Selecting a Sample of Households for the Consumer Expenditure Survey

The Design and Selection of the Survey's Sample for 2015-2024

By Susan L. King

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 consists of two separate surveys, the Diary and Quarterly Interview surveys. Each quarter of the year, approximately 3,000 households are visited in the Diary survey and approximately 12,000 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 reviewing the CE's sample design and the selection process for the next ten years (2015-2024).

Survey Description

The CE is an important economic survey. One of the primary uses of the 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. In addition, the data are used to calculate poverty thresholds for the Supplemental Poverty Measure, which is a measure of poverty that augments the official poverty measure.

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 replaced by another household.

Each household in the Interview survey is interviewed every 3 months for 4 consecutive quarters. Field representatives ask household members about their expenditures over the previous 3 months, and their responses are entered into a laptop computer. Each interview takes approximately 60 minutes to complete. The households in the Interview survey are on a rotating schedule, with approximately one-fourth 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 is selected to be in the survey. Then after the PSUs are defined and selected, the second stage of sampling is determining the number of households to be visited in each PSU. The CE's budget allows a certain number of households to be visited each year nationwide, and, in this stage that number is allocated to the individual PSUs that were 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 every 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 centered around an 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.

Since OMB does not group rural counties into small clusters of adjacent counties, CE defines its own rural PSUs. CE requires a rural PSU to be within a state border, consist of adjacent rural counties, have a land area less than 3,000 square miles, and have a population of at least 7,500 people. The last two constraints are guidelines used by the Census Bureau for establishing the maximum workload for a single field representative.

After the PSUs are defined, they are categorized according to their population and Census region and division of the country. The Census Bureau divides the United States into four geographic regions (Northeast, Midwest, South, and West); and each region has two divisions except the South which has three divisions, which makes a total of nine divisions. There are three PSU size-classes since smaller metropolitan and micropolitan PSUs are stratified together:

- "S" PSUs, which are metropolitan CBSAs with a population over 2.5 million people.
- "N" PSUs, which are metropolitan CBSAs with a population under 2.5 million people and micropolitan CBSAs.
- "R" PSUs, which are non-CBSA areas, and are often referred to as "rural" PSUs.

By definition, the "S" PSUs are "self-representing" and, therefore, have a 100 percent probability of selection in the survey. The "N," and "R" PSUs are "non-self-representing," and a sample of them is randomly selected for the survey. The non-self-representing PSUs are grouped together into groups of PSUs (called "strata") according to a 4-variable geographic model whose variables are: median household income, median household property value, latitude, and longitude. An average "N" stratum has approximately 26 PSUs, and all of the PSUs are in the same "division-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 stratum.

For example, table 1 shows a fictitious stratum, N35Q, which is a group of seven "N" PSUs in the South, Region 3, and the South Atlantic Division, Division 5. According to the 2010 Census, their populations ranged from 27,731 to 2,217,012, for a total stratum population of 3,090,499 people. Both Charlotte and Charleston are metropolitan PSUs, whereas the remaining PSUs are micropolitan CBSAs. One PSU was randomly selected to represent the entire stratum. Charleston, South Carolina has 21.5 percent of the stratum's population (0.215 = 664,607 / 3,090,499), hence it had a 21.5 percent chance of being selected, and a random number generator selected it for the sample.

Table 1. The PSUs in Stratum N35Q

	PSU	Population
	Charlotte-Concord-Gastonia, NC-SC	2,217,012
✓	**Charleston-North Charleston, SC**	**664,607**
	Gaffney, SC	55,342
	Henderson, NC	45,422
	Douglas, GA	42,556
	Americus, GA	37,829
	Wauchula, FL	27,731
	Total	**3,090,499**

For stratification, Alaska and Hawaii are separated from the continental United States because they have homogeneous markets with unique pricing behaviors and weak correlation with price changes of the other non-self-representing PSUs in the western United States. For this reason, in the previous design, both Anchorage, AK and Honolulu, HI were self-representing PSUs even though their populations were below the cut-off. In the current design, the four CBSAs in Alaska were grouped into a state stratum and Anchorage was selected to represent the state stratum. Likewise, the four CBSAs in Hawaii were grouped into a state stratum and Honolulu was selected to represent the stratum.

PSU definitions for the current CE sample (2015-2024) are based on information from the 2010 Census, while PSU definitions from the previous sample (2005-2014) were based on information from the 2000 Census. The two sample designs are called the "2010 Census-based sample design" and the "2000 Census-based sample design," respectively. The 2010 Census-based sample design consists of 91 PSUs, of which 75 urban PSUs are designated as "CPI areas." The CE and CPI share the sample design with the exception of the "R" PSUs. The CE survey covers the entire nation ("S," "N," and "R" PSUs), while the CPI survey covers only the urban portion of the nation ("S" and "N," but not "R" PSUs.) See table 2 for the number of PSUs by region, division, and size-class in CE's 2010 Census-based sample design.

Table 2. 2010 Census-based Sample Design (91 PSUs)

		PSU Size-Class			
Region	Division	S	N	R	Total
1. Northeast	1. New England	1	2	1	4
	2. Middle Atlantic	2	4	1	7
2. Midwest	3. East North Central	2	8	2	12
	4. West North Central	2	4	2	8
3. South	5. South Atlantic	5	12	2	19
	6. East South Central	0	6	2	8
	7. West South Central	2	8	2	12
4. West	8. Mountain	2	4	3	9
	9. Pacific	7	4	1	12
Total		23	52	16	91

A map of the PSUs in the 2010 Census-based sample design is shown in figure 1.

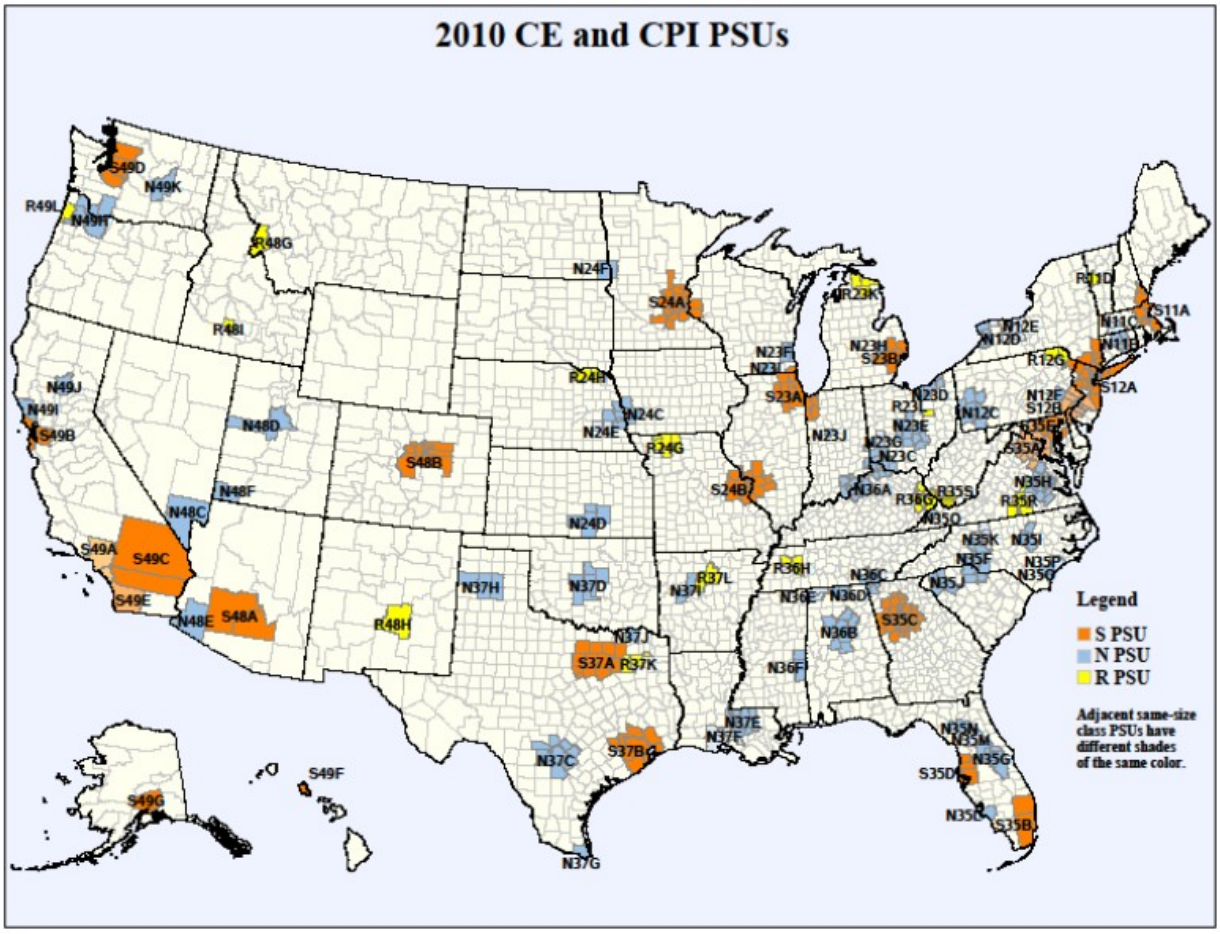

Figure 1. Spatial distribution of CE PSUs across the United States. The "S" (orange) PSUs are the large population centers. The southern United States has a large number of "N" (blue) PSUs. Every Census Division has at least one "R" (yellow) PSU.

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 2010 Census-based sample design, CE's budget allowed for 12,000 addresses to be visited per year in the Diary survey and 12,000 addresses to be visited per quarter in the Interview survey at the national level. In this stage of sampling, those 12,000 households are allocated (divided) among the 91 PSUs in the 2010 Census-based sample design.

The objective of the two-step allocation process is to allocate the 12,000 addresses to the PSUs to minimize CE's nationwide variance. In the first step, the 91 sample PSUs are placed into 41 "index areas" defined by CPI and the addresses are allocated directly proportional to the population represented by each of the CPI index areas. The 41 index areas consist of the 23 self-representing PSUs plus the 18 non-self-representing division size-classes (9 Census divisions x 2 size-classes). In the second step, the addresses are sub-allocated to individual PSUs in the index areas.

The nonlinear optimization program below allocates the 12,000 addresses to the 41 index areas. The objective function minimizes the sum of squared differences between each index area's share of the national population and its share of the addresses. Both the total U.S. population, *P*, and the population of each index area, p_i, are known. The expected number of interviewed households is $n_i r_i$, where n_i is the number of addresses, the decision variable to be determined in the optimization model, and r_i is the participation rate for index area *i*. The participation rate is the response rate times the eligibility rate. The eligibility rate is the percent of addresses on the sampling frame with occupied housing units and is calculated using the most recent five years of data from the American Community Survey (ACS), while the response rate is the percent of occupied housing units in CE's sample that give completed interviews and is calculated using the most recent five years of data from the CE survey. The total number of interviewed households is *NR*. The first constraint is linear and restricts the number of addresses to 12,000. The lower bound constraints require at least 80 interviewed households in each of the 32 urban index areas (*i = 1 to 32*) and 40 interviewed households in each of the 9 rural index areas (*i = 33 to 41*).

$$Minimize \sum_{i=1}^{41} \left(\frac{n_i r_i}{NR} - \frac{p_i}{P} \right)^2$$

$$Subject¿: \sum_{i-1}^{41} n_i = 12{,}000$$

$$n_i r_i \geq 80 \; for \; i = 1¿32$$

$$n_i r_i \geq 40 \; for \; i = 33¿41$$

where:

n_i = number of addresses to be allocated to the i[th] index area;

p_i = population of the i[th] index area;

r_i = participation rate of the i[th] index area, ($0 \leq r_i \leq 1$);

$P = \sum_{i=1}^{44} p_i$, the population of the United States;

$n_i r_i$ = expected number of interviewed households in the i[th] index area;

$NR = \sum_{i=1}^{41} n_i r_i$, the expected total number of interviewed households.

Since the response rates are different for the Interview and Diary Surveys, an optimization model is run for each survey. In the 2010 Census-based sample design, the number of addresses is calculated annually using the most current response and eligibility rates.

In the second step, the addresses are sub-allocated to each PSU based on its population percentage in the index area.

Selecting the Households to Visit

After the number of households to visit in each PSU is determined, the final stage of sampling is selecting specific households to visit. The Census Bureau has a list of households across the nation (called a "sampling frame"), and the specific households to visit are selected from that list. CE samples from the U.S. civilian non-institutional population, which includes people living in houses, condominiums, apartments, and group quarters such as college dormitories. However, military personnel living on base, nursing home residents, and prison inmates are excluded. The addresses are from two sampling frames maintained by the Census Bureau: the Unit and Group Quarters (GQ) frame. Both frames are derived from the Master Address File (MAF), which is basically a list of all residential addresses identified in the 2010 census and is updated twice per year with information from the U.S. Postal Service. It contains an accurate, up-to-date inventory of all known living quarters in the United States. The Unit frame is the larger frame and it contains both existing housing units and new housing units. It has approximately 99% of the MAF's civilian non-institutional addresses and is updated twice per year. The GQ frame is also created from the MAF but it is much smaller; it has the remaining 1% of the addresses and it is updated less frequently, every three years.

For each county within each PSU, a "systematic sample" of households is selected from each of the two frames. The Unit frame uses a stratification variable (the sorting variable) created from the number of occupants in each household, their housing tenure (owner/renter), and the market value of their home (for owners) or the rental value of their apartment or home (for renters). These variables are used because they are correlated with expenditures: households with more people tend to be wealthier than those with fewer people; homeowners tend to be wealthier than renters; and people living in high-price housing units tend to be wealthier than those living in low-price housing units. The GQ frame uses a geographic and block level sort on "percent of college housing." These stratification variables ensure that every economic segment of the population is equally represented in the sample.

Once the list of housing units within a county are sorted using the within-PSU stratification variable, the first housing unit is randomly selected using a random number generator. Then the remaining housing units are selected by taking every k^{th} housing unit on the ordered list. The number k is the sampling interval for the county and it is computed independently for each PSU by dividing the total number of housing units from the MAF by the desired sample size.

The Interview and Diary households are selected jointly, in one sample selection process for each frame, with the even numbered addresses assigned to the Interview survey and the odd numbered addresses assigned to the Diary survey. The number of selected addresses, sample size, is the larger of the two sample sizes and through a sample reduction process, addresses are randomly removed for the survey requiring the smaller number of addresses.

Conclusion

This article describes 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 non-selected PSUs. Next, the number of interviewed households is determined by the budget and allocated to the individual PSUs to minimize CE’s nationwide variance. Finally, the specific households to be visited are selected from the complete list of households (called the “sampling frame”) using a systematic selection procedure. This 3-stage sampling process provides the CE with a well-balanced and representative sample of American households.

References

King, S. and Johnson-Herring, S. (2008). Selecting a Sample of Households for the Consumer Expenditure Survey. Consumer Expenditure Survey Anthology, 14-19.