DATA
BROWSE AND SELECT DATA
DOWNLOAD OR REVISE EXTRACTS
ANALYZE DATA ONLINE
IPUMS ABACUS
IPUMS REGISTRATION

DOCUMENTATION
VARIABLES
SAMPLES
USER'S GUIDE
GEOGRAPHIC TOOLS
FAQ

RESOURCES
ENUMERATION FORMS
PUBLISHED CENSUS VOLUMES
ERRATA AND REVISIONS

RESEARCH
CITATION AND USE
BIBLIOGRAPHY
RELATED SITES

CONTACT US
HELP
USER FORUM
IPUMS STAFF
HOW TO HELP
DONATE TO IPUMS

SAMPLE DESIGNS

Go back to the IPUMS User's Guide

The component samples of the IPUMS employ a variety of sample designs. The links below lead to concise descriptions of the design of samples from each year. Most of the short descriptions provide links to more detailed descriptions, when available.

- Overview of sample design
- 1850
- 1860 and 1870
- 1880
- 1900
- 1910
- 1910 Puerto Rico
- 1920
- 1920 Puerto Rico
- 1930
- 1940
- 1950
- 1960
- 1970
- 1970 Puerto Rico
- 1980
- 1980 Puerto Rico
- 1990
- 1990 Puerto Rico
- 2000 (United States and Puerto Rico)
- 2010 (United States and Puerto Rico)
- American Community Survey (ACS) and Puerto Rico Community Survey (PRCS)

OVERVIEW OF SAMPLE DESIGN

All the IPUMS samples are **cluster samples**. The most interesting census information describes characteristics of individuals, but the IPUMS samples are not individual-level samples; instead, they are samples of **households** or **dwellings**. Individuals are sampled as parts of households because many important topics of analysis-such as fertility, household composition, and nuptiality-require information about multiple individuals within the same household.

All the samples are also **stratified** to some degree. That is, they divide the population into strata based on key characteristics, and then sample separately from each stratum. This ensures that each stratum is proportionately represented in the final sample. The samples for years prior to 1960 are geographically stratified: the original source materials were divided up prior to sampling in such a way as to ensure a more even geographical distribution of cases than would be expected from a true random sample. The 1960 and subsequent samples employed more elaborate stratification schemes based not only on geography but also on such characteristics as household size, race, and group quarters membership. The effects of clustering and stratification are described in more detail in "Sampling Error." The ways in which each sample is stratified are described in the sample-by-sample discussion below.

The sample designs for all years are constrained by the available **units of enumeration**. In the pre-1940 censuses, all individuals were assigned to a **family**. The definition of the census family varied only slightly from census to census between 1850 and 1930. Generally, a family was an individual or group of individuals living together in the same dwelling place. Two or more families could reside in a single structure, provided they occupied separate parts of it and their housekeeping was separate. However, all the permanent occupants of any hotel, military barracks, or large institution, were considered members of a single family. Census enumerators likewise counted boarders, lodgers, and servants as part of the family occupying the dwelling place where they slept, regardless of their housekeeping arrangements.

In 1940 the basic unit of enumeration shifted from the census family to households and quasi-households. A **household** consisted of the group of persons occupying a dwelling place or part of a dwelling place with either separate cooking equipment or an outside entrance. The maximum number of boarders and lodgers in a household was ten; if the dwelling place contained more than ten boarders and/or lodgers, it was enumerated as a **quasi-household**. Quasi-households also included hotels, military barracks, dormitories, and other large institutions. The 1950 census was similar, except that quasi-households included all units with five or more persons unrelated to the household head.

In the years since 1950, the term **group quarters** has been substituted for the term quasi-household, and there have been minor variations in the criteria used to distinguish separate households within the same structure. In 1980, the number of unrelated persons required for group quarters classification was raised from five to ten.[1]

For the census years before 1950, the sample units - households and dwellings - are actually subsets of the original enumeration units. The following sample-by-sample discussion describes the sample units used for each census year, and briefly defines the procedures used to select cases from the original sources for inclusion in each sample.

1850: The manuscript census of the 1850 free population consists of roughly 560,000 census pages recorded on 976 reels of microfilm. Each census page has eighty-four lines, and the information pertaining to each individual appears on a separate line.

The sample was drawn systematically from each microfilm reel, ordinarily at intervals of six pages. On each selected census page, one line was randomly selected and designated as the sample point. Any valid sample unit beginning at the sample point or within four subsequent lines was included in the sample, yielding a 1-in-100 sample with equal probabilities of inclusion for all individuals and households. Valid sample units are defined as follows:

1. **Dwellings**: structures containing fewer than 31 residents, with or without multiple families.
2. **Households**: census families with fewer than 31 members in dwellings containing 31+ residents.
3. **Related groups in large units**: groups related by blood or marriage in census families with 31+ members. Family relationships are inferred from surnames.
4. **Individuals in large units**: unrelated individuals in census families with 31+ members.

For more information on the 1850 Census sample design, click here.

1860 and 1870: The manuscript censuses for 1860 and 1870 consist of approximately two million census pages recorded on 3,186 reels of microfilm. Each census page has eighty lines, and the information pertaining to each individual appears on a separate line.

The 1860 and 1870 samples employ the same sampling scheme as the 1850 sample, except that households containing any black person are sampled at twice the density of other households (1-in-50).

1880: The IPUMS offers three samples of the 1880 census: the 1880 1% sample, 1880 5% sample with oversamples (preliminary version), and the 1880 100% database. The 1% and 5%

samples may combined for a total of a 6% sample. The 100% database was originally constructed for genealogical purposes and therefore has limited variable availability.

The manuscript census for 1880 consists of about 600,000 enumeration pages. Each form consisted of an "A" side and a "B" side, with each side containing 50 lines. These records are contained on 1,454 reels of microfilm.

For the 1880 1% sample, one line was randomly selected from each 100-line page and designated as the sample point. Sample units were included only if they began at the designated sample point. Valid sample units were defined the same as in 1850, except that related groups in group quarters could be identified through the family relationship variable as well as by surname. This procedure yielded a 1-in-100 sample with equal probabilities of inclusion for all individuals, households, and dwellings. For more information on the 1880 1% sample design, click here.

For the 5% 1880 sample with 10% minority oversamples, each 100-line page was divided into four 25-line sampling windows (e.g., 1-25, 26-50, 51-75, and 76-100). Every 10th window encountered was considered for sampling. Any household that began on one of the 25 lines of a 10th window was included in the sample.

The Preliminary sample includes data only from odd-numbered microfilm reels (e.g., 1, 3, 5, 7). Each reel generally contains census manuscripts for several counties, with counties organized alphabetically by state. For this reason, counties on even-numbered reels are not represented in this dataset. All counties will be represented in the final 10% 1880 sample.

The sampling strategy produced a 5% sample of all households. Additionally, any minority household that began on one of the 25 lines of a 10th+1 window were also included in the sample, thus producing an oversample of minority households. This produced an additional 5% sample of minority households, for a total of a 10% minority household sample.

For the purposes of sampling in the 1880 5% sample, minority households were defined as a households that contained at least one individual whose race was Native American or African-American, whose race or birthplace indicated that they were Chinese, or whose name or birthplace indicated Hispanic origins. In the 1880 5% sample, households not including a minority have PERWT/HHWT values of 20 (since they are part of a 5% sample), where as households including a minority have PERWT/HHWT values of 10 (since they are part of a 10% sample).

The 1880 100% population database was originally entered by the Church of Jesus Christ of Latter Day Saints. The data was cleaned and coded at the Minnesota Population Center. The cleaned dataset was first released by the North Atlantic Population Project (NAPP). Several key variable groups were never entered, including items relating to school, literacy, unemployment, disability, month of birth, marriage within the past year, and street address. Group quarters containing more than 60 people were split into 1-person households. The Minnesota Population Center is currently in the process of entering all variables for a 10% sample of the database.

As a part of IPUMS data processing for the 1880 100% database, group quarters units containing more than 60 people were split into 1-person households, each with their own unique value in the IPUMS SERIAL variable. For this reason, the SERIAL variable makes it impossible to identify any group of people who share a group quarters unit containing more than 60 people. Researchers needing to analyze intact group quarters units containing more than 60 people (such as the residents of a large prison or poorhouse) should use SERIAL80 and PERNUM80, instead of SERIAL and PERNUM.

1900: The IPUMS offers three samples of the 1900 census: the 1900 1% sample, 1900 1% sample with oversamples, and the 1900 2.5% sample (preliminary version). The 1% samples contain many of the same cases as one another and may not be combined. Either of the 1% samples may be combined with the 2.5% sample, for a total sample size of 3.5%.

The two 1900 1% IPUMS samples include different combinations of data from four independently drawn component samples. Each of the four component samples can be identified by using the SAMP1900 variable. The component samples include:

- a 1-in-100 national sample
- a 1-in-5 Alaskan sample
- a 1-in-5 Hawaiian sample
- a 1-in-5 sample of the American Indian schedules

The 1900 1% sample with oversamples consists of all records from all four samples, with appropriate weights assigned. Since the oversamples are at a higher density than the national sample, users MUST apply weighting variables such as PERWT and HHWT to get accurate statistics when using this data.

The 1900 1% sample also consists of data from all samples listed above. To make the sample self weighting (i.e., all records have the same weight), we selected approximately 5% percent of the records from each of the 1-in-5 samples. The 1900 1% Unweighted sample does not require the use of weights.

The 1900 2.5% preliminary sample draws on 1-in-20 households from every odd-numbered reel. Since each reel contains one or more counties, entire counties are excluded from this dataset. The final 1900 5% dataset will include data from all counties. The 2.5% sample may be combined with either of the 1% samples.

The paragraphs below describe the sampling strategies for the four component samples of the 1900 IPUMS samples.

The 1900 manuscript census consists of some 900,000 census pages contained on 1,854 microfilm reels. Each page contains one hundred lines. The sample 1-in-100 national sample was drawn systematically from each microfilm reel, ordinarily at intervals of five pages. On each selected census page, one line was randomly selected and designated as the sample point. Any valid sample unit beginning at the sample point or within four subsequent lines was included in the sample, yielding a 1-in-100 sample with equal probabilities of inclusion for all individuals and households. Valid sample units are defined as follows:

1. **Dwellings**: structures containing fewer than 31 residents, with or without multiple families.
2. **Households**: census families with fewer than 31 members in dwellings containing 31+ residents.
3. **Related groups in large units**: groups related by blood or marriage in census families with 31+ members. Family relationships are inferred from relationship to head information and surnames.
4. **Individuals in large units**: unrelated individuals in census families with 31+ members.

The oversamples of Alaskans, Hawaiians, and American Indians followed the same rules as the 1-in-100 national sample , with the following exceptions:

The schedules for Alaska and Hawaii have 50 lines per page number, numbered 1-25 on the A side and 26-50 on the B side. Five-line sample windows were randomly generated for every side of the census page.

The schedules used to enumerate the American Indian population had 40 line per page. Eight-line sample windows were randomly generated for every side of the census page.

The 1900 1% sample and the 1900 1% sample with oversamples include a small number of records that have a PERWT of 0. These records are part of fragmentary households. Some members of these households were located in sample windows and were originally sampled as household fragments (see SAMPRULE). These records received a non-zero PERWT. However, other members of these households were enumerated elsewhere on the original manuscripts (i.e., not contiguous with the sample window). When possible, we located these individuals and reunited them with the remainder of their household and assigned a PERWT of 0. Adding these records was useful in order to construct household level variables that require information about all members of a given household.

1910: The IPUMS offers two samples of the 1910 census: the 1910 1% sample and the 1910 1.4% sample with oversamples.

The two 1910 IPUMS samples include different combinations of data from seven independently drawn component samples. Each of the seven component samples can be identified by using the SAMP1910 variable. The component samples include:

- a 1-in-100 national sample
- a 1-in-5 Alaskan sample
- a 1-in-5 Hawaiian sample
- a 1-in-5 sample of the American Indian schedules
- a 1-in-250 national sample
- a Black oversample of selected counties
- a Hispanic oversample of selected counties

The 1910 1.4% sample with oversamples consists of all records from all seven samples with appropriate weights. The samples employed a variety of different sampling schemes and densities, so users MUST apply weighting variables such as PERWT and HHWT to get accurate statistics when using this data.

The 1910 1% sample consists of data from the first four samples listed above. To make the sample self weighting (i.e., all records have the same weight), we selected approximately 5% percent of the records from each of the 1-in-5 samples. The 1910 1% sample does not require the use of weights.

The first four component samples listed above were produced at the University of Minnesota between 2001-2006. The last three component samples listed above were produced at the University of Minnesota, the University of Texas, and the University of Pennsylvania in the 1980s and 1990s. The last three samples were previously available via the IPUMS Extract system as the "1910 General sample," the "1910 Black Oversample," and the "1910 Hispanic Oversample."

Each of the component samples employs slightly different sampling rules and sampling densities. More information about how each of the component samples was drawn is available here.

1910 Puerto Rico: The IPUMS offers a 12% sample of the 1910 Puerto Rican census. This sample consists of three component samples:

- a 1-in-10 sample of households outside the municipality of Loiza and the coffee regions (described below)
- a 1-in-5 sample of households in the municipality of Loiza (an African-descent enclave)
- a 1-in-5 sample of households located in coffee regions. Coffee regions in 1910 consisted of the following municipalites: Adjuntas, Ciales, Lares, Las Marias, Maricao, Mayaguez, San Sebastian, Utuado; and the following townships within the following municipalities; townships are listed first, and the corresponding municipality is listed following the township name in parentheses: Casy Arriba (Anasco), Cercado (Anasco), Corcobada (Anasco), Corcovada (Anasco), Rio Arriba (Anasco), Caonillas Arriba (Juana Diaz), Collores (Juana Diaz), Hato Puerco Arriba (Juana Diaz), Vacas (Juana Diaz), Villalba Arriba (Juana Diaz), Anon (Ponce), Guaraguao (Ponce), San Patricio (Ponce), Frailes (Yauco), Naranjo (Yauco), Reio Prieto (Yauco), and Rubias (Yauco).

Users MUST apply weights when using these samples (see PERWT and HHWT).

The original data for the 1910 Puerto Rican sample was compiled by the University of Wisconsin Madison Survey Center between January 2002 to July 2005. The sample was created with the purpose of producing public use samples analogous and comparable to the United States census samples available through the IPUMS. The original UW dataset is available from ICPSR as Study Number 4343.

The 1910 Puerto Rican census manuscripts are stored on 28 microfilm reels. The images on each reel are divided into two separate sequences, where each sequence contains about 415 numbered pages. Each page contains 50 lines, the first 25 lines on the "A" side and the second 25 lines on the "B" side.

1920: The manuscript census of the 1920 population consists of about 1.2 million census pages recorded on 2,076 reels of microfilm. Each census page has one hundred lines, and the information pertaining to each individual appears on a separate line.

The 1920 sample employs the same sampling scheme as the 1850 sample, except that family relationships as well as surnames are used to identify related groups in group quarters.

1920 Puerto Rico: The IPUMS offers a 12% sample of the 1920 Puerto Rican census. This sample consists of three component samples:

- a 1-in-10 sample of households outside the municipality of Loiza and the coffee regions (described below)
- a 1-in-5 sample of households in the municipality of Loiza (an African-descent enclave)
- a 1-in-5 sample of households located in coffee regions. Coffee regions in 1920 consisted of the following municipalities: Adjuntas, Ciales, Lares, Jayuya, Lares, Las Marias, Maricao, Mayaguez, San Sebastian, Utuado; and the following townships within the following municipalities; townships are listed first, and the corresponding municipality is listed following the township name in parentheses: Casey Arriba (Anasco), Cercado (Anasco), Corcovado (Anasco), Rio Arriba (Anasco), Collores (Juana Diaz), Anon (Ponce), Anon/Oeste (Ponce), Guaraguao (Ponce), San Patricio (Ponce), Caonillas Arriba, Parte Norte (Villalba), Caonillas Arriba, Parte Sur (Villalba), Hato Puerco Arriba, Parte Noroeste (Villalba), Hato Puerco Arriba, Parte Sureste (Villalba), Vacas, Parte Este (Villalba), Vacas, Parte Oeste (Villalba), Villalba Arriba (Villalba), Villalba Arriba, Parte Noreste (Villalba), Villalba Arriba, Parte Sur (Villalba), Frailes (Yauco), Naranjo, East (Yauco), Naranjo, Parte Oeste (Yauco), Rio Prieto (Yauco) and Rubias (Yauco).

Users MUST apply weights when using these samples (see PERWT and HHWT).

The original data for the 1920 Puerto Rican sample was compiled by the University of Wisconsin Madison Survey Center between January 2002 to July 2005. The sample was created with the purpose of producing public use samples analogous and comparable to the United States census samples available through the IPUMS. The original UW dataset is available from ICPSR as Study Number 4344.

The 1920 Puerto Rican census manuscripts are stored on 33 microfilm reels. The images on each reel are divided into three separate sequences, where each sequence contains about 265 numbered pages. Each page contains 50 lines, the first 25 lines on the "A" side and the second 25 lines on the "B" side.

1930: The manuscript census of the 1930 population consists of roughly 1,240,000 census pages recorded on 2625 reels of microfilm. Each census page has one hundred lines, and the information pertaining to each individual appears on a separate line.

The sample was drawn systematically from each microfilm reel. Although the interval was ordinarily every fifth census page, it is useful to think of the interval as every tenth side of a census page. Each census page consists of an 'A' and a 'B' side (typically the 'A' side has lines 1-50, and the 'B' side has lines 51-100). The sample window consisted of either the top half (lines 1-25 or lines 51-75) or the bottom half (lines 26-50 or lines 76-100) of every tenth side. In addition, the window alternated between top halves and bottom halves of page sides. Thus, we have the equivalent of 25 sample lines per five census pages (or 500 lines). This sample design will ultimately result in a 1-in-20 sample. The current release, however, is a 1-in-100 sample. For this release, we generated five-line sample windows within the original 25-line windows for sample inclusion.

The sample design described above is generally comparable to other samples constructed at the Minnesota Population Center. The basic premise is that any valid sample unit beginning in the 25-line sample window (or five-line window for the current 1-in-100 release) was included in the sample. One significant difference is that we no longer sample all dwellings in their entirety if the dwelling contained fewer than 31 persons. Thus, the definition of valid sample units becomes:

1. **Households**: census families with fewer than 31 members (regardless of dwelling size).
2. **Related groups in large units**: groups related by blood or marriage in census families with 31+ members. Family relationships are inferred from surnames.

3. **Individuals in large units**: unrelated individuals in census families with 31+ members.

1940: The population schedules of the 1940 census are preserved on 4,576 microfilm reels. Each census page contains information on forty individuals. Two lines on each page were designated as "sample lines" by the Census Bureau; the individuals falling on those lines - 5 percent of the population - were asked a set of supplemental questions that appear at the bottom of the census page.

Two of every five census pages were systematically selected for examination. On each selected census page, one of the two designated sample lines was randomly selected. Data entry personnel then counted the size of the sample unit containing the targeted sample line. Sample units containing fewer than seven persons were included in the sample in inverse proportion to their size. Thus, every one-person unit was included in the sample, every second two-person unit, every third three-person unit, and so on. Units with seven or more persons were included with a probability of 1-in-7: every seventh household of size seven or more was selected for the sample. Sample units for 1940 were defined as follows:

1. **Households:** dwelling places with fewer than five persons unrelated to a household head, excluding institutions and transient quarters, were sampled as households.
2. **Group quarters:** dwelling places with five or more persons unrelated to a household head, and individual residents of institutions and transient quarters, were sampled as group quarters-that is, only at the individual level.

This design ensures that each selected sample unit contains one individual who was asked the supplemental sample questions at the bottom of the enumeration form. It yields a flat 1-in-100 sample of persons in units of size seven or less. Persons in units larger than seven are over-represented in the 1940 sample; they must be weighted downward to achieve a representative distribution of household size. Analyses of sample-line individuals who answered supplemental questions must also be weighted. Appropriate weights are included in the IPUMS variables HHWT and SLWT.

For more information on the 1940 Census sample design, click here.

1950: The 1950 census schedules are contained on 6,278 microfilm reels. Each census page contains information on thirty individuals. Every fifth line on the census page was designated as a sample line, and additional questions for the sample-line individuals on each page appear at the bottom of the form. For the last sample-line individual on each page, there was a block of additional supplemental questions. Thus, 20 percent of individuals were asked a basic set of supplemental questions, and 3.33 percent of individuals were asked a full set of supplemental questions.

One in eleven pages within each enumeration district was selected randomly. On each selected census page, the sixth sample-line individual (the one with the full set of questions) was selected for inclusion in the sample. Any other members of the sample unit containing the selected individual were also included. Sample units are defined as in the 1940 sample.

As in the 1940 sample, each household in the 1950 sample includes one individual who was asked supplemental questions. The sampling procedure yielded a flat 1-in-330 sample of these sample-line individuals. But the sampling procedure is not flat for persons who were not sample-line individuals. The probability of inclusion in the sample is directly proportional to the size of the unit. Thus, when analyzing the entire population of the persons in units with more than one individual, cases must be weighted in inverse proportion to household size. An appropriate weight is included in the IPUMS variables HHWT and SLWT.

For more information on the 1950 Census sample design, click here.

1960: The 1960 census used a machine-readable household form instead of the traditional census schedule. Census information was collected on a separate form for each housing unit. For the first time, housing questions were included on the same form as the population items, and are thus included in the census samples. Every fourth enumeration unit received a "long form," which contained supplemental sample questions that were asked of all members of the unit. Since the public use microdata files are drawn entirely from these long forms, the sample questions are available for all individuals in every unit, instead of for a single member of each unit as in 1940 and 1950. Of the units receiving a long form, four-fifths received one version (the 20% questionnaire), and one-fifth received a second version with the same population questions but slightly different housing questions (the 5% questionnaire). The 1-in-100 1960 sample is drawn from both questionnaires. 1960 sample units are defined the same as for the 1940 and 1950 samples.

The sample employed a three-step procedure to select cases from the long-form questionnaires, which collectively formed a 25 percent sample. First, the entire census was divided into 33,000 geographic units, called smallest weighting areas (SWAs). The population of each SWA was broken into 44 categories, based on age group, sex, race, headship, and home ownership. For each category a weight was calculated representing the ratio of persons in the full population count to persons in the 25 percent sample. These weights were used in calculating most census tabulations of sample characteristics for small geographic areas.

Next, the sample weights generated for each SWA were used to select a stratified 5 percent sample from the 25 percent sample. The 25 percent sample of the long forms was divided into 38 strata, based on household size, home ownership, race, and group quarters residence. Within each stratum, the cumulative sum of weights for each household head was calculated, and a case was selected for inclusion in the sample each time the cumulative sum passed a multiple of twenty. This procedure yielded a flat 5 percent sample that was used to produce many of the census publications pertaining to the general population. As part of the Minnesota Population Center's and Census Bureau's 1960 Restoration Project, a new 5% Public-Use Microdata Sample (PUMS) was drawn from a restored 25 percent internal sample. The new 5% PUMS has a sample design consistent with original 5% PUMS flat sample.

Finally, a 1 percent sample was selected from the 5 percent sample, using essentially the same procedure to select every fifth case within each of 38 strata. The strata used in this selection were the same as those used to select the five percent sample, except that they employed a slightly different classification of household size. The 1 percent 1960 sample is divided into 100 subsamples, each of which incorporates the same stratification. This elaborate three-step selection scheme yielded a flat sample with very small standard errors, especially for race and home ownership.

For more information on the 1960 Census sample design, click here.

1970: Sample units in the 1970 samples are defined the same as in the 1940, 1950, and 1960 samples. One in five housing units in 1970 received a long form containing supplemental sample questions. There were two versions of the long form, with different inquiries on both housing and population items; 15 percent of households received one version, and 5 percent received the other. Six independent 1 percent public use samples were produced for 1970, three from the 15 percent questionnaire and three from the 5 percent questionnaire. Each of the three samples drawn from each questionnaire provide somewhat different geographical information.

The procedures used to select cases for inclusion in the 1970 public use samples were similar to those used in 1960 but were slightly more elaborate. Again, weights were constructed for the SWA as the ratio of persons with selected characteristics in the full population count to persons with the same characteristics in the 15 percent and 5 percent samples. In 1970, these weights were calculated in three stages that controlled for household size, sex of head, presence of own children of head, group quarters residence, headship, race, age, and sex.

To select the six 1 percent samples (three from the 15 percent sample and three from the 5 percent sample), the weighted population for each sample was divided into seventy-five strata, based on home ownership, race, sex of head, household size, presence of own children, inmate status, and other residence in group quarters. Within each stratum, the sum of weights for household heads was cumulated. The weights represent the ratio of persons in the full count to persons in each sample; because three 1 percent extracts were required for each sample, a case was selected each time the cumulated total of weights passed a multiple of thirty-three. As in 1960, each sample was divided into one hundred subsamples, all of which incorporate the same stratification.

For more information on the 1970 Census sample design, click here.

1970 Puerto Rico: Data collection and sample design in the 1970 Puerto Rican census was nearly identical to that in the 1970 sample described above. There were three major exceptions:

- The 1970 Puerto Rican census contained only one version of the long-form questionaire, administered to 20 percent of the population. The U.S. Census contained two versions of the long-form, one administered to 15 percent of the population and the other administered to 5 percent of the population.
- The Census Bureau's public use microdata samples from Puerto Rico included only 3 percent of the population. This data was released as three independent 1 percent samples: a "State" sample, a "Neighborhood" sample, and a "Municipio" sample.
- The 1970 Puerto Rican census did not collect data on race. This affected techniques used to generate published population estimates and to draw the public use microdata samples. Each of these processes involved dividing the population up into numerous strata. For the Puerto Rican data, these strata were never based on racial characteristcs. This is the only difference between the sampling strategies used to create the Puerto Rican and United States data in 1970.

1980: The 1980 census employed a single long form questionnaire completed by one-half of housing units in places with a population under 2,500 and one-sixth of other housing units. Overall, 19.4 percent of housing units were included in the sample. Sample units were defined the same as in 1970, except that the threshold for sampling as group quarters was raised from five or more persons unrelated to the head to ten or more persons unrelated to the head. Five samples were produced in 1980: a 5% State sample, a 1% Metro sample, a 1% Urban/Rural sample, a 1% Labor Market Area sample, and a 1% Detailed Metropolitan/Nonmetropolitan sample. Each of these samples aims to preserve different types of geographic information.

The 1980 census used the same procedures as the 1970 census to select long-form sample cases for inclusion in the sample, but each step was more elaborate. As in 1970, a three-stage ratio estimation procedure was used to assign weights to sample cases representing the ratio of the full population count to the sample count for persons with particular characteristics in smallest weighting areas. For the 1980 samples, the weights were designed to control for 179 characteristics and combinations of characteristics, including household size, presence of own children, group quarters residence, householder status, detailed race and Spanish origin, age, and sex. The weighted population was divided into 102 strata, including breakdowns by race, Spanish origin, home ownership, sampling rate, and presence of own children. As in 1960 and 1970, cases were selected by cumulating the weights within each stratum, and one hundred stratified subsamples were identified within each of the 1980 samples.

For more information on the 1980 Census sample design, click here.

1980 Puerto Rico: Data collection and sample design in the 1980 Puerto Rican census was nearly identical to that in the 1980 sample described above. There were two major exceptions:

1. The 1980 Puerto Rican census did not collect data on race or Hispanic origin. This affected techniques used to generate published population estimates and to draw the public use microdata samples. Each of these processes involved dividing the population up into numerous strata. For the Puerto Rican data, these strata were never based on race or Hispanic origin.
2. The strata for the estimation and sampling of the 1980 Puerto Rican data contained another slight difference from that used in the 1980 United States data. Specifically, the categories specified in the stratification matrix labelled "Stage II--Tenure/Value or Rent", have different cutoffs in Puerto Rico than they did in the United States. The differing United States and Puerto Rico tables can be viewied on the 1980 sample design page.

1990: The 1990 census used a single long-form questionnaire for sample questions completed by one-half of persons in places with a population under 2,500, one-sixth of persons in other tracts and block numbering areas with fewer than 2,000 housing units, and one-eighth of all other areas. Overall, about one-sixth of housing units completed a long form. Sample units were defined the same as in 1980. Four samples were produced: a 5 percent sample, a 1 percent sample containing somewhat different geographic codes, a 3 percent sample of the elderly, and a 0.5 percent sample of labor market areas.

The ratio estimation procedure used to assign weights to sample cases in 1990 was virtually identical to the procedure used in 1980. The stratification scheme, however, continued the trend toward increasing complexity: the number of separate strata was increased from 102 to 1,049, mainly because of additional detail on age and race.

At this point, the 1990 selection procedure broke with the precedent established in the previous three census years. The previous censuses used the weights to extract a flat sample from each stratum, so the final public use samples had equal probabilities of inclusion for all individuals and households. For 1990, the Census Bureau opted instead to produce weighted samples. Within each state, the Bureau divided the sample questionnaires into an appropriate number of 1 percent samples. For example, if 20 percent of the population of a state completed long forms, the sample questionnaires for that state were divided into twenty subsamples of equal size. Each subsample would then consist of every twentieth case drawn from each stratum. The 5 percent, 1 percent, and 3 percent files were then selected at random from the 1 percent subsamples for each state. Weights were attached to each case representing the number of individuals in the general population represented by any particular case in the sample; these weights range from 0 to 1,138.

The advantage of the weighted sample design adopted for 1990 is that it provides maximum precision for persons residing in small localities. The disadvantages are significant, however. The sample is not only more cumbersome to use than those previously produced by the Census Bureau, but precision is actually reduced for the general population.

For these reasons, the IPUMS provides a 1 percent unweighted sample, extracted from the 1990 5 percent file (the state sample). This was created using the same method that the Census Bureau used to create the 1960 and 1970 samples. We kept a running total of household weights for each of the 100 subsamples in the 1990 5 percent sample. Whenever the sum of a given subsample's household weights crossed the threshold of 100, we included that household (and its associated person records) in the 1 percent unweighted sample. We then subtracted 100 from the running total of household weights for that subsample and continued the process. Since group quarters cases had household weights of 0 in the original 1990 5 percent sample, we used the person weights to decided whether or not to include those cases in the 1 percent unweighted sample.

Using this process, we selected approximately 1-in-5 households from the 1990 5 percent sample to be included in the 1 percent unweighted sample. All cases in the resulting 1 percent unweighted sample have household and person weights of 100; the use of weights is optional with this sample.

For more information on the 1990 Census sample design, click here.

1990 Puerto Rico: Data collection and sample design in the 1990 Puerto Rican census was nearly identical to that in the 1990 sample described above. There was one major exception:

1. The 1990 Puerto Rican census did not collect data on race or Hispanic origin. This affected techniques used to generate published population estimates and to draw the public use microdata samples. Each of these processes involved dividing the population up into numerous strata. For the Puerto Rican data, these strata were never based on race or Hispanic origin. They were otherwise identical to the 1990 United States strata described on the 1990 sample design page.

2000 (United States and Puerto Rico): Approximately 1 out of every 6 housing units in the country were included in the long form Census 2000 sample. Four different sampling rates for housing units were based on occupied housing estimates in a method similar to the 1990 sample design. Census blocks or tracts with less than 800 housing units were sampled at a 1-in-2 rate; blocks with more than 800 and fewer than 1,200 housing units were sampled at a 1-in-4 rate; blocks with more than 1,200 and fewer than 2,000 were sampled at a 1-in-6 rate; and blocks with over 2,000 housing units were sampled at a 1-in-8 rate. The sampling unit is the household and all persons residing in the household. The Census Bureau uses the variable sampling rates to improve the reliability of small area estimates. Nationwide the average sampling rate is 1-in-6 housing units. Two PUMS samples, a 1 percent and a 5 percent file, were produced in 2000 from the long-form data.

Like the 1990 PUMS, the 2000 PUMS is not a self-weighted sample but uses person and household weights to adjust for mixed sampling procedures and differential rates of nonresponse. Weights were assigned to persons and households by ratio adjustment to equal 100-percent totals for certain groups within defined weighting areas. Weighting areas are contiguous areas within counties that have a minimum of 400 people.

The ratio adjustment procedure for persons followed four stages and for households followed three stages with a separate procedure for vacant housing. The stages adjusted for household type, sampling rates, and householder age/race/sex/Hispanic origin groups. The total number of possible strata for housing units and group quarters was 36,932 separate strata (34,080 strata for occupied housing units, 2,840 strata for persons housed in group quarters, and 12 strata for vacant housing units). In addition to gains in efficiency and reduction of possible bias, the weights produce estimates that are consistent with the complete count of persons and housing units at the county level and higher.

The advantage of the weighted sample design adopted for 1990 and 2000 is that it provides maximum precision for persons residing in small localities. The disadvantages are significant, however. The sample is not only more cumbersome to use than those produced by the Census Bureau prior to 1990, but precision is actually reduced for the general population. For these reasons, the IPUMS provides a 1 percent unweighted sample, extracted from the 2000 5 percent file. This was created using the same method that we used to create the 1990 1 percent unweighted sample (see above under '1990' heading).

The sampling for 2000 Puerto Rico data was identical to that for 2000 United States data. For more information on the 2000 Census sample design, click here.

2010 (United States and Puerto Rico): The annual ACS survey was designed to replace the Census long-form, and as a result, the 2010 Decennial Census consisted of a single short-form questionnaire. The short form asked age, sex, race, ethnicity (Hispanic or non-Hispanic), relationship to head, and whether the housing unit was rented or owned by a member of the household. The 2010 Decennial PUMS file is a 10% sample of these responses, which means it contains an extremely limited set of variables. Researchers interested in 2010 data from a long form questionnaire should consult the 2010 ACS.

While the 1990 and 2000 censuses produced weighted samples, in 2010 the Census Bureau released an unweighted sample, which means that the final public-use sample has equal probabilities of inclusion for all individuals and households. Flat samples such as this one have greater precision for the general population, but precision is reduced for persons residing in small localities. Overall, however, self-weighted samples are easier to use, as each person record represents a consistent number of people.

In 2000 a fraction of the population was given the long-form questionnaire, and the PUMS sample was selected from that pool of people, which can be considered a sample of a sample. In 2010, the short-form was sent out universally, so the selection of the public-use data was from the entire population. There were three subsampling universes (occupied housing units, vacant housing units, and persons in group quarters), and the sample selections were completed separately for each of these universes. Selections were made ten times by all 50 states, the District of Columbia, and Puerto Rico, resulting in ten different 10-percent samples for each state. The 10-percent state sample that would be included in the PUMS was selected at random.

The sampling for 2010 Puerto Rico was identical to that for 2010 United States data. For more information on the 2010 Census sample design, click here.

American Community Survey (ACS) and Puerto Rico Community Survey (PRCS): The American Community Survey (ACS) and the Puerto Rico Community Survey (PRCS) are monthly rolling samples of households that were designed to replace the Census long form. Nationally-representative ACS data have been available each year since 2000; PRCS data have been available annually since 2005.

The ACS and PRCS samples include about 3 million households nationwide. About 1 percent of the national group quarters population are also included in the 2006-onward ACS/PRCS samples. The public use samples of the ACS and PRCS are extracted from the Census Bureau's larger internal data files and are thus subject to additional sampling error and further data processing (such as imputation and allocation).

The sampling unit is the household and all persons residing in the household. To protect individual confidentiality, geographic identifiers are currently restricted to the state level, and individual variables, such as income and housing values, are Top coded.

The ACS/PRCS sample design approximates the Census 2000 long form sample design and oversamples areas with smaller populations. Each month a systematic sample is drawn to represent each U.S. county or county equivalent. The selected monthly sample is mailed the ACS/PRCS survey at the beginning of the month. Nonrespondents are contacted via telephone for a computer assisted telephone interview (CATI) one month later. One third of the nonrespondents to the mail or telephone survey are contacted in person for a computer assisted personal interview (CAPI) one month following the CATI attempt. Weights included with the ACS PUMS for the household and person-level data adjust for the mixed geographic sampling rates, nonresponse adjustments, and individual sampling probabilities. Estimates from the ACS IPUMS samples may not be consistent with summary table ACS estimates due to the additional sampling error.

For more information on the ACS sample design, click here.

ENDNOTES

1. For further details on the comparability of census enumeration units, see Steven Ruggles, "Comparability of the Public Use Files of the U.S. Census of Population," *Social Science History*, 15 (1991): 123-158; and Daniel Scott Smith, "The Meanings of Family and Household: Change and Continuity in the Mirror of the American Census," *Population and Development Review*, 18 (1992): 421-456.

Go back to the IPUMS User's Guide

▲ Back to Top

SUPPORTED BY

NIH National Institutes of Health | NSF | ancestry | Stat/Transfer | University of Minnesota