ANALYSIS AND VARIANCE ESTIMATION WITH THE IPUMS

As described in Chapter 2 and Chapter 3 of the IPUMS documentation, IPUMS employs a variety of sample designs which have a measurable impact on sampling standard errors. While appropriate use of sampling weights will produce correct point estimates (e.g., means, proportions), many researchers believe that it is also necessary to use additional statistical techniques that account for the complex sample design to produce correct standard errors and statistical tests. IPUMS has provided the STRATA and CLUSTER variables for this purpose.

For samples prior to 1960, pseudo strata were created based on microfilm page ranges (1850-1930) or enumeration districts (1940-1950), which serve as a proxy for geographic stratification (see "Construction of Strata in the IPUMS Samples"). Using these variables in analysis performs well as an alternative to relying solely on the original design variables (HHWT, PERWT, and SLWT).

For the 1960-2000 samples, strata were created based on the stratification criteria used to select PUMS samples such as household size, age, race, ethnicity, tenure, group quarters membership, and vacancy status. For more information on the creation of strata, see this link: "Construction of Strata in the IPUMS Samples".

For the American Community Survey samples strata were created based on the lowest level of geography available in each sample. For the 2000-2004 samples, each state forms a stratum. In the 2005 onward ACS samples, strata are defined as unique Public Use Micro-data Areas (PUMA). In addition, IPUMS also provides replicate weights for use with the 2005-onward ACS samples. More information on using replicate weights can be found here: Replicate Weights in the American Community Survey/Puerto Rican Community Survey.

IPUMS TECHNICAL VARIABLES FOR ANALYSIS AND VARIANCE ESTIMATION

Three Technical Variables Are Needed for Analysis of the IPUMS Data:

1. A sampling weight must be chosen (PERWT, HHWT, or SLWT), based on the type of question a researcher is trying to answer. Analysts should review variable descriptions of the variables of interest and the "Sampling Weights" note for more information about which weight to use.
2. STRATA is an integrated variable that represents the impact of the sample design stratification on the estimates of variance and standard errors.
3. CLUSTER is an integrated variable which uniquely identifies each household record in a given sample.

GENERAL SYNTAX TO ACCOUNT FOR SAMPLE DESIGN

The following general syntax will allow users to account for sampling weights and design variables when using STATA, SAS, or SAS-callable SUDAAN to estimate, for example, means using IPUMS data.

STATA

```
svyset cluster [pweight=perwt], strata(strata)
svy: mean var1
```

SAS

```
proc sort data = datasetname;by strata cluster;

run;

proc surveymeans data = datasetname;

weight perwt;

strata strata ;

cluster cluster;

var var1;

run;
```

SAS-Callable SUDAAN

```
proc sort data = datasetname;

by strata cluster;

run;

proc descript data = datasetname filetype = sas design = wr;

nest strata cluster;

weight perwt;

var var1;

print nsum wsum mean semean / nohead;

run;
```

SYNTAX FOR SUBPOPULATION ANALYSIS

The following syntax demonstrates, generally, how an analyst can conduct subpopulation analysis using IPUMS data without compromising the design structure of the data. This approach has the effect of producing estimates for the population of interest, while incorporating the full sample design information for variance estimation.

STATA

```
svyset cluster [pweight=perwt], strata(strata)

svy, subpop(if age >= 65): mean var1
```

SAS

```
subpopvar = 1 if age ge 65;

else subpopvar = 0;

proc sort data = datasetname;

by strata cluster;

run;

proc surveymeans data = datasetname;

weight perwt;

strata strata ;

cluster cluster;

domain subpopvar;

var var1;

run;
```

SAS-Callable SUDAAN

```
proc sort data = datasetname;

by strata cluster;

run;

proc descript data = datasetname filetype = sas design = wr;

nest strata cluster;

weight perwt;

subpopn age >= 65/NAME = "Population 65 years and older";

print nsum wsum mean semean / nohead;

run;
```

▲ Back to Top