1990 CENSUS PUBLIC USE MICRODATA SAMPLE DESIGN ISSUES

Richard A. Griffin and Alfredo Navarro[1]

Introduction

Public Use Microdata Samples (PUMS) are computer files which contain records for a sample of housing units, with information on the characteristics of each unit and the people in it. In order to protect confidentiality of respondents, the Census Bureau excludes identifying information from the records. Within the limits of the sample size and geographic detail provided, these files permit users with special needs to prepare virtually any tabulations of the data they may desire. For 1990, the Bureau of the Census will produce 5% and 1% files for the United States and Puerto Rico as standard products. In addition to the obvious size difference between the two files, the major distinction is the geography around which each file is built. The 5% files are basically county/county equivalent files, whereas the 1% files are metropolitan area files. Each file will show as many of the various levels within the geography hierarchy as possible while still preserving the disclosure rules of only releasing geographic units which have at least 100,000 persons. In addition, the Census Bureau also will produce PUMS files for Guam and, perhaps, cost reimbursable PUMS files for the Elderly Population. The 1990 PUMS files furnish nearly all the detail recorded on long form questionnaires in the census. Thus, with only minor exceptions, PUMS files contain the full range of population and housing information collected in the 1990 census. Some examples are: age by single years up to 90, marital status, sex, educational attainment, industry, occupation, income, rent/value, source of water, yearly cost of electricity, gas and property tax, and much more.

PUMS samples will be useful to users (1) who are doing research that does not require the identification of specific small geographic areas or detailed cross tabulations for small populations, and (2) who have access to programming and computer time needed to process the samples. Microdata users frequently study relationships among census variables not shown in existing census tabulations, or concentrate on the characteristics of certain specially defined populations, such as unemployed homeowners or families with four or more children.

Protecting Confidential Information

Records on PUMS files contain no names or addresses. Also the Bureau limits the detail on place of residence, place of work, high incomes, and selected other items to further protect the confidentiality of the records. As mentioned above, PUMS records identify no geographic area with fewer than 100,000 inhabitants. Microdata samples include only a small fraction of the population, drastically limiting the chance that the record of a given individual is even contained in a microdata file, much less could be linked to the corresponding respondent.

Bias

For the 1980 PUMS, a stratified systematic selection procedure with probability proportional to a measure of size was used to select each sample. The measure of size was the full sample weight that resulted from the 1980 census ratio estimation procedure. For occupied housing units, the full sample person weight assigned to the householder of the unit was used. For GQ persons, the full sample person weight was used, while for vacant housing units, the full sample housing unit weight was used.

The 1980 PUMS were self-weighting. The data user could estimate the frequency of a particular characteristic for the entire population by tallying records from the microdata files that had the characteristic and multiplying the result by the inverse of the sampling rate, e.g., multiplying raw counts from a 5% PUMS by 20. Ten years ago, the Bureau felt that this self weighting property was important for a substantial portion of potential PUMS users.

Sample selection with probability proportional to census sample weight was done primarily to accommodate self-weighting. Equal probability sampling from the census sample would require differential weights on the PUMS files for each sample person or housing unit. However, in the case of occupied housing units, using the census weight of the householder results in a slight bias for estimates for persons or housing units. Note that it is necessary to choose one weight to represent the occupied housing unit and using the householder weight is a good choice. Estimates for persons (housing units) are biased to the extent non-householder person (housing unit) weights differ from householder person weights. The nature of this bias is as follows:

We would like the conditional expected value of the PUMS estimate, given the census sample, to be equal to the census sample estimate. Let

n = the number of persons in the census sample.

---

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect that of the Census Bureau.

w_{hhi} = the census weight for the householder in the occupied housing unit in which census sample person i lives.

w_i = the census weight for census sample person i.

a_i = 1 if person i is in the PUMS sample given the person is in the census sample; 0 otherwise.

E_1 = conditional expectation given the census sample.

TE = the PUMS take every (i.e., 20 for a 5% PUMS).

x_i = the value of characteristic X for census sample person i.

$\hat{X} = TE \sum_i a_i x_i$. This is the self-weighting PUMS estimate for X.

$\hat{X}_c = \sum_i w_i x_i$. This is the final census sample estimate for X.

$$\text{Bias}_1 \hat{X} = E_1(\hat{X}) - \hat{X}_c = TE \sum_i (W_{hhi}$$

$$= \sum_i x_i (w_{hhi} - w_i) .$$

Thus, the conditional bias for a person estimate is a function of the differences between the weight of the householder and the weights of the non-householders within census sample households.

It was recognized that it was possible to correct for some of this bias by multiplying the take every by the ratio of the estimated number of persons using the head of household weight for each person to the estimated number of persons using the full census sample person weights. This adjustment is most helpful for person characteristic estimates. For housing characteristic estimates, the adjustment is not as good since ideally the ratio would be formed based on analogous estimates of the number of housing units.

Preliminary plans for the 1990 Census called for using this adjustment to the take every (calculated separately for each PUMS sampling stratum) and also including on the PUMS files the approximate unbiased weight (i.e., inverse of probability of selection) for each person and housing unit. Users would have been given the option of accepting the bias, much of it hopefully corrected due to the take every adjustment, by using self-weighting or using the unbiased weight which would be different for each sample person or housing unit.

Overlap

The 1990 Census sample is about 17% of all persons and housing units. At the time of PUMS sample selection, we wanted to partition most of the census sample into PUMS samples. From these PUMS samples the 1%, 5% and elderly PUMS files would be created and the remaining PUMS files kept in reserve for future PUMS file requests. This multiplicity of files will put the Census Bureau in a position to provide users with many files in which we can vary levels of detail on specific items and vary geography. For example, we might release a national level file with extremely fine detail on all characteristics. In fact, a national level file may be able to have elevated topcodes. For the 5% and the 1% files the wages or salary character-istic, for example has a topcode of $150,000. In a file with less geographic detail, this topcode could be raised to say, $200,000. For any person with a wage or salary value greater than this, the median wage or salary of all persons in the state with a wage or salary above the topcode is shown instead of the value collected in the census. In addition to a national file, the Census Bureau may release a regional file with less detail than on the national file. This idea can be extended to state files, metro-files, and so on; however, at each level there is a measured trade-off between geography and variable detail [1].

There is a potential overlap problem when selecting multiple PUMS files with probability proportional to full census sample weights. An overlap occurs when a census sample person is in more than one PUMS sample. As each census sample person on the sample edited detail file is processed through PUMS sample selection, the weight of the householder (or individual if group quarters) is added to the cumulation of all previous weights and compared to multiples of the take-every. If the cumulation exceed a multiple, the person is in that PUMS sample. The take- every multiples are kept separate for each PUMS sample. Thus, a census sample person can be designated in more than one PUMS sample. For example, consider a 5% and 1% sample selection. The take-every for the 5% sample is 20 and the take-every for the 1% sample is 100. Suppose the cumulation of previous weights is 97, the take-every counter for the 1% sample is 97 (i.e., the random start was 0 and there have been no hits), the take-every counter for the 5% sample is 17 (i.e., the random sample was also 0 and there have been 4 hits) and the next census sample person has a full census sample weight of 5. Then this person will be in both the 5% and 1% PUMS samples [2].

Due to disclosure avoidance concerns, the Census Bureau wants the probability of a record being in more than one PUMS sample to be very small, zero if possible. Preliminary indications were that the Census Bureau's Microdata Review Committee would accept a probability of overlap as high as 5%

for any individual. Suppose we desired to designate one 5% PUMS and ten 1% PUMS. The more PUMS samples to be designated, the higher the probability of overlap. It did not appear that the usual sampling scheme of applying a take-every to the census head of household weights could be used to select this many PUMS samples while maintaining a probability of overlap less than 5%. For example, using a 1980 Census weight distribution from the State of Rhode Island, one 5% and five 1% PUMS samples can be selected so that the probability of overlap for a record in the 5% sample is about 4% and the probability of overlap for a record in the 1% sample is about .5%. Adding additional 1% samples makes it difficult to keep the overlap probabilities low enough. Other weight distributions that would occur in the 1990 Census would present the same problem.

While the probability proportional to census weight selection scheme does have some variance reduction properties, the primary reason for its use in PUMS in the past has been, as mentioned earlier, to enable users to use the self-weighting option without a significant bias. We asked data users if they would be comfortable using weights and discovered that they felt that using weights would be no problem. If we kept the self-weighting option, we would need to reduce the amount of census sample available for PUMS samples in order to maintain the desired probability of overlap. This would reduce our capability of being responsive to future requests for PUMS files. If we were willing to not have a self-weighting option, we could reduce the probability of overlap to 0 by sampling the census sample with equal probability. In doing this, we could partition the entire census sample into PUMS samples. The only price for this benefit would be the self-weighting option. We decided to pay this price.

Properties of Basic Sample Design

The 1990 PUMS sample design will be basically as follows (details are given in the Detailed Sample Design section of this paper). First, within each PUMS stratum the records are sorted by census designated sampling fraction (1/2, 1/6, or 1/8) and within sampling fraction by weight of head of householder. Suppose the overall census sample is K percent and the random start in a PUMS stratum is 1. Then the first record in the stratum is in the first PUMS sample, the second record is in the second PUMS sample and so forth up to the Kth record in the Kth PUMS sample. Then the procedure starts over again. Each sample person is given a PUMS weight equal to the census full sample person weight times K and each sample housing unit is given a PUMS weight equal to the census full sample housing unit weight times K.

Properties of this scheme are:

1. It produces K 1% PUMS samples with no overlap.
2. It will be necessary for a PUMS user to use the weights on the file. The bias from self weighting would be much to large.[2]
3. The conditional expected value (given the census sample) of person or housing unit estimates from PUMS will be equal to the full census sample estimate of that characteristic.
4. Each PUMS sample will have a distribution of census weights that is probably very close to the full census sample weight distribution.
5. For the 5% PUMS samples, 5 of the 1% samples can be combined and each weight divided by 5.

Effect of Possible Adjustment of the Census on PUMS

Due to the possibility of adjustment it was necessary for us to consider the effects of adjustment on PUMS. If the census has not been adjusted then these planned procedures are documented in case they are ever needed. Plans, if needed, are as follows.

Sample data is obtained for each 100% count adjustment record (overcount or undercount) by matching to a sample data record for a person in the same geographic area on the basis of their 100% characteristics. Once a match is found, the sample data is substituted to the adjustment record. The overcount cases will have a full census sample weight of negative 1 and the undercount cases will have a full census sample weight of positive 1. Since the PUMS weight is the full census sample weight times K, overcount cases selected in a PUMS sample would have a negative weight.

Most data users have expressed their preference to not have negative weights on the PUMS files. This preference is primarily based on software considerations. Thus, we want to make it possible for users not to have to use negative weights without biasing the PUMS estimates.

Count adjustment persons with negative weights will be single persons, one record for each (overcount) person. A mechanism was developed to establish a link between the donor (census sample record) and the donee (overcount adjustment record). Overcount adjustment records will be ignored for PUMS sample selection. At the time of sample selection for PUMS, the weight of the donor records (for the overcount) will be reduced for PUMS purposes only. The weight reduction is defined by

---

[2] For the 1% PUMS, Bias = $\sum X_i \left(\frac{100}{K} - W_i \right)$ is the conditional bias given the census sample were X_i is the characteristics, W_i is the full census sample weight and the sum is over the census sample. The variability in census weights makes this bias too large.

the number of times the person was used as a donor. No donor for an overcount adjustment record will be used more times than the donors original census weight (to avoid a negative census weight after reduction and a resulting negative PUMS weight). In general, if r is the number of times a census sample record is used as the donor for overcount records and the census weight is w_i then the PUMS weight is $K(w_i-r)$, where 1/K is the probability of a record being in a particular PUMS sample given it is in the census sample. In general, r will equal 1, except under some special circumstances.

If both the 100% count adjustment overcount record and the matching sample record are in the PUMS tabulation area, then the conditional expected value (given the census sample) of the PUMS estimates will be equal to the corresponding full census sample estimates. If the matching sample record is in the tabulation area but the 100% count adjustment overcount record is out, there will one a negative bias equal to the value of the sample characteristic. If the 100% count adjustment overcount record is in the tabulation area but the matching sample record is out, there will be a positive bias equal to the value of the characteristic. In most cases both persons in n match pair will be in the same PUMS tabulation area.

In the event of count adjustment, count adjustment records will not have the relationship item. Thus, PUMS estimates for sample data by relationship should not be calculated using these adjusted person weights (i.e., $K(w_i-r)$. Due to this fact, both adjusted and unadjusted PUMS person weights will be placed on the PUMS files. The unadjusted weights are to be used for relationship tabulations and the adjusted weights are to be used for all other tabulations. This should not cause users a problem. Note that for the vast majority of PUMS sample records the two weights will be the same.

Detailed Sample Design - Overview

There are K 1% PUMS samples selected. K is a function of the full census sample observed sampling rate and is calculated separately for each state. For example if the full census sample observed sampling rate in a state is 17.2%, then K, for that state is 17. This is done in order to ensure that, for example, the 1% PUMS file for each state has a sample size very close to 1% of the population of the state. The observed sampling rate for the census varies from state to state due in large part to the full census sample design. Thus, if K were calculated once at the national level and used for all states then some states would have considerably more and some considerably less sample than desired. This problem will still occur for substate geographic areas but would be too complex to correct for all possible PUMS tabulation areas. From these K 1% samples, one 5%, one 1%, and one 5% elderly PUMS files are selected. The remaining K-11 percent files will be used to create subsequent PUMS files, as requested.

Puerto Rico is treated just like a state except no elderly PUMS file is selected. Guam has no census sample. That is all questions including those equivalent to the stateside sample questions are asked of everyone in the population. For Guam, one 10% PUMS file is selected. Since, in effect, all "regular" Guam census weights are 1, each PUMS record will have a PUMS weight of 10. Note that since the Guam PUMS sample will be a "perfect" sample within demographic strata (i.e., the observed sample within strata is exactly the same as the designated sample), no ratio estimation is necessary. Thus, the Guam PUMS sample is self-weighting.

The details that follow pertain to the stateside PUMS.

Stratification

A stratified systematic selection procedure with equal probability independent within select the PUMS samples. The sampling universe is defined as all occupied housing units including all occupants, vacant housing units and group quarters (GQ) persons, including count adjustment persons with a weight of 1 (undercount) but excluding count adjustment persons with a weight of -1 (overcount).

The sample units are stratified during the selection process. The stratification is intended to improve the reliability of estimates derived from the 1%, 5%, and the elderly samples by defining strata within which we know from experience there is a high degree of homogeneity among the households within each strata with respect to characteristics of major interest.

A total of 1101 strata are defined; 936 household strata, 156 strata for GQ persons, and 9 strata for vacant housing units. First, the units are divided into three major groups: households, vacant housing units, and GQ population. The household universe is stratified be family type and non-family, race/hispanic origin of the householder, tenure, and age within sampling stratum. The family type strata are family with own children under 18 and family without children under 18. The race/Hispanic origin strata are as follows: White/other by Hispanic and Non-Hispanic; Black/American Indian, Eskimo or Aleut; Chinese; Filipino; Hawaiian; Korean: Vietnamese; Japanese; Asian Indian; Somoan; Guamanian; and other Asian and Pacific Islanders. The tenure strata are owner and renter. Each household is placed in a age stratum based on the age of the oldest household member. The age stratification is included to ensure that households with elderly individuals are sampled at exactly the correct rate for the elderly PUMS. The age strata are 0-59, 60-74, 75-89, and 90+. The sampling strata are by census sampling rate 1-in-2, 1-in-6, and 1-in-8. For the census sample selection the

population was stratified by geographic size into three sampling strata, i.e., units in small governmental units were sampled at 1-in-2, units in small tract/block numbering areas were sampled at 1-in-6 and the remainder of the units were sampled at 1-in-8. Stratifying by sampling rate helps ensure that the full census sample weight distribution of each PUMS sample is close to the distribution of weights in the full census sample.

The vacant housing units universe is stratified by vacancy status (for sale, for rent, other) and sampling rate. Finally, the GQ population is stratified by GQ type (Institutions, non-institutions, undercount adjustment persons), race, Hispanic origin, and age.

Allocation of Units to PUMS Samples

For each stratum i, a random number, R_i, between 1 and K is selected and the units are sorted by county/place/MCD/Tract or Block Numbering Area and householder weight or person's weight for GQ units or housing unit weight for vacants. Each unit is assigned to only one of the K PUMS. The first unit in the stratum is assigned to random integer R_i. The next unit is assigned the number $R_i + 1$. This is continued until a unit is assigned the number K at which time the process starts over with 1 and is continued in this fashion until the last unit is assigned a number. All units with the same K number are merged into one of the K PUMS. For instance, units that are assigned the number 5 constitute the fifth of the K samples.

Calculation of PUMS Weights

Let w_{pc} denote the full census sample person weight and w_{hc} denote the full census sample housing unit weight. PUMS weights are defined as a function of these weights and K. For the 1% PUMS file the PUMS person weight is defined as $w_p = Kw_{pc}$ and the PUMS housing unit weight is defined as $w_h = Kw_{hc}$. For the 5% PUMS file and the 5% elderly PUMS file, the PUMS person weight is defined as $w_p = Kw_{pc}/5$ and the PUMS housing unit weight is defined as $w_h = Kw_{hc}/5$. (Note: These are the unadjusted weights, changes in weights in the event of count adjustment were explained in the section on the effects of count adjustment).

Creation of the PUMS Files

Eleven distinct random numbers from 1 to K are generated. The samples corresponding to the first 5 random numbers are merged and sorted and this file is designated as the 5% PUMS file. The sample corresponding to the sixth random number is placed in the proper sort and designated as the 1% PUMS file. All the units in the 60 and over age strata (60-74, 75-89, 90+) of the samples corresponding to the last 5 random numbers are stripped off, sorted and designated as the 5% elderly PUMS file. The remaining K-11 samples are reserved for future data requests.

Reliability Considerations

Reliability considerations were based on the coefficient of variation (CV). The CV is the relative value of the standard error, specifically, the ratio of the standard error of an estimate to the expected value of the estimate.

1% and 5% PUMS samples [3] CV's were evaluated by race and Hispanic origin and for the total population at the national, state and city level. City population sizes considered were 100,000; 250,000; 500,000; and 1,000,000. The CV's were calculated for a 5% to 20% data item (p values) with an increment of 5%.

The CV for estimates of the total population at the national, state and city level range from less than 1 tenth of a percent for a 20% data item at the national level for the 5% PUMS sample to 14% for a 5% data item for a city of population 100,000 for the 1% PUMS sample. These results are very encouraging from a reliability point of view. The CV increases dramatically in some cases for estimates for race and Hispanic origin group. For tabulation areas where the concentration of minorities is very small, the CV's for estimates of minorities are not adequate. We will warn users to make use of this data with caution. For example, the CV's for the American Indian, Eskimo and Aleut population in a city of 500,000 are over 50% for the 1% PUMS for the p values we examined. The CV's for estimates of American Indians, Eskimos and Aleuts are also over 50% in the state of Wyoming in the 1% PUMS for p values of .15 or less. This is to be expected since American Indians, Eskimos and Aluets only account for (on the average) 2 tenths of 1% of the total population in cities with over 500,000 population and they represent less than 5 tenths of 1% of the population in Wyoming. However, in most cases, the CV's are under 20%, and thus, the reliability for estimates from the 5% and 1% PUMS are adequate.

5% Elderly PUMS sample [4]

At the start, we assumed that the elderly PUMS sample would not be larger than 5%. We felt that 5% would give sufficient reliability and taking more would hurt our plans for keeping a reasonable number of 1% PUMS files in reserve for future file requests.

The smallest tabulation area will be 100,000 persons. For a given data item (say p=.1), the reliability of the estimate is determined by the sampling fraction and the size of the area. Given sampling fraction and p value estimates for larger areas will have better reliability than estimates for smaller areas.

Data users have expressed interest in PUMS for the population group defined as 85 and older in addition to the 65 and older group and the 60 and older group. Since the 85 and older group is

obviously smaller, we based our primary reliability analysis on this group. We used estimates of population by age and sex for 1989 in our analysis. We assumed simple random sampling and that the distribution of population by age for cities of 100,000 or more is similar to the size distribution at the state level (since there are no publications with percentages for the 85+ group below the state level).

Since the lowest percentage of elderly (age 85+) is .25% and the largest is about 2%, we evaluated the CV's from .25% to 2% in increments of .25%. The 65+ group is between 5 and 15% of the total population and the 60+ group is at least 15% of the total population. The 65+ group was evaluated for 5% and 10% of the population and the 60+ group was evaluated for 15% of the population.

Several potential data users of elderly data are interested in estimates for different size areas. In our analysis, we considered three size categories ranging from 100,000 to 500,000 in population. The 100,000 is the worst case scenario from a reliability point of view, the CV's for such an area for an estimate of p=.05 or p=.1 and a sample size of 1% tend to be over 100%. In order to reduce these CV's, we decided to select a 5% elderly PUMS sample. Even at the 5% sample size, CV's tend to be over 50% for a p=.05, area size of 100,000 and percent of population who are elderly (85+) less than 1.5. For cities larger than 100,000 (250,000 or 500,000) CV's decrease and thus reliability increases. For a p value less than 5% the corresponding CV's will be higher. A number of states were evaluated based upon their size and proportion of population in the 85+ group. With a 5% PUMS most of the CV's were reasonable at the state level for estimates for the 85+ group. Estimates for the 65+ and 60+ groups were found to be good at the city and state level with a 5% elderly PUMS.

Guam PUMS sample [5]

We performed an evaluation to see if a PUMS sample larger than 5% was necessary for Guam. The goal for the Guam PUMS was to achieve a CV no greater than 50% for an estimate of a 10% population characteristic for all ethnic groups. The criterion CV = .5% implies that we should choose the sample size n to be sufficiently large that the standard error of the proportion p being estimated equals one half of p.

We felt that the sample size required to achieve the proposed reliability goal might be significantly greater than 5% due to Guam's multiethnicity. There are 14 ethnic groups in Guam for which users will be interested in PUMS estimates. These range from Chamorro with 41.8% of the population and Filipino with 21.2% to Guamanian with .8% of the population and other islander with .5%. The estimated 1990 total population of Guam is about 115,000.

Calculation of CV's indicated that the design goal is achieved with a sample size of 10%. The CV for an estimate of a 10% population characteristic is less than 35% for all ethnic groups except other islander for whom the CV is 46%. By contrast, other islander has a CV of about 67% for a sample size of 5%.

A 10% PUMS will be selected for Guam. The use of a 10% sampling rate for a PUMS file is not a precedent. The 1980 Indian PUMS was a 10% sample.

Variance Estimation Considerations

States and Puerto Rico

As in 1980, we will use the same design effects as produced for the full census sample along with a simple random sample standard error formula/table for the appropriate sampling rate. The ratio of the PUMS sample design standard error to this appropriate simple random sample standard error should be about the same as the census design effects since the weights are simply multiplied by a subsampling factor and the ultimate cluster is still the household.

Guam

Since there is no census sample for Guam, a set of design effects calculated specifically for the PUMs file will be calculated. The actual variance formula for a 1-in-10 systematic sample will be used for selected characteristics. This is possible since all the census data is available for each of the 10 systematic clusters. In each case, the actual standard error will be divided by a 1-in-10 simple random sampling standard error to produce design effects for Guam. The selected characteristics will be grouped and the group design effects averaged to produce the design effects that will be published. The simple random sample standard error formula/table for a 10% sample will also be published so that data users will calculate confidence intervals for Guam PUMS estimates in the same way that they are calculated for all other census data products.

References

[1] Greenberg, B., Creating 1990 Decennial PUMS, internal Census Bureau memorandum, 9/19/89.

[2] Garland, M.G., Sampling Specifications for New Public-Use Microdata Samples - D Sample, E Sample, and Puerto Rico A, B, and C Samples, internal Census Bureau memorandum, 1/24/85.

[3] Kohn, F., Public Use Microdata (PUMS), internal Census Bureau Memorandun, 8/24/90.

[4] Kohn, F., 1990 Elderly Public Use Microdata Sample (PUMS), internal Census Bureau memorandum, 5/14/90.

[5] Thompson, J. H., Reliability and Sample Size analysis for the Guam Public use Microdata Sample (PUMS), 9/13/90.