CHAPTER 4

SAMPLE DESIGN AND ESTIMATION

This chapter discusses selecting the public use microdata samples (PUMS) and forming estimates.

SAMPLE DESIGN

The 2010 PUMS was designed to include 10 percent of the housing units and 10 percent of the group quarters (GQ) persons from the entire 2010 Census population in the United States and Puerto Rico. The PUMS sample of persons in households was selected by keeping all persons in selected PUMS housing units. The 2010 PUMS sample design is different from the 2000 PUMS because the 2000 PUMS was selected from the long-form questionnaires; hence, the 2000 PUMS was a sample of a sample. For 2010, only short-form data was collected from every person and housing unit.

SELECTING THE PUBLIC USE MICRODATA SAMPLE

A 1-in-10 systematic selection procedure with equal probability was used to select the PUMS. The sampling universe was defined as all occupied housing units (including all occupants), vacant housing units, and GQ persons in the census. The sampling units were sorted during the selection process. The sorting was intended to improve the reliability of estimates derived from the 10-percent sample by implicitly defining strata within which there is a high degree of homogeneity among the census households with respect to characteristics of major interest.

The sample selection was done separately for each of the three subsampling universes: occupied housing units including all people in them, vacant housing units, and GQ persons. The sorting within these universes was done within each state in the United States, as well as the District of Columbia and Puerto Rico. Ten 10-percent samples were created from the full census population for a given state. The 10-percent PUMS for that state was designated at random from those 10 samples.

In the case of occupied housing units, the primary sampling units were housing units, and all person records were extracted after the housing units were chosen. The occupied housing unit universe was sorted in the following order:

- Race of householder
- Hispanic origin of householder
- Family type (with own children under 18, without own children under 18, nonfamily)
- Tenure
- Age group for the maximum age in the household (0 – 59, 60 – 74, 75 – 89, 90+)
- Unique housing unit identification code

The vacant housing unit universe was sorted in the following order:

- Vacancy status (for rent, for sale, other)
- Unique housing unit identification code

Finally, the GQ person universe was sorted in the following order:

- Race
- Hispanic origin
- GQ type (institutional or military, noninstitutional and nonmilitary)
- Age group (0 – 59, 60 – 74, 75 – 89, 90+)
- Unique GQ person identification code

SELECTING SUBSAMPLES OF THE PUMS FILES

Nationwide, the PUMS files have records for over 30 million people and 13 million housing units. Since processing a smaller sample is less resource intensive, some users may prefer to use a smaller sample. Within each state sample, 100 representative subsamples were designated during the PUMS sample selection. Two-digit subsample numbers from 00 to 99 were assigned consecutively to each sample case in the PUMS. The subsample numbers allow for 1) the designation of various size subsamples, and 2) the calculation of standard errors directly from the PUMS sample.

Reliability improves with increases in sample size, so the choice of sample size must represent a balance between the level of precision desired and the resources available for working with microdata files. To gauge the impact on the reliability when deciding sample size, use the following formula to approximate the increase in the sampling error for various subsampling rates of the full PUMS microdata.

$$Increase = \frac{\sqrt{\left(\frac{1}{f_1 \times f_2} - 1\right)}}{\sqrt{\left(\frac{1}{f_1} - 1\right)}} \qquad (1)$$

where
f_1 is the PUMS sampling rate, 0.10, and
f_2 is the rate at which the PUMS microdata records are subsampled.

For example, if selecting half of the PUMS sample, that is $f_2 = 0.5$, equivalently a 5-percent sample of the census, the increase in the sampling error would be a factor of 1.45 or 45 percent.

Samples of the total census population of any size between 10 percent (the PUMS sample) and 0.1 percent (1-percent sample of the PUMS records) may be selected by using appropriate two-digit subsample numbers assigned to the microdata sample. As an example, if the user wants to extract 10 of the 100 subsamples from the PUMS files, the choice of records having the same "units" digit in the subsample number (e.g., the 2 "units" digit includes subsample numbers 02, 12, 22, …, 92) will provide a 10-percent sample of the PUMS records or a 1-percent sample of the total census population. Care must be exercised when selecting such samples. If only the "units" digit is required, the "units" digit should be randomly selected. If two "units" digits are required, the first should be randomly selected and the second should be either 5 more or 5 less than the first. Failure to use this procedure, e.g., selection of records with the same "tens" digit

instead of records with the same "units" digit, would provide a 1-in-100 subsample of the total census population, but one that would be somewhat more clustered and, as a result, subject to larger sampling error.

PRODUCING ESTIMATES OR TABULATIONS

To produce estimates or tabulations of census characteristics from the PUMS files, add the weights of all persons or housing units that possess the characteristic of interest.

Equivalently, one can take advantage of the 2010 PUMS being a self-weighting sample. All persons or housing units in the PUMS have a weight of 10. Therefore, to produce estimates of characteristics from the PUMS files, multiply the number of PUMS persons or housing units that possess the characteristic of interest by 10. For instance, if the characteristic of interest is "total number of Hispanic males aged 5-17," determine the sex, age, and Hispanic origin of all persons, and multiply the number of Hispanic males aged 5-17 by 10.

To get estimates of proportions, divide the estimate of persons or housing units with a given characteristic by the base sample estimate. For example, the proportion of "owner-occupied housing units" is obtained by dividing the PUMS estimate of owner-occupied housing units by the PUMS estimate of total housing units.

To get estimates of characteristics such as the "total number of related children in households," sum the value of the characteristic across all household records and multiply by 10. If the desired estimate is the "number of households with at least one related child in the household," count all households with a value not equal to zero for the characteristic and multiply by 10.

The PUMS estimates are subject to sampling error, which is the source of any difference between a 2010 PUMS estimate and the 2010 census count of the same characteristic. The impact of sampling error ranges from being negligible for larger characteristics to being relatively larger for small characteristics. While sorting is a means for reducing sampling error for the sort characteristics, the impact is more evident for the primary sort variables relative to the secondary variables, particularly for small geographic areas and small characteristics. Chapter 5 discusses sources of error in the PUMS sample.

ADJUSTING WEIGHTS FOR SUBSAMPLING

To produce estimates of characteristics from a subsample of the PUMS files, the weights of all persons or housing units that possess the characteristic of interest must be adjusted according to the subsampling rate used. All persons or housing units in the PUMS have an original weight of 10. To determine the new weight for persons or housing units in a subsample, multiply the reciprocal of the probability of selection by 10. In general, let f_1 be the sampling rate for the PUMS (0.10) and f_2 be the subsampling rate. Then

$$new\ weight = \frac{1}{f_1} \times \frac{1}{f_2} \qquad (2)$$

For example, using equation (2), if the user wants a 20-percent sample of the PUMS records, the new weight would be

$$\frac{1}{0.10} \times \frac{1}{0.20} = 10 \times \frac{1}{0.20} = 50.$$