Public Use Microdata Sample (PUMS)

Accuracy of the Experimental Data

2020

INTRODUCTION

The Public Use Microdata Sample (PUMS) are a subset of the 2020 American Community Survey (ACS) sample. Note that for 2020, data from the Puerto Rico Community Survey (PRCS) sample will not be released.

The 2020 PUMS contains a sample of the group quarters (GQ) population. Group quarter data were included in PUMS beginning in 2006. The 2020 PUMS GQ sample includes imputed records. Imputed GQ records were first included in PUMS in 2011. The 2020 ACS selected sample from which PUMS is drawn covers all counties across the nation.

Estimates created from the 2020 PUMS data are expected to be different from the 2020 published ACS estimates because the PUMS data are subject to additional sampling error and further data processing operations. The additional sampling error results from the additional stage of sampling of selecting the PUMS housing and person records.

In the public use file, the basic unit is an individual housing unit, except for the sample from GQs. For the GQ sample, the basic unit is the person. The population sample is defined as all persons living in households selected in the housing unit sample, plus the persons selected from the GQ sample. Note that microdata records in this sample do not contain names, addresses, or any information that can identify a specific housing unit, GQ, or person.

Table of Contents

CONFIDENTIALITY OF THE DATA

The Census Bureau has implemented a series of steps to protect the confidentiality of the data. Title 13 United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's data can be identified.

The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases[1]. A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed.

Title 13, United States Code

Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 is maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to five years in prison and up to $250,000 in fines for wrongful disclosure of confidential census information.

Disclosure Avoidance

Disclosure avoidance is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual that has provided information under a pledge of confidentiality. For data tabulations, the Census Bureau uses disclosure avoidance procedures to modify or remove the characteristics that put confidential information at risk for disclosure.

Data Swapping

Data swapping is a method of disclosure avoidance designed to protect confidentiality. Data swapping is done by editing the source data or exchanging records for a sample of cases when creating a table. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics (such as the same number of adults and same number of children). Because the swap often occurs within a neighboring area, there is no effect on the marginal totals for the area or for totals that include data from multiple areas. Data swapping procedures were first used in the 1990 Census.

Synthetic Data

The goals of using synthetic data are the same as the goals of data swapping, namely to protect the confidentiality. Persons are identified as being at risk for disclosure based on certain

[1] The Census Bureau's Disclosure Review Board approved the 2020 PUMS 1-year data for release with DRB Clearance number CBDRB-FY21-183.

characteristics. The synthetic data technique then models the values for another collection of characteristics to protect the confidentiality of that individual.

PUMAs

The Census Bureau takes further steps to prevent the identification of specific individuals, households, or housing units on the PUMS files. The main disclosure avoidance method used is to limit the geographic detail shown in the files. The smallest geographic unit that is identified is the Public Use Microdata Area (PUMA). The current PUMAs were formed based on data and location collected in the 2010 Census and have been used by the ACS PUMS files since the 2012 data year. With the completion of the 2020 Census, new PUMA boundaries are being prepared.

PUMAs do not cross state boundaries. The Census Bureau provides maps for the PUMAs, and users can identify geographies of interest by zooming in on selected areas. These maps can be found at: https://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_main.html. See the PUMS ReadMe document for step-by-step instructions for using TIGERweb.

Additional Measures

Other disclosure avoidance measures used in the PUMS files include top-coding, age perturbation, weight perturbation, and collapsing of detail for categorical variables. The answers to open-ended questions where an extreme value might identify an individual are top-coded (or bottom-coded). Top-coding (and bottom-coding) substitutes the value of extreme cases with the mean of the highest (or lowest) cases. Top-coded questions include age, income, and housing unit value. Age perturbation disguises original data by randomly adjusting the reported ages for a subset of individuals. Weight perturbation disguises the probability of selection for some records. Users should exercise caution when forming estimates near top-coded or bottom-coded values. More information on the variables that receive top or bottom coding in the 2020 PUMS can be found at: https://www.census.gov/programs-surveys/acs/data/experimental-data.html.

SAMPLE DESIGN

The 2020 PUMS was designed to include one percent of the housing units and one percent of the non-imputed GQ persons in the United States and Puerto Rico. The PUMS sample was selected from the full sample ACS records separately for Housing Units (HUs) and GQ persons. The PUMS sample sizes were based on the Population Estimates Program estimates for housing units and GQ persons.

Note that for the 2020 PUMS 1-year data not all states were able to meet the target of a 1-percent sample due to the pandemic's impact on the ability to collect ACS responses.

The PUMS sample of persons in households was selected by keeping all persons in selected PUMS HUs. The systematic sampling method used sampling intervals chosen to yield roughly a one percent weighted housing unit sample and one percent group quarter person sample. The ACS estimates for Housing Units may be found in table XK202502 (Housing Units). The ACS estimate for Group Quarters may be found in table XK202601 (Group Quarter Population). Data for both tables may be found using https://www.census.gov/programs-surveys/acs/data/experimental-data.html.

Note: The 2020 PUMS 1-year data will not be published on data.census.gov or via the online tool found at https://data.census.gov/mdat/.

The GQ population sample has been supplemented by a large-scale whole person imputation for not-in-sample GQ facilities. The goal of the imputation process was to establish representation of the major GQ type groups within county and tract to agree better with the ACS GQ sample frame. The interviewed GQ person records were selected at random to become donor records which were imputed into the selected not-in-sample GQs. The imputed records were given new values for the geography and GQ type fields.

For details on the ACS GQ estimation methodology, see the 2020 ACS 1-year Accuracy of the Data document located at: https://www.census.gov/programs-surveys/acs/data/experimental-data.html.

Note that the PUMS carries PUMA and state codes, but does not carry variables that identify the major GQ type groups or the county and tract information of the imputed records. By including these records, the PUMS will agree better with the full sample ACS for population totals by state and PUMA.

Housing Units

The sampling for PUMS HUs was performed on the ACS sample of HUs as follows:

1. Records of ACS HUs were sorted within each state by: PUMA, ACS weighting area, interview mode, type of vacant, tenure, building type, household type, householder demographics (race, Hispanic origin, sex and age), county, tract, and housing unit weight.

2. Systematic sampling was applied to ACS HUs:

    a. Within each state, a random number is chosen between zero and the sampling interval. A counter is initialized with the random number.

    b. At each HU record, the value of the counter is incremented by one and compared to the sampling interval.

        i. If the counter's new value is greater than the sampling interval, the HU record is selected for the PUMS and a flag is set to 1. The counter is

decreased by the sampling interval with the new value passed to the next record.

ii. If the counter is less than the sampling interval, the HU record is not selected for the PUMS and the value of the counter is passed to the next record without altering its value.

3. All HUs selected for PUMS were placed in the PUMS HU sample file.

The PUMS HU sample file was matched to the ACS sample of persons. All persons in selected HUs were placed in the PUMS person sample.

Group Quarters

The sampling for PUMS GQ persons was performed on the ACS sample of GQ persons as follows:

1. Interviewed GQ persons were sorted within each state by the size of their GQ facility (large vs. small), the type of GQ facility, PUMA, demographics (race, Hispanic origin, sex and age), county, tract, and GQ person weight.

2. Systematic sampling was applied in the same manner as described above for HUs.

3. All selected GQ interviewed persons were added to the PUMS person sample. All imputed records derived from the selected PUMS interviews were also kept in the PUMS person sample. A placeholder record was also placed in the PUMS HU file for each PUMS GQ person record.

WEIGHTING

Data Year 2020 presented unique challenges. Due to the difficulties of data collection, the sample obtained does not adequately reflect a representative sample of the full population. In order to address this, the PUMS weighting incorporated an alternative weighting methodology in addition to some of our standard ratio estimation procedures. An explanation of this methodology may be found here: https://www.census.gov/programs-surveys/acs/data/experimental-data.html.

In addition, because the weighting was designed primarily to produce estimates for states and large counties, estimates for PUMAs which can combine or split counties should be used with caution as the experimental weights are not optimized to produce estimates for these areas.

Weights for PUMS person records are a product of the final full ACS weight, the PUMS subsampling factor, and ratio-estimate factors. The PUMS subsampling factors are the sampling intervals used to sample the PUMS HU or GQ person records within a state. The ratio-estimate factors bring the PUMS estimates into closer agreement with the published ACS estimates for several characteristics explained below.

Group Quarters Person Weighting

The group quarters (GQ) person weighting for the PUMS 2020 1-year estimates was similar to the previous years' PUMS weighting in that it included both the sampled interviews and the imputed records described in the section on Sample Design. However, imputed records were treated the same as PUMS sample interviews in the weighting.

The procedure used to assign the weights to the GQ persons is performed independently within each state. The steps are as follows:

Initial Weight for GQ Persons

The PUMS initial weight is the product of the ACS unrounded weights for the record and the PUMS subsampling factor. Each imputed record received the same subsampling factor as its donor interview.

GQ Person Weighting Factors

GQ Person Post-stratification Factor

This factor adjusts the GQ person weights so that the weighted sample counts equal ACS published estimates at the state level. Due to the ACS GQ sample design and noise added for disclosure avoidance reasons, only state level PUMS GQ person estimates will agree closely with published ACS 2020 estimates. This adjustment uses the following groups:

State × Institutional/noninstitutional × Sex × Age Category

Rounding of GQ Person Weights

The final GQ person weight is rounded to an integer. Rounding is performed so that the sum of the rounded weights is within one person of the sum of the ACS total GQ person estimate for the state.

Housing Unit and Household Person Weighting

The housing unit and household person weighting used the previously referenced alternative methodology.

Initial Weight for Persons and HUs

The PUMS initial weight is the product of the ACS alternative weight for the record and the PUMS subsampling factor.

Rounding of Housing Unit Weights

The Initial weight is rounded to an integer. Rounding is performed so that the sum of the rounded weights is within one housing unit of the sum of the ACS total HU's estimates within state.

Information on the methodology used to create the experimental weights may be found in the Technical Working paper titled "Addressing Nonresponse Bias in the American Community Survey During the Pandemic Using Administrative Data" which may be found on the ACS Experimental Data Webpage (located at https://www.census.gov/programs-surveys/acs/data/experimental-data.html).

ESTIMATION

To produce estimates or tabulations of characteristics from the PUMS, add the weights of all persons or HUs that possess the characteristic of interest.[2] For instance, if the characteristic of interest is "total number of black teachers", simply determine the race and occupation of all persons and cumulate the weights of those who match the characteristics of interest. To obtain estimates of proportions, divide the weighted estimate of persons or HUs with a given characteristic by the weighted estimate of the denominator. For example, the proportion of "black teachers" is obtained by dividing the weighted estimate of black teachers by the weighted estimate of teachers.

Due to the variance properties of the experimental estimation methodology, the variance estimates for some PUMS estimates may be smaller than expected when compared to the equivalent variance estimates from previous years.

PUMS estimates are expected to be different from published ACS estimates that are based on the full set of data because of the additional sampling. In general, the exception will be characteristics controlled by the ratio-estimate factors at the PUMA level for HUs and persons in HUs and at the state level for GQ persons.

Note that the housing unit file contains some records with zero weights. These are the GQ placeholder records[3]. The housing unit weights were set to zero for these records since they are not housing units, but persons. For confidentiality reasons, the GQ data are not provided at the level of an address but only at the person-level. All of the GQ person data are included in the PUMS person file except for the variable FS ("Yearly food stamp/Supplemental Nutrition Assistance Program (SNAP) recipiency"), which is included on the GQ placeholder records in the housing unit file. For food stamp recipiency estimates of persons in GQs, you will need to match the placeholder records to the person file to obtain the person weights.

[2] Users should exercise caution when forming estimates near top-coded or bottom-coded values. More information on the variables that receive top or bottom coding in the PUMS can be found at: https://www.census.gov/programs-surveys/acs/data/experimental-data.html. For previous years' top- and bottom-coded values, see https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html.

[3] To identify HU and GQ placeholder records on the PUMS housing file, see RELSHIPP and TYPEHUGQ variables in the PUMS data dictionary: https://www.census.gov/programs-surveys/acs/data/experimental-data.html. For previous years' Data Dictionaries, see https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html.

ERRORS IN THE DATA

Every sample survey is subject to two types of error: sampling error and nonsampling error.

Sampling Error

The data in the ACS products are estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology. The estimates from the chosen sample also differ from other samples of HUs and persons within those HUs. Sampling error in data arises due to the use of probability sampling, which is necessary to ensure the integrity and representativeness of sample survey results. The implementation of statistical sampling procedures provides the basis for the statistical analysis of sample data.

Estimates made with PUMS data are subject to additional sampling error because the PUMS data consists of a subset of the full ACS sample. Thus, standard errors and margins of error of PUMS estimates can be larger than the standard errors or margins of error that would be obtained using the full ACS microdata.

Nonsampling Error

In addition to sampling error, data users should realize that other types of errors might be introduced during any of the various complex operations used to collect and process survey data. For example, operations such as data entry from questionnaires and editing may introduce error into the estimates. These and other sources of error contribute to the nonsampling error component of the total error of survey estimates.

Nonsampling errors may affect the data in two ways. Errors that are introduced randomly increase the variability of the data. Systematic errors, which are consistent in one direction, introduce bias into the results of a sample survey. The Census Bureau protects against the effect of systematic errors on survey estimates by conducting extensive research and evaluation programs on sampling techniques, questionnaire design, and data collection and processing procedures. In addition, an important goal of the ACS is to minimize the amount of nonsampling error introduced through nonresponse for sample HUs. One way of accomplishing this is by following up on mail nonrespondents during the CAPI phase.

The 2020 ACS 1-year data had evidence of non-response bias due to the pandemic. To learn more about data collection disruptions, the modifications to standard weighting and estimation to combat the collection issues, and the resulting data quality issues that informed the decision to not release the standard 1-year ACS data products, please see the report titled "An Assessment of the COVID-19 Pandemic's Impact on the 2020 ACS 1-Year Data", which may be found at https://www.census.gov/programs-surveys/acs/data/experimental-data.html.

More information about the control of nonsampling error can be found in the ACS Accuracy of the Data at: https://www.census.gov/programs-surveys/acs/data/experimental-data.html.

Previous years' ACS Accuracy documents may be found at: https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html.

MEASURING SAMPLING ERROR

Standard Error

A measure of the deviation of a sample estimate from the average of all possible samples. Sampling error and some types of nonsampling error, such as undercoverage and item nonresponse, are estimated by the standard error. The sample estimate and its estimated standard error permit the construction of interval estimates with a prescribed confidence that the interval includes the average result of all possible samples.

Due to the use of experimental weights, the variance for some PUMS estimates may be smaller than expected when compared to equivalent variances from previous years' estimates.

Usually, two methods are provided for estimating the standard errors of PUMS estimates: a successive difference replicate (SDR) method using replicate weights and a generalized variance function (GVF) method using design factors. Due to the unique circumstances of the 2020 production year, the 2020 PUMS 1-year design factors are not published. The SDR method should be used for estimating standard errors.

Confidence Intervals

A sample estimate and its estimated standard error may be used to construct confidence intervals about the estimate. These intervals are ranges that will contain the average value of the estimated characteristic that results over all possible samples, with a known probability.

For example, if all possible samples that could result under the PUMS sample design were independently selected and surveyed under the same conditions, and if the estimate and its estimated standard error were calculated for each of these samples, then:

1. Approximately 68 percent of the intervals from one estimated standard error below the estimate to one estimated standard error above the estimate would contain the average result from all possible samples.

2. Approximately 90 percent of the intervals from 1.645 times the estimated standard error below the estimate to 1.645 times the estimated standard error above the estimate would contain the average result from all possible samples.

3. Approximately 95 percent of the intervals from 1.96 times the estimated standard errors below the estimate to 1.96 times the estimated standard errors above the estimate would contain the average result from all possible samples.

These intervals are referred to as 68 percent, 90 percent, and 95 percent confidence intervals, respectively. An example of how to construct a 90 percent confidence interval follows:

Add and subtract 1.645 times the standard error (SE) of the estimate to yield the lower and upper bounds of a 90% confidence interval around the estimate.

LB = Lower bound = Estimate - 1.645 x SE(Estimate)

UB = Upper bound = Estimate + 1.645 x SE(Estimate)

The 90% confidence interval is the interval (LB, UB).

Limitations

The user should be careful when computing and interpreting standard errors and confidence intervals.

Nonsampling Error

The estimated standard errors included in this data product do not include all portions of the variability due to nonsampling error that may be present in the data. In particular, the standard errors do not reflect the effect of correlated errors introduced by interviewers, coders, or other field or processing personnel. Nor do they reflect the error from imputed values due to missing responses. Thus, the standard errors calculated represent a lower bound of the total error. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence (i.e., 68, 90, or 95 percent). Thus, some care must be exercised in the interpretation of the data in this data product based on the estimated standard errors.

Very Small (Zero) or Very Large Estimates

The value of almost all PUMS characteristics is greater than or equal to zero by definition. For zero or small estimates, use of the method given previously for calculating confidence intervals relies on large sample theory, and may result in negative values which for most characteristics are not admissible. In this case the lower limit of the confidence interval is set to zero by default. A similar caution holds for estimates of totals close to a control total and estimated proportions near one, where the upper limit of the confidence interval is set to its largest admissible value. In these situations, the level of confidence of the adjusted range of values is less than the prescribed confidence level.

Calculating Standard Errors with Replicate Weights

The standard error may be calculated using the successive difference replicate (SDR) method using the replicate weights provided in the PUMS file. The advantage of using the SDR method is that a single formula is used to calculate the standard error of many types of estimates. Generally, using the SDR method will produce a more accurate estimate than the GVF method.

Each PUMS housing unit and person record contains 80 PUMS replicate weights. These replicate weights are based on the ACS replicate weights adjusted for PUMS subsampling. For any estimate X, 80 replicate estimates are also computed using the replicate weights. For this discussion, we refer to X as the 'full sample estimate.' The first replicate estimate, X_1, is computed using the first replicate weight, the second replicate estimate, X_2, is computed using the second replicate weight, and so on. Each replicate estimate is computed using the replicate weights in the same way that the full sample estimate X is calculated.

NOTE: When programming the replicate weight standard errors, users will find the 80 replicate weights can be positive, zero or negative. The negative replicate weights are due to the addition of the Group Quarters (GQ) population to the full ACS weighting process. Within a weighting cell, GQ estimates were subtracted from population totals, sometimes resulting in negative values for the cell. The cells were collapsed in such a way as to prevent a final cell from being zero or negative for the full sample weights. The full sample weights always have a value of at least one. This restriction was not placed on the replicate weights since their only purpose is to represent the variability of the sample. PUMS replicate weights are based on ACS replicate weights so negative values may occur. Keep in mind that the replicate weights are only to be used to estimate standard errors.

The standard error of X can be calculated after the replicate estimates X_1 through X_{80} are computed. The standard error is estimated using the sum of squared differences between each replicate estimate X_r and the full sample estimate X. The standard error formula is:

$$SE(X) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(X_r - X)^2}$$

Data users who wish to see worked examples may consult the documentation for the ACS Variance Replicate Estimate (VRE) tables, located here: https://www.census.gov/programs-surveys/acs/technical-documentation/variance-tables.html. Additional resources for using the SDR method are listed in the PUMS ReadMe document.

As we mentioned earlier, the standard error can be used to form a 90% confidence interval around the estimate (X) as follows:

$$LB = \text{Lower bound} = X - 1.645 \times SE(X)$$

$$UB = \text{Upper bound} = X + 1.645 \times SE(X)$$

The 90% confidence interval is the interval (LB, UB).

Examples of PUMS estimates with their SE and MOE may be found by clicking on PUMS Estimates for User Verification at: https://www.census.gov/programs-surveys/acs/data/experimental-data.html. For previous years' PUMS, see: https://www.census.gov/programs-surveys/acs/microdata/documentation.html

Users can check national and state level estimates with associated standard errors and margin of errors by comparing to values shown in these files. The SE and MOE are calculated using the SDR method with the PUMS replicate weights.

Note on Design Factors for the 2020 PUMS 1-year files

Due to the unique circumstances of the 2020 production year, the 2020 PUMS 1-year design factors will not be published. Data users should use the SDR method for calculating standard errors and margins of error.

WORKING WITH DOLLAR AMOUNTS

Dollar variables must be adjusted by the inflation adjustment factors supplied on the PUMS files before they are used to form estimates. Also, when comparing the 2020 PUMS data to other PUMS years, the dollars must be converted into a common year.

Adjustment Factors on the PUMS File

The PUMS data dictionary includes two adjustment factors for dollar values:

ADJINC – inflation adjustment factors for income variables, such as household income, self-employment income, retirement income and wages.

ADJHSG – inflation adjustment factor for most housing dollar variables, such as utility costs, rent, food stamps, and condominium fees.

The adjustment factor for income and earnings dollar amounts (ADJINC) is applied to income and earning variables. The factor is necessary because the ACS collects data on the past twelve months of income in each of the twelve months of the year. Responses therefore often include amounts from both the current year and the previous year. The adjustment factor will convert amounts into consistent 2020 dollars.

The adjustment factor for housing dollar amounts (ADJHSG) is applied to variables related to housing costs. See the PUMS ReadMe document and PUMS Data Dictionary for more information. Both are located at https://www.census.gov/programs-surveys/acs/data/experimental-data.html. Previous years' documents may be found at: https://www.census.gov/programs-surveys/acs/microdata/documentation.html.

Dollars from Different Years

When working with dollar amounts from different PUMS years, it is necessary to convert the amounts into dollars from a common year (after applying the adjustment factors described in the previous paragraph). We use the CPI-U-RS adjustment factors from the Bureau of Labor Statistics. These factors can be found in "All Items" tab under the column labelled "AVG" of the table "Updated CPI-U-RS, All items, 1977-2020" located at:

https://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm. For example, to express year 2000 dollars in terms of 2020 dollars, multiply the 2000 dollars by 381.2 / 252.9 = 1.507.

REFERENCES

[1]. ACS Accuracy of the Data (2020): https://www.census.gov/programs-surveys/acs/data/experimental-data.html

[2]. Design and Methodology of the American Community Survey: April 2014: https://census.gov/programs-surveys/acs/methodology/design-and-methodology.html

[3]. Updated CPI-U-RS, All Items, 1977-2020: https://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm