a systematic sample of HHs is selected. The first HH selected from the list is randomly selected using a random number generator to select one of the first k HHs on the list. Then the remaining HHs are selected by taking every $k^{th}$ HH 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 HHs in the PSU by the number of HHs in the PSU that will be visited.

For example, in stratum X344 (Greenville, South Carolina) the sampling frame has 176,654 HHs, and the CE draws a sample of 276.29 HHs per year in that area, hence the sampling interval is k=639.38.

$k$ = PSU sampling interval
= (Number of HHs in the PSU) / (Designated sample size)
= 176,654 / 276.29
= 639.38

This means the first HH selected for the sample is one of the first 639 HHs on the list. Then after the initial HH is randomly selected, every $639^{th}$ HH on the list after it is selected for the sample as well. Thus if the $r^{th}$ HH on the list is randomly selected ($1 \leq r \leq 639$), then the other HHs will be r + 639, r + (2×639), r + (3×639), etc. The selected HHs are assigned to the CED and CEQ surveys on an alternating basis.

**Conclusion**

In this article we have shown how the CE selects a representative sample of American HHs to find out how Americans spend their money. The first stage of sampling is the definition of geographic areas called "PSUs," which are groups of counties. The PSUs are grouped into "strata," and one PSU is randomly selected from each stratum. The randomly selected PSU represents itself plus the other non-selected PSUs. Then the number of interviewed HHs targeted for the entire nation is allocated to the individual PSUs, and those numbers are inflated to account for survey "non-participation." Finally, the specific HHs to visit are selected from the complete list of HHs (the "sampling frame") using a systematic selection procedure.

So if your HH is in a selected PSU and your HH matches one of the HHs selected in the final stage of sampling, you will be contacted to participate in the CE survey.