Public Use Microdata Sample (PUMS)

Accuracy of the Data (2017)

INTRODUCTION

The Public Use Microdata Sample (PUMS) are a subset of the 2017 American Community Survey (ACS) and Puerto Rico Community Survey (PRCS) samples. Unless otherwise specified, the term "ACS" in this document will refer to both the ACS and PRCS.

The 2017 PUMS contains a sample of the group quarters (GQ) population. Group quarter data were included in PUMS beginning in 2006. The 2017 PUMS GQ sample includes imputed records. Imputed GQ records were first included in PUMS in 2011. The 2017 ACS selected sample from which PUMS is drawn covers all counties across the nation and all municipios in Puerto Rico.

Estimates created from the 2017 PUMS data are expected to be different from the 2017 published 2017 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

INTRODUCTION 1
CONFIDENTIALITY OF THE DATA 3
Title 13, United States Code 3
Disclosure Avoidance 3
Data Swapping 3
Synthetic Data 4
PUMAs 4
Additional Measures 4
SAMPLE DESIGN 4
Housing Units 5
Group Quarters 6
WEIGHTING 6
Group Quarters Person Weighting 6
Housing Unit and Household Person Weighting 7
ESTIMATION 9
ERRORS IN THE DATA 10
Sampling Error 10
Nonsampling Error 10
MEASURING SAMPLING ERROR 11
Standard Error 11
Confidence Intervals 11
Limitations 12
Approximating Standard Errors with Replicate Weights 13
Approximating Generalized Standard Errors with Design Factors 14
Examples of Standard Error Calculations using Generalized Standard Error Formulas 19
WORKING WITH DOLLAR AMOUNTS 22
Adjustment Factors on the PUMS File 22
Dollars from Different Years 22
REFERENCES 23

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 in tables of frequency data (the number or percent of the population with certain characteristics). 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. Because of data swapping, users should not assume that tables with cells having a value of one or two reveal information about specific individuals. Data swapping procedures were first used in the 1990 Census, and were used again in Census 2000.

[1] The Census Bureau's Disclosure Review Board approved the 2017 PUMS 1-year data for release with DRB Clearance number CBDRB-FY18-506.

Synthetic Data

The goals of using synthetic data are the same as the goals of data swapping, namely to protect the confidentiality in tables of frequency data. Persons are identified as being at risk for disclosure based on certain 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 within a state based on data and location collected in the 2010 Census and have been used by the ACS PUMS files since the 2012 data year. 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://www.census.gov/programs-surveys/acs/geography-acs/areas-published/other-mapping-resources.html.

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 2017 PUMS can be found at: https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html.

SAMPLE DESIGN

The 2017 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.

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 B25001 (Housing Units) at: https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_1YR/B25001/0100000US.04000. The ACS estimate for Group Quarters may be found in table B26001 (Group Quarter Population) at: https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_1YR/B26001/0100000US.04000.

The sampling interval for each state and HU/GQ sample is the ratio of the number of interviewed records available for sampling and the required sample size (sampling intervals are not rounded to integers).

The 2017 PUMS GQ sample was similar to previous years' PUMS sample. The GQ population sample has been supplemented by a large-scale whole person imputation into 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 2017 ACS Accuracy of the Data at: https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html

Since the PUMS intended sample size of the original GQ interviews was set at one percent and the number of imputed records was similar to the number of interviews, the PUMS total sample size is about two percent. 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

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 2017 1-year estimates was similar to the previous year's 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 2017 estimates. This adjustment uses the following groups:

State × Institutional/noninstitutional × Sex × Age Category

Rounding for 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 estimation procedure used to assign the HU and person weights is performed independently within each **PUMA**.

Initial Weight for Persons and HUs

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

Person Weighting Factors

The person weights are adjusted to agree better with ACS published estimates for householders, spouses, race, Hispanic origin, sex and age by a series of two steps, which are repeated until a stopping criterion is met. This is an iterative proportional fitting, or raking, process. The person weights are individually adjusted at each step as described below.

The two steps are as follows:

Spouse Equalization/Householder Equalization Raking Factor

This factor is applied to individuals based on the combination of their status of being in a married-couple or unmarried-partner household and whether they are the householder. All persons are assigned to one of four groups:

1. Householder in a married-couple or unmarried-partner household
2. Spouse or unmarried partner in a married-couple or unmarried-partner
3. household (non-householder)
   Other householder
4. Other non-householder

The weights of persons in the first two groups are adjusted so that their sums are each equal to the ACS estimate of married-couple or unmarried-partner households using the ACS housing unit weight. The weights of persons in the third group are adjusted so that the sum is equal to the ACS estimate of occupied housing units not having a partner using the housing unit weight. The weights of persons in the fourth group are adjusted to agree with the ACS total population minus the first three groups. The goal of this step is to produce more consistent estimates of spouses or unmarried partners and married-couple and unmarried-partner households while simultaneously producing more consistent estimates of householders, occupied housing units, and households.

Demographic Raking Factor

This factor is applied to individuals based on their age, race, sex, and Hispanic origin. It adjusts the person weights so that the weighted sample counts equal ACS population estimates by age, race, sex, and Hispanic origin at the PUMA level. Because of collapsing of groups in applying this factor, only total population is assured of agreeing precisely with the published ACS 2017 population estimates at the PUMA level.

This uses the following groups within each PUMA (note that there are 13 Age groupings):

Race / Ethnicity (non-Hispanic White, non-Hispanic Black, non-Hispanic American Indian or Alaskan Native, non-Hispanic Asian, non-Hispanic Native Hawaiian or Pacific Islander, and Hispanic (any race)) × Sex × Age Groups.

These two steps are repeated several times until the estimates at the PUMA level achieve their optimal consistency with regard to the spouse and householder equalization. The final Person Weighting Factor is then equal to the product of the factors from all of the iterations of these two adjustments. The unrounded person weight is then equal to the product of Person Weighting Factor times the initial person weight.

Rounding of Person Weights

The person weight after the Person Weighting Factor has been applied 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 persons from HU's estimates within state and PUMA.

Householder Adjustment Factor (HHRF)

This factor, applied to occupied housing units, is the same as the Person Weighting factor from the person weighting. After this stage the weight of the housing unit is identical to the unrounded person weight of the householder after the Person Weighting Factor is applied.

Housing Unit Control Factor

This factor adjusts PUMS housing unit estimates to agree with the published ACS housing unit estimates for housing units with married couples (or partners), occupied housing units without partners and vacant housing units.

Rounding of Housing Unit Weights

The Housing Unit weight after the Housing Unit Control Factor is applied 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 and PUMA.

For a detailed description of how the original ACS weights are computed see the 2017 ACS Accuracy of the Data at: https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.2017.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 get estimates of proportions, divide the weighted estimate of persons or HUs with a given characteristic by the weighted estimate of the base. For example, the proportion of "black teachers" is obtained by dividing the weighted estimate of black teachers by the weighted estimate of teachers.

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. 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.

[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/technical-documentation/pums/documentation.html.

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 the food stamp recipiency variable, 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 get the person weights.

A note to GQ data users. There are limitations to the usefulness of GQ estimates at the PUMA level. The PUMS weighting controls the GQ estimates to agree with the ACS state level estimates. Depending on the application or analysis, GQ data users should consider working with state level estimates rather than PUMAs.

As was done since the 2012 1-year PUMS file, the complete plumbing facilities recode (PLM) was assigned the value of '9' in all PUMAs in Puerto Rico to mean not applicable.

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 of PUMS estimates can be larger than standard errors that would be obtained using all of the ACS data.

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,

[3] To identify HU and GQ placeholder records on the PUMS housing file, see the TYPE variables in the PUMS data dictionary https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html.

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 CATI and CAPI phases.

More information about the control of nonsampling error can be found in the 2017 ACS Accuracy of the Data at: https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.2017.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.

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. Replicate weights have been provided with the ACS PUMS files since the 2005 PUMS. Design factors method is used by the Census PUMS and also in use by the ACS PUMS since 2000. **It is important to keep in mind that there will be differences between the standard error approximations computed by the two methods.** Generally, using the SDR method will produce a more accurate estimate of a standard error.

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 two estimated standard errors below the estimate to two 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 of the estimate to yield the lower and upper bounds of a 90% confidence interval around the estimate (EST).

LB=Lower bound = EST - 1.645*SE(EST)

UB=Upper bound = EST + 1.645*SE(EST)

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.

Approximating Standard Errors with Replicate Weights

The standard error may be calculated using the SDR method using the replicate weights provided in the PUMS file. SDR standard errors will often be more accurate than GVF standard errors, although they may be more inconvenient for some users to calculate. The advantage of using the SDR method is that a single formula is used to calculate the standard error of many types of estimates.

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 computed, as described in the Estimation section.

NOTE: When programming the replicate weight standard errors, users will find the eighty 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 are always one or more. 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 with the formula provided in the PUMS accuracy document.

The standard error of X can be approximated 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:

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

If X is zero, then use the generalized variance method for zero estimates given in Standard Errors for Totals and Percentages, to approximate the standard error. Data users who wish to see worked examples may consult the documentation for the ACS Variance Replicate Tables, located here: https://www.census.gov/programs-surveys/acs/technical-documentation/variance-tables.html.

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

LB=Lower bound = X - 1.645*SE(X)

UB=Upper bound = X + 1.645*SE(X)

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

As mentioned previously, we consider the SDR SEs using the replicate weight to be more accurate than the GVF SEs using the design factor. For exceptions, please note the following:

When using SDR SEs, occasionally the SE is zero for an estimate. Except for controlled estimates, all PUMS estimates are based on a sample of the population and should not have a SE of zero. If the SDR method using the replicate weight gives a SE of zero for an estimate that is not controlled, use the GVF method with the design factors to obtain a SE for that estimate. If the estimate is controlled, then a SE of zero is appropriate. Data users may check it an estimate is controlled by finding the equivalent ACS estimate on American FactFinder. If the ACS estimate is controlled it will have a MOE of five asterisks (*****).

If your estimate is a median, the SDR method using replicate weights may yield a SE of zero. Medians should always have a non-zero SE. A median with a SE of zero may occur when several records in the middle of the distribution were rounded to the same value, or when the characteristic contains few records, such as a median based on less than five records. Rounding by respondents, as well as rounding by PUMS edits may mask the variability in the median. In order to yield a more adequate standard error for that case, use the GVF method with the design factors to estimate the SE of a median.

Examples of PUMS estimates with their SE and MOE may be found by clicking on PUMS Estimates for User Verification at:
https://census.gov/programs-surveys/acs/technical-documentation/pums/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.

Approximating Generalized Standard Errors with Design Factors

Totals and Percentages

The design factors provided in Attachment A can be used to approximate the standard errors of most sample estimates of *totals* and *percentages*. Design factors are given by subject for the United States, all 50 states, the District of Columbia, and Puerto Rico. The term "subject" refers to a characteristic, such as age for persons and tenure for HUs. The design factors reflect the effects of the actual sample design and estimation procedures used for the ACS. To approximate the standard error use the following formulas:

Total Formula

$$SE(\hat{Y}) = DF \times \sqrt{99 \times \hat{Y}\left(1 - \frac{\hat{Y}}{N}\right)}$$

Where:
DF = Design Factor
N = Size of Geographic Area
$\hat{Y}$ = Estimate of Characteristic Total

Percent Formula

$$SE(\hat{p}) = DF \times \sqrt{\frac{99}{B} \times \hat{p}(100 - \hat{p})}$$

Where:
DF = Design Factor
B = Base of Estimated Percentage
$\hat{p}$ = Estimated Percentage

The values of N and the design factor can be determined as follows:

1. For the value of N, obtain the number of persons, number of households or number of HUs respectively for the geography (geographies) you are interested in. If the estimate is of HUs then use the number of HUs; if the estimate is of families or households then use the number of households; otherwise use the number of persons.

2. Select the appropriate table from Attachment A. Use the table for the United States (Attachment A-1) when estimating characteristics for the United States or geographic areas that cover more than one state. Use the table for a specific state when estimating characteristics for that state or geographic areas that are contained entirely within that state.

3. Then use the selected table to obtain the appropriate design factor for the characteristic: for example, educational attainment or ancestry. If the estimate is a combination of two or more characteristics, we suggest the following guideline: Use the largest design factor for this combination of characteristics. **The only exception to this is for items crossed with race or Hispanic Origin. For an item(s) crossed with race or Hispanic Origin use the largest design factor not including the race or Hispanic Origin design factor**.

An inspection of the formulas used to calculate standard errors under simple random sampling suggests that when dealing with zero estimates or very small estimates of totals and percentages the standard error estimates approach zero. This is also the case for very large estimates of totals and percentages. Zero or small estimates, like any other sample estimates,

are still subject to sampling variability and therefore an estimated standard error of zero or close to zero is not adequate. Use one of the following procedures for estimates of this type:

1. **An estimated total is less than 425 or within 425 of the total size of the tabulation area**. Use a basic standard error of 246 multiplied by the design factor for the type of estimate.

2. **For estimated percentages that are less than 2 or greater than 98.** Use a value of 2 for the estimated percentage in the percent formula.

3. **The denominator of a percentage is zero.** There are no sample observations available to compute an estimate of a proportion or an estimate of its standard error.

Sums and Differences

For the sum or difference between two estimates, the standard error is approximately the square root of the sum of the two individual standard errors squared:

$$\mathrm{SE}(\hat{X} + \hat{Y}) = \mathrm{SE}(\hat{X} - \hat{Y}) = \sqrt{[\mathrm{SE}(\hat{X})]^2 + [\mathrm{SE}(\hat{Y})]^2}$$

This method is, however, an approximation as the two estimates of interest in a sum or a difference are likely to be correlated. If the two quantities X and Y are positively correlated, this method underestimates the standard error of the sum of X and Y and overestimates the standard error of the difference between the two estimates. If the two estimates are negatively correlated, this method overestimates the standard error of the sum and underestimates the standard error of the difference.

Ratios

Frequently, the statistic of interest is the ratio of two variables, where the numerator is not a subset of the denominator. An example is the ratio of students to teachers in public elementary schools. The standard error of the ratio between two sample estimates is approximated as follows:

$$\mathrm{SE}\left(\frac{\hat{X}}{\hat{Y}}\right) = \left(\frac{\hat{X}}{\hat{Y}}\right) \times \sqrt{\frac{[\mathrm{SE}(\hat{X})]^2}{\hat{X}} + \frac{[\mathrm{SE}(\hat{Y})]^2}{\hat{Y}}}$$

If the ratio is a proportion, that is, the numerator is a subset of the denominator, then it should be transformed into a percentage and the procedure outlined in Standard Errors for Totals and Percentages followed.

Medians

The sampling variability of an estimated median depends on the form of the distribution and the size of its base. The standard error of an estimated median is approximated by

constructing a 68 percent confidence interval. Estimate the 68 percent confidence limits of a median based on sample data using the following procedure.[4]

1. Obtain the weighted frequency distribution for the selected variable. Cumulate these frequencies to yield the base.

2. Approximate the standard error of a 50 percent proportion using the formula in Standard Errors for Totals and Percentages.

$$SE(50\,percent) = DF \times \sqrt{\frac{99}{B} \times 50^2}$$

Subtract from and add to 50 percent the standard error determined in step 2.

p_lower = 50 – SE(50 percent)

p_upper = 50 + SE(50 percent)

3. Determine the categories in the distribution that contain p_lower and p_upper. If p_lower and p_upper fall in the same category, follow step 5. If p_lower and p_upper fall in different categories, go to step 6.

4. If p_lower and p_upper fall in the same category, do the following:

   a. Define A1 as the smallest value in that category.
   b. Define A2 as the smallest value in the next (higher) category.
   c. Define C1 as the cumulative percent of units strictly less than A1.
   d. Define C2 as the cumulative percent of units strictly less than A2.

   Use the following formulas to approximate the lower and upper bounds for a confidence interval about the median:

$$\text{Lower Bound} = \left[\frac{\text{p}_{lower} - \text{C1}}{\text{C2} - \text{C1}}\right] \times (\text{A2} - \text{A1}) + \text{A1}$$

$$\text{Upper Bound} = \left[\frac{\text{p}_{upper} - \text{C1}}{\text{C2} - \text{C1}}\right] \times (\text{A2} - \text{A1}) + \text{A1}$$

[4]The design factor method shown here for medians is preferred over the replicate weight method whenever the replicate weight method gives a standard error of zero. This may happen due to having several records in the middle of the range that have exactly the same value. Be aware that PUMS dollar values are rounded to the nearest 100 for values between 1,000 and 50,000 and rounded to the nearest 1,000 above 50,000. This increases the numbers of respondents with exactly the same value. The amount of rounding done by respondents is unknown, but could be substantial. Since rounding may cause the number of records with exactly the same value to increase, and might cause all 80 replicates to yield the same median, the replicate weight formula can give a standard error of zero. To avoid this, it is possible to calculate the medians using a categorical method with linear interpolation for all 80 replicates, OR simply use the design factor method to estimate the standard errors.

5. If p_lower and p_upper fall in different categories, do the following:

    a. For the category containing p_lower: Define A1, A2, C1, and C2 as described in step 5. Use these values and the formula in step 5 to obtain the lower bound.

    b. For the category containing p_upper: Define new values for A1, A2, C1, and C2 as described in step 5. Use these values and the formula in step 5 to obtain the upper bound.

    Use the lower and upper bounds approximated in steps 5 or 6 to approximate the standard error of the median.

$$\text{SE(median)} = \frac{1}{2} \times \text{(Upper Bound-Lower Bound)}$$

Means

A mean is defined here as the average quantity of some characteristic (other than the number of people, HUs, households, or families) per person, housing unit, household, or family. For example, a mean could be the average annual income of females age 25 to 34. The standard error of a mean can be approximated by the formula below. Because of the approximation used in developing this formula, the estimated standard error of the mean obtained from this formula will generally underestimate the true standard error.

$$SE(\bar{Y}) = DF \times \sqrt{\frac{99}{B} \times s^2}$$

Where: B is the base (denominator) of the mean, and s^2 is the sample variance of the characteristic based on weighted data. The value of s^2 can be computed using the formula:

$$s^2 = \frac{\sum_{i}^{n} = 1 \, w_i {y_i}^2 - [(\sum_{i=1}^{n} w_i y_i)^2 / \sum_{i=1}^{n} w_i]}{(\sum_{i=1}^{n} w_i) - 1}$$

Where:

w_i is the weight of the i^{th} sample record
y_i is the value of the characteristic for the i^{th} sample record
n is the number of sample records

Note that $\sum_{i=1}^{n} w_i$ is the weighted estimate of persons/HUs in the sample (ex. the number of females age 25 to 34, and $\sum_{i=1}^{n} w_i y_i$ is the weighted aggregate estimate for the characteristic of interest (ex. the aggregate income of females age 25 to 34).

Examples of Standard Error Calculations using Generalized Standard Error Formulas

Note that the examples presented below use 2015 PUMS data to demonstrate the use of the generalized standard error formulas.

Example 1 – Using Design Factors to Estimate a Total

Suppose the estimated number of people 15 years or over who were never married is 2,219,061 from the 2015 PUMS data for the state of Virginia. To calculate the standard error, we use the total formula for Totals and Percentages. In this formula, Y is our estimate of 2,219,061 and N is the total 2015 PUMS population for the state of Virginia, which is 8,382,993. The design factor (from Attachment A-48) for "Marital Status" is 1.3.

$$\text{SE} = 1.3 \times \sqrt{99 \times 2{,}219{,}061 \times \left(1 - \frac{2{,}219{,}061}{8{,}382{,}993}\right)} = 16{,}522.47$$

To calculate the lower and upper bounds of the 90 percent confidence interval around 2,219,061 using the standard error, simply multiply 16,522.47 by 1.645, then add and subtract the product from 2,219,061. Thus the rounded 90 percent confidence interval for this estimate is [2,219,061 - 1.645(16,522.47)] to [2,219,061 + 1.645(16,522.47)] or 2,191,882 to 2,246,240.

Example 2 – Using Design Factors to Estimate a Proportion or Percentage

Suppose the estimated percent of people 25 years or over with a bachelor's degree or higher in Louisiana is 23.2033 from the PUMS data and the base of the estimated percentage is 3,096,763. To calculate the standard error, we use the percent formula for Totals and Percentages. The design factor (from Attachment A-20) for "Educational Attainment" is 1.4.

$$\text{SE} = 1.4 \times \sqrt{\frac{99}{3{,}096{,}763} \times 23.2033 \times (100 - 23.2033)} = 0.3341$$

To calculate the lower and upper bounds of the 90 percent confidence interval around 23.2033 percent using the standard error, simply multiply 0.3341 by 1.645, then add and subtract the product from 23.2033. Thus the 90 percent confidence interval for this estimated percentage is [23.2033 - 1.645(0.3341)] to [23.2033 + 1.645(0.3341)] or 22.6537 to 23.7529.

Example 3 – Calculating the Standard Error of a Median

Suppose the table below shows a weighted frequency distribution for adjusted household income in Massachusetts.

Adjusted Household Income	Frequency	Cumulative Frequency	Cumulative Percent
Less than $10,000	151,095	151,095	5.90%
$10,000 to $14,999	121,879	272,974	10.66%
$15,000 to $19,999	104,443	377,417	14.74%
$20,000 to $24,999	96,288	473,705	18.50%
$25,000 to $29,999	94,499	568,204	22.20%
$30,000 to $34,999	103,927	672,131	26.26%
$35,000 to $39,999	91,039	763,170	29.81%
$40,000 to $44,999	85,698	848,868	33.16%
$45,000 to $49,999	85,307	934,175	36.49%
$50,000 to $59,999	173,460	1,107,635	43.27%
$60,000 to $74,999	227,331	1,334,966	52.15%
$75,000 to $99,999	308,271	1,643,237	64.19%
$100,000 to $124,999	264,619	1,907,856	74.53%
$125,000 to $149,999	175,919	2,083,775	81.40%
$150,000 to $199,999	212,867	2,296,642	89.71%
$200,000 or more	263,311	2,559,953	100.0%

The base is the cumulative sum of the weighted frequencies, which is 2,559,953.

Determine the standard error of a 50 percent proportion. For this example, the design factor for household income is 1.5 from Attachment A-23.

$$\text{SE}(50 \text{ percent}) = 1.5 \times \sqrt{\frac{99}{2{,}559{,}953} \times 50^2} = 0..47$$

Calculate p_lower and p_upper:

p_lower = 50 – 0.47 = 49.53

p_upper = 50 + 0.47= 50.47

Determine the categories that contain p_lower and p_upper. The first category with a cumulative percentage that is greater than 49.53 is $60,000 to $74,999. The first category with a cumulative percentage that is greater than 50.47 is $60,000 to $74,999. Since p_lower and p_upper fall in the same category, follow the instructions given in step 5 of the procedure for medians.

Define A1, A2, C1, and C2: A1 = 60,000, A2 = 75,000, C1 = 43.27 and C2 = 52.15 Calculate the lower bound and upper bound using these values.

$$\text{Lower Bound} = \left[\frac{49.53 - 43.27}{52.15 - 43.27}\right] \times (75{,}000 - 60{,}000) + 60{,}000 = 70{,}574$$

$$\text{Upper Bound} = \left[\frac{50.47 - 43.27}{52.15 - 43.27}\right] \times (75{,}000 - 60{,}000) + 60{,}000 = 72{,}162$$

Finally, calculate the standard error of the median:

$$\text{SE(median)} = \frac{1}{2} \times (72{,}162 - 70{,}574) = 794$$

Example 4 – Calculating the Standard Error of a Mean

Suppose we wish to estimate mean adjusted person income of females age 25 to 34 in Alabama. The table below summarizes the computation of the terms in the formula for s^2. The PUMS data for Alabama has 2,575 records for females age 25 to 34 that have a non-missing value for person income.

Sample Record	y_i	w_i	$w_i\, y_i$	$w_i\, y_i{}^2$
1	24,931	27	673,150	16,782,616,148
2	32,241	112	3,610,958	116,419,832,264
3	42,053	32	1,345,699	56,590,790,731
4	27,435	59	1,618,643	44,406,888,125
.				
.				
2,575	55,170	80	4,413,572	243,495,190,712
Total		321,622	6,925,235,327	343,669,131,330,670

The mean adjusted income is

$$\bar{Y} = \frac{6{,}925{,}235{,}327}{321{,}622} = 21{,}532.22$$

s^2 is computed as follows:

$$s^2 = \frac{343{,}669{,}131{,}330{,}670 - 6{,}925{,}235{,}327^2/321{,}622}{321{,}622 - 1} = 604{,}915{,}251$$

The design factor for person income in Alabama, from Attachment A-2, is 1.5. The standard error of the mean can now be calculated:

$$\text{SE}(\bar{Y}) = 1.5 \times \sqrt{\frac{99}{321{,}622} \times 604{,}915{,}251} = 647$$

WORKING WITH DOLLAR AMOUNTS

Dollar variables must be adjusted by the adjustment factors supplied on the PUMS file before they are used to form estimates. Also, when comparing the 2017 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 for 2017 includes two adjustment factors for dollar values[5]:

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.

For example, apply ADJINC to the household income variable. 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 2016 and 2017. The adjustment factor will convert amounts in 2017 dollars. Note that the value of ADJINC is the same for all sample cases (its value for 2017 is 1.011189). This is for disclosure avoidance reasons, as otherwise the month of the interview could be identified by the adjustment factor.

For example, the monthly rent estimate needs to be multiplied by ADJHSG to adjust rent into 2017 dollars. On this file, ADJHSG is a factor of 1.000000.

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 column AVG of the first table "All items" in the PDF file at: https://www.bls.gov/cpi/research-series/ [For example, to express year 2000 dollars in terms of 2017 dollars, multiply the 2000 dollars by 361.0/252.9 = 1.427].

[5] See http://census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html for details.

REFERENCES

[1]. ACS Accuracy of the Data (2017): https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.2017.html

[2]. Design and Methodology of the American Community Survey: April 2009, revised 2010: https://census.gov/programs-surveys/acs/methodology/design-and-methodology.html

[3]. Updated CPI-U-RS, All Items, 1977-2017: https://www.bls.gov/cpi/research-series/

[4]. Attachment A: 2017 PUMS 1-Year Design Factors: https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html