

# Chapter 5.

## Sample Design and Estimation

---

### **PRODUCING ESTIMATES OR TABULATIONS**

To produce estimates or tabulations of 100 percent characteristics from the PUMS files, simply add the weights of all persons or housing units that possess the characteristic of interest.

To create person estimates, use the person weight. To create estimates of households or families, use the person weight of the householder. Use the housing unit weight for housing unit estimates.

For instance, if the characteristic of interest is total number of Hispanic males, aged 5-17, simply determine the sex, age, and Hispanic origin of all persons and cumulate the weights of those who match the characteristic of interest. The PUMS weight is a function of the full census sample weight and the PUMS sample design. The Census 2000 PUMS design is not a self-weighting design.

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

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

### **LONG FORM SAMPLE DESIGN**

The Public Use Microdata Samples are chosen from the universe of Census 2000 Long Form records. Every person and housing unit in the United States was asked basic demographic and housing questions (for example, race, age, and relationship to householder). A sample of these people and housing units was asked more detailed questions about items, such as income, occupation, and housing costs. The sampling unit for Census 2000 was the housing unit, including all occupants. There were four different housing unit sampling rates: 1-in-8, 1-in-6, 1-in-4, and 1-in-2 (designed for an overall average of about 1-in-6). The Census Bureau assigned these varying rates based on precensus occupied housing unit estimates of various geographic and statistical entities, such as incorporated places and interim census tracts. For people living in group quarters or enumerated at long form eligible service sites (shelters and soup kitchens), the sampling unit was the person and the sampling rate was 1-in-6.

The sample designation method for housing units depended on the data collection procedures. The majority of the population was enumerated by the mailback procedure. In these areas, the Census Bureau used the Decennial Master Address File (DMAF) to select a probability sample. The questionnaires were either mailed or hand-delivered to selected addresses with instructions to complete and mail back the form.

---

The housing unit sampling rate varied by census block. Long Form Sampling Entities (LFSEs) were used to determine sampling rates in Census 2000 similarly to the way governmental units were used in the 1990 census sample design. LFSEs were defined to be:

- Counties and county equivalents (such as parishes in Louisiana).
- Cities.
- Incorporated places (including consolidated cities).
- Census designated places in Hawaii only.
- Minor civil divisions in certain states only (Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and Wisconsin).
- School districts (based on the 1995-1996 school year).
- American Indian reservations.
- Tribal Jurisdiction Statistical Areas (now known as Oklahoma Tribal Statistical Areas).
- Alaska Native village statistical areas.

Size estimates for LFSEs were based on housing unit counts from the DMAF and occupancy rates from the 1990 census. If the smallest LFSE that included all or any part of a block had an estimated housing unit count of less than 800, the housing units in the block were sampled at a 1-in-2 rate. If the smallest LFSE that included all or any part of a block had an estimated housing unit count of 800 or more but less than 1,200, housing units in the block were sampled at a 1-in-4 rate. If a block was not in either of the two previous sampling rate categories, and was part of an interim census tract with 2,000 or more estimated housing units, the housing units in the block were sampled at a 1-in-8 rate. Housing units in all remaining blocks (those not assigned to 1-in-2, 1-in-4, or 1-in-8 rates) were sampled at a 1-in-6 rate.

In List/Enumerate areas (accounting for less than 0.5 percent of the housing units), each enumerator was given a blank address register with designated sample lines. Only two sampling rates, 1-in-2 and 1-in-6 were used in these areas. Beginning about Census Day (April 1, 2000), the enumerator systematically canvassed an Assignment Area (AA) and listed all housing units in the address register in the order they were encountered. Completed questionnaires, including sample information for any housing unit listed on a designated sample line, were collected. If an AA contained any blocks that would qualify for a 1-in-2 or 1-in-4 rate, all households in the AA were sampled at 1-in-2. Housing units in all other AAs were sampled at 1-in-6.

Housing units in American Indian reservations, Tribal Jurisdiction Statistical Areas (now known as Oklahoma Tribal Statistical Areas), and Alaska Native villages were sampled according to the same criteria as other LFSEs, except the size estimates of these LFSEs were based on the American Indian and Alaska Native population in those areas, as measured in the 1990 census. Trust lands were sampled at the highest rate of any part of their associated American Indian reservations. If the associated American Indian reservation was entirely outside the state containing the trust land, then the trust land was sampled at a 1-in-2 rate. All remote Alaska assignment areas were sampled at a rate of 1-in-2. All housing units in Puerto Rico were sampled at a 1-in-6 rate.

Variable sampling rates provide relatively more reliable estimates for small areas and decrease respondent burden in more densely populated areas, while maintaining data reliability. When all sampling rates were taken into account across the nation, approximately 1 out of every 6 housing units was included in the Census 2000 sample.

## **ESTIMATION PROCEDURE**

The weights that appear on the PUMS files are the product of the long form weight and the PUMS sampling weight. The long form weights were obtained from an iterative ratio estimation procedure (iterative proportional fitting) resulting in the assignment of a weight to each sample person and housing unit record. For any given tabulation area, a characteristic total was estimated by

summing the weights assigned to the people or housing units possessing the characteristic in the tabulation area. Estimates of family or household characteristics were based on the weight assigned to the family member designated as householder. Each sample person or housing unit record was assigned exactly one weight to be used to produce estimates of all characteristics. For example, if the weight given to a sample person or housing unit had the value 6, all characteristics of that person or housing unit would be tabulated with a weight of 6. The estimation procedure, however, did assign weights varying from person to person and housing unit to housing unit.

The estimation procedure used to assign the weights was performed in geographically defined *weighting areas*. Generally, weighting areas were formed of contiguous geographic units within counties. Weighting areas were required to have a minimum sample of 400 people. Also, weighting areas never crossed county boundaries. In small counties with a sample count below 400 people, the minimum sample size condition was relaxed to permit the entire county to become a weighting area.

Augmentation of the Census 2000 sample occurred in a relatively small number of weighting areas where the realized sample size was determined to be inadequate. A systematic sample of person and housing unit records was selected and sample data was imputed for these records.

### People

Within a weighting area, the long form sample was ratio-adjusted to equal the 100-percent totals for certain data groups. There were four stages of ratio adjustment for people. The first stage used 21 household-type groups. The second stage used three groups with the following sampling rates: 1-in-2, 1-in-4, and less than 1-in-4. The third stage used the dichotomy householders/nonhouseholders and the fourth stage used 312 aggregate age-sex-race-Hispanic origin groups. The stages were defined as follows:

#### Stage I: Type of Household

Group	Family with own children under 18: Number of people in housing unit
1 . . . . .	2
2 . . . . .	3
3 . . . . .	4
4 . . . . .	5
5 . . . . .	6-7
6 . . . . .	8 or more
	Family without own children under 18:
7-12 . . . . .	2 through 8 or more
	All other housing units:
13 . . . . .	1
14-19 . . . . .	2 through 8 or more
20 . . . . .	People in group quarters
21 . . . . .	Service Based Enumerations

#### Stage II: Sampling Type

Group	
1 . . . . .	1-in-2
2 . . . . .	1-in-4
3 . . . . .	1-in-6 or 1-in-8

---

### Stage III: Householder Status

Group	
1 .....	Householder
2 .....	Nonhouseholder

### Stage IV: Age/Sex/Race/Hispanic origin

People of Hispanic origin: Black or African American: Male:

Group	Age
1 .....	0-4
2 .....	5-14
3 .....	15-17
4 .....	18-19
5 .....	20-24
6 .....	25-29
7 .....	30-34
8 .....	35-44
9 .....	45-49
10 .....	50-54
11 .....	55-64
12 .....	65-74
13 .....	75+
14-26 .....	Female: Same age categories as 1-13
27-52 .....	American Indian or Alaska Native: Same gender and age categories as 1-26
53-78 .....	Asian: Same gender and age categories as 1-26
79-104 .....	Native Hawaiian or Pacific Islander: Same gender and age categories as 1-26
105-130 .....	White: Same gender and age categories as 1-26
131-156 .....	Some Other Race: Same gender and age categories as 1-26
157-312 .....	People not of Hispanic origin: Same race, gender, and age categories as 1-156

Note: Multiple race respondents were included in one of the six race groups for estimation purposes only, however the PUMS files include the full set of responses to the race item.

The ratio estimation procedure for people was conducted within a weighting area in four stages. Prior to performing the four stage adjustment, the following steps were taken:

1. Each sample person record was assigned an initial weight approximately equal to the inverse of the observed sampling rate for the weighting area.
2. Prior to iterative proportional fitting, the categories within each final weighting area described above were combined, if necessary, to increase the reliability of the ratio estimation procedure. Any group that did not meet pre-specified criteria for the unweighted sample count or for the ratio of the 100-percent to the initially weighted sample count was combined with another group according to a specified collapsing pattern. There was an additional criterion concerning the number of complete count people in each race/Hispanic origin category in the second estimation stage.

### Ratio Adjustment

The initial weights underwent four stages of ratio adjustment applying the grouping procedures described above.

**Stage I.** At the first stage, the ratio of the complete census count to the sum of the initial weights for each sample person was computed for each Stage I group. The initial weight assigned to each person in a group was then multiplied by the Stage I group ratio to produce an adjusted weight.

**Stage II.** The Stage I adjusted weights were again adjusted by the ratio of the complete census count to the sum of the Stage I weights for sample people in each Stage II group.

**Stage III.** The Stage II weights were adjusted by the ratio of the complete census count to the sum of the Stage II weights for sample people in each Stage III group.

**Stage IV.** The Stage III weights were adjusted by the ratio of the complete census count to the sum of the Stage III weights for sample people in each Stage IV group.

The four stages of ratio adjustment were repeated in the order given above until the predefined stopping criteria were met. The weights obtained from the final iteration of Stage IV were assigned to the sample person records. However, to provide whole numbers of persons and housing units for tabulated data, integer weights were assigned. For example, if the final weight of the people in a particular group was 7.25, then 1/4 of the sample people in this group were randomly assigned a weight of 8, while the remaining 3/4 received a weight of 7.

### Housing Units

The ratio estimation procedure for housing units was essentially the same as that for people, except that vacant housing units were treated separately. The occupied housing unit ratio estimation procedure was done in three stages. The first stage for occupied housing units used 19 household type groups while the second stage used three sampling type groups. The third stage used 24 race Hispanic origin-tenure groups. The vacant housing unit ratio estimation procedure was done in a single stage with three groups. The stages for ratio estimation for housing units were as follows:

#### Occupied Housing Units

##### Stage I: Type of Household

Group	Family with own children under 18: Number of people in housing unit
1 .....	2
2 .....	3
3 .....	4
4 .....	5
5 .....	6-7
6 .....	8 or more
	Family without own children under 18:
7-12 .....	2 through 8 or more
	All other housing units:
13 .....	1
14-19 .....	2 through 8 or more

##### Stage II: Sampling Type

Group	
1 .....	1-in-2
2 .....	1-in-4
3 .....	1-in-6 or 1-in-8

### Stage III: Race and Hispanic Origin of Householder/Tenure

Group	Owner: Hispanic origin:
1 . . . . .	Black or African American
2 . . . . .	American Indian or Alaska Native
3 . . . . .	Asian
4 . . . . .	Native Hawaiian or Pacific Islander
5 . . . . .	White
6 . . . . .	Some Other Race
7-12 . . . . .	Owner: Not of Hispanic origin: Same race categories as 1-6
13-24 . . . . .	Renter: Same Hispanic origin and race categories as 1-12

### Vacant Housing Units

Group	
1 . . . . .	Vacant for rent
2 . . . . .	Vacant for sale
3 . . . . .	Other vacant

As was done for persons, both occupied and vacant housing unit records were assigned an initial weight, and the groupings within each final weighting area went through a similar collapsing procedure.

The weights produced by this estimation procedure realize some of the gains in sampling efficiency that would have resulted if the population had been stratified into the ratio-estimation groups before sampling, and if the sampling rate had been applied independently to each group. The net effect is a reduction in both the standard error and the possible bias of most estimated characteristics to levels below what would have resulted from simply using the initial, unadjusted weight. Also, this estimation procedure produces estimates that are consistent with the complete count of persons and housing units at the county level and higher.

### SELECTION OF THE PUBLIC USE MICRODATA SAMPLES

A stratified systematic selection procedure with equal probability was used to select each of the public use microdata samples. The sampling universe was defined as all occupied housing units including all occupants, vacant housing units, and group quarters people in the census sample. The sample units were stratified during the selection process. The stratification was intended to improve the reliability of estimates derived from the public use microdata samples by defining strata, within which there is a high degree of homogeneity among the census sample households with respect to characteristics of major interest.

The occupied housing unit stratification was performed using a matrix containing 34,080 cells made by combining 71 race groups, 5 Hispanic origin groups, 3 family types, 2 tenure groups, 4 groups based on maximum age of household members, and the 4 long form sampling rates. In the case of occupied housing units the primary sampling units selected by the systematic selection process are housing units and all person records are extracted after the housing units are chosen. Therefore, the race and Hispanic origin correspond to the householder. The maximum age variable, in contrast, can come from any household member. For group quarters people, the race, Hispanic origin, and age will be those of the individual group quarters person. Table A contains a representation of the occupied housing unit stratification matrix.

The vacant housing unit stratification was performed within a matrix consisting of 12 cells made by combining the four long form sampling rates with three vacancy statuses. Table B contains a representation of the vacant housing unit stratification matrix.

---

The group quarters stratification used a matrix of 2,840 cells made by combining 71 race groups, five Hispanic Origin groups, four group quarters person age groups, and two types of group quarters. Table C contains a representation of the group quarters person stratification matrix.

### **SUBSAMPLING THE PUMS FILES**

The sample selection procedures were performed separately for each of the three subsampling universes: occupied housing units (including all people in them), vacant housing units, and group quarters persons, as follows. The number of 1-percent public use microdata samples for a given state was determined by the full census sample size for that state. For instance, if the full census sample for a state was 20 percent, then the census sample was divided into 20 subsamples of approximately equal size. The 1-percent public use microdata sample was designated at random from the 20 subsamples. From the remaining 19 subsamples, five 1-percent subsamples were designated at random and merged to produce the 5-percent public use microdata sample.

During the sample selection operation, consecutive two-digit subsample numbers from 00 to 99 were assigned to each sample case in the 5-percent and 1-percent samples to allow for the designation of various size subsamples and, as discussed in the preceding chapter, to allow for the calculation of standard errors. As an example, for a 1-percent public use microdata sample, the choice of records having subsample numbers with the same “units” digit (e.g., the two “units” digit includes subsample numbers (2,12,22,....,92) will provide a 1-in-1000 subsample.

Samples of any size between 1/20 and 1/10000 may be selected in a similar manner by using appropriate two-digit subsample numbers assigned to either of the microdata samples. Care must be exercised when selecting such samples. If only one “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-10 subsample but one that would be somewhat more clustered and as a result subject to larger sampling error.

### **SERVICE-BASED ENUMERATION**

Service Based Enumeration was designed to account for people without a usual residence who use service facilities (i.e., shelters, soup kitchens and mobile food vans). Only people using the service facility on the interview day were enumerated. In addition, people enumerated in Targeted Non-Shelter Outdoor Locations (TNSOLS) and people without a usual residence that filed Be Counted Forms (BCF) augmented the enumeration. Note that only people enumerated in shelters and soup kitchens were eligible for selection in the initial census sample. **This component of the enumeration should not be interpreted as a complete count of the population without a usual residence.**

**Table A: Census 2000 PUMS Stratification Matrix - Occupied Households**

Sampling rate (1-in-2, 1-in-4, 1-in-6, 1-in-8)					
Household type	Maximum age in HH	White		.....	(71 detailed race groups)
		Hispanic origin (5 cells for the 5 categories)		.....	
		Owner	Renter	.....	(Tenure)
Family with own children under 18 .	0-59				
	60-74				
	75-89				
	90+				
Family without own children under 18 .	0-59				
	60-74				
	75-89				
	90+				
Other household (nonfamily) .....	0-59			.....	
	60-74			.....	
	75-89			.....	
	90+			.....	

**Table B: Census 2000 PUMS Stratification Matrix - Vacant Housing Units**

Vacancy status	Sampling rate			
	1-in-2	1-in-4	1-in-6	1-in-8
Vacant, for sale.....				
Vacant, for rent.....				
Vacant, other.....				

**Table C: Census 2000 PUMS Stratification Matrix - Group Quarters People**

GQ type	Institutional or military				Noninstitutional and Nonmilitary			
	White/other	Black	.....	.....	.....	(71 Detailed race groups)		
Race	Hispanic origin (5 categories)	.....	.....	.....	.....			
Hispanic origin/age								
0-59.....								
60-74.....								
75-89.....								
90+.....								