Stata: Survey Analysis in Stata

Topics: Adjusting for stratification, clustering, oversampling and other complex survey design characteristics in Stata

```
svyset [pweight=weight], psu(v021) strata(v023)
```

1. Complex Sampling Characteristics

This lecture introduces the survey-set statement, a single line of code at the start your Stata data analysis .do file which specifies the survey design of your dataset. All subsiquent analyses, in particular descriptive and bivariate statistics, must account for the survey design to produce unbiased mean and accurate variance estimates.

This lecture draws directly from online materials developed by the Carolina Population Center at the University of North Carolina-Chapel Hill, and it has been produced with their permission.

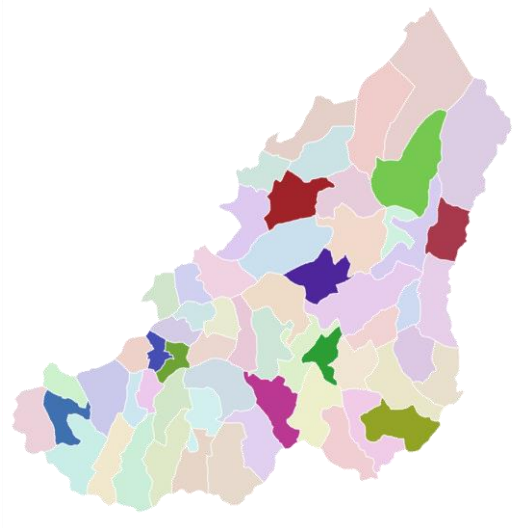
www.cpc.unc.edu/research/tools/data_analysis/statatutorial/sample_surveys

1a. Clustering

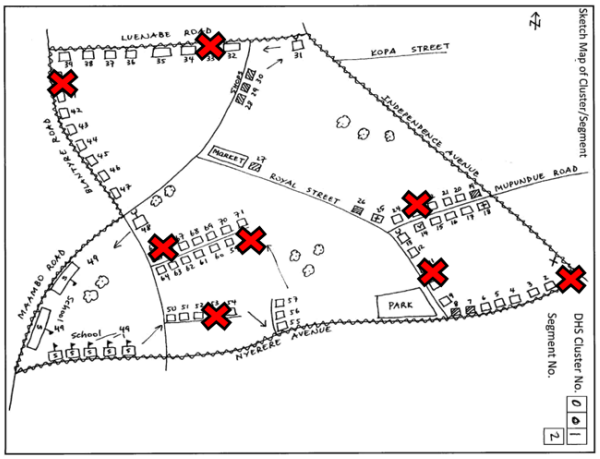
A truly random selection of households would involve listing all the households in the country and randomly selecting the desired number of households from that list. Using this approach, each household would have an independent and equal chance of being included in the survey. While this approach is statistically ideal, the cost of first enumerating all households and then visiting the selected households that would be scattered all over the country make this approach impractical.

A more practical approach is cluster sampling. For example, in the Demographic and Health Surveys, "enumeration areas" from the Census or similar national surveys are first selected randomly from a list of all such areas in the country (or within strata if stratification is being used). These areas are often referred to as "clusters" or as "primary sampling units" (PSUs). They may be towns or villages, or they may be census tracts in cities.

First, sample clusters (primary sampling units)

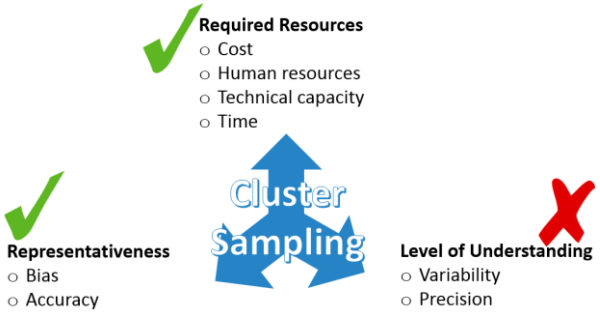


Second, sample households from clusters



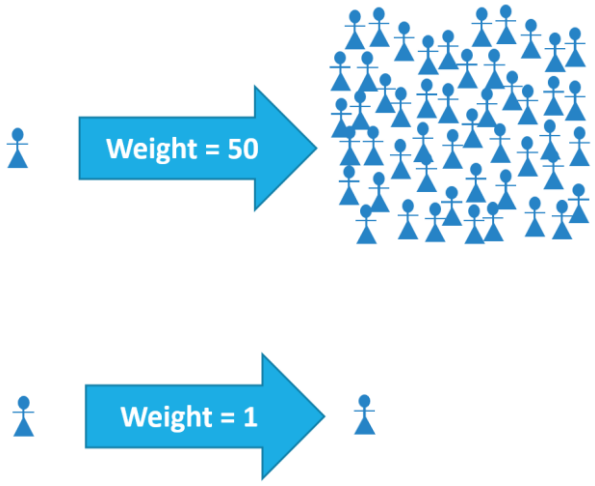
Generally each cluster contains roughly the same number of households.

The next step is to enumerate (count and label) all the households in the cluster. Then a random-selection process is used to select households within each cluster. This is the sample of households that will be visited for the survey.



While cluster sampling is much more practical, it also means that the households are not statistically independent. Instead, the characteristics of a given household (and its household members) are more like other households in the same cluster, and are less like households in other clusters. This effect of a non-independent sampling process, called the "sample survey design effect", shows up in the standard error of estimation statistics (means, regression coefficients). Clustering tends to incorrectly decrease the size of standard errors, leading to a greater chance of finding statistical differences that are not real. By correcting for the effect of clustering (design effect) we account for the fact that people in the sample are similar in some way because they are from the same communities.

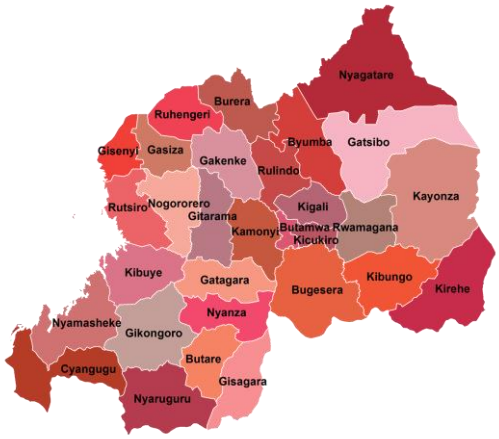
1b. Sampling Weights



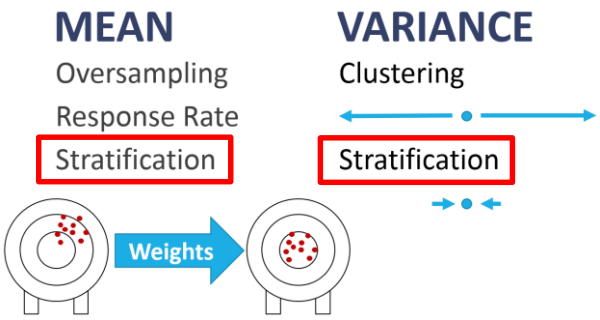
Each observation in the sample is chosen using a method of random selection. An important property of this method is that the probability of selection may not be equal for all members of the population. The sampling weight for each observation is computed as the inverse of the selection probability. Additional adjustments (such as non-response) may be made to the sampling weights. An observation with a sampling weight of 1000 represents one thousand individuals from the target population while another observation with a sampling weight of 50 represents only fifty individuals. This is why the DHS normalizes the sampling weight around one; an average person in

the sample represents one person in the real world.

1c. Stratification



The population can be divided into sections (strata) that are internally more homogeneous, or similar. This may be done in order to over-sample smaller groups in a target population. Examples of strata are region of country, urban/rural residence, or education level. A separate sample is selected from each stratum. Like oversampling, stratification can bias the mean estimates and is therefore incorporated into sampling weights. And like clustering, the observations within strata are not statistically independent so stratification can effect variance estimates. As a result, stratification is adjusted for in two ways: as part of the sampling weights, and a separate adjustment on the variances.

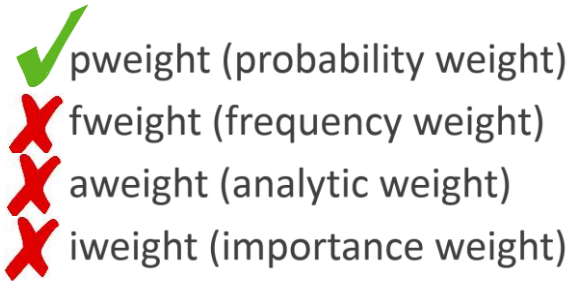


Researchers may not adjust for the effect of stratification on variances in two circumstances: (1) when conducting subpopulation analyses because there may be too few observations per strata to produce variance estimates, or (2) when performing multi-survey analyses where the units of stratification differed across surveys. When we do not specify the strata in our survey design, confidence intervals might be widened very slightly, resulting in slightly more conservative estimates and conclusions, which is fine.

2. Choosing the Correct Weight: pweight

One of the most common mistakes made when analyzing data from sample surveys is specifying an incorrect type of weight for the sampling weights. Only one of the four weight keywords provided by Stata, pweight, is correct to use for sampling sampling weights. The pweight command causes Stata to use the sampling weight as the number of subjects in the population that each observation represents when computing estimates such as proportions, means and regressions parameters. A automatically be used to adjust for the design characteristics so that variances, standard errors and confidence intervals are correct.

In the Demographic and Health Surveys, the pweight variable is stored with no decimal points,



so you must divide the relevant weight variable by 1,000,000. The weight variables include the women and children's weight v005, male weight mv005, household weight hv005, and domestic violence module weight d005.

The fweight (frequency weight), aweight (analytic weight), and iweight (importance weight) are incorrect and should not be used for survey data analysis. Visit the UNC Carolina Population Center website for a concise explanation of why these other weights should not be used.

3a. svyset statement

Before any of the survey estimation commands can be used, the svyset command should be used to specify the variables that describe the stratification, sampling weight, and primary sampling unit variables. The svyset statement follows this syntax: `svyset`, square bracket around the pweight where pweight equals the probability weight variable, comma, psu option with the psu variable in parentheses, and if desired, strata option with the strata variable in parentheses.

Alternatively, the syntax can be written `svyset` statement, PSU variable, square bracket around pweight equals…, comma, and then the strata option.

3b. svyset statement applied

```
* Survey Design
gen weight = v005/1000000
svyset [pweight=weight], psu(v021) strata(v023)
```

Here is an example that uses the survey set statement from the 2010 Rwanda Demographic and Health Survey. First, we introduce six decimal values to the sampling weight variable, v005, by generating a new variable called weight equal to v005 divided by 1,000,000. Then we specify the survey design in a svyset statement where the probability weight is equal to the variable weight, the primary sampling unit is equal to v021, and the strata is equal to v023. You can use these same variables to specify the study design in other DHS datasets. Find the equivalent variables if working with a different population survey dataset.

Finally, after we have specified the survey design, we tell Stata to account for this survey design in any subsequent analyses by inserting `svy:` before

```
* % of children in homes with charcoal fuel
tabulate charcoal // NOT accounting for survey design
svy: tabulate charcoal, percent
```

the rest of the command. For example, to estimate the percent of children whose families cook with charcoal *NOT accounting* for the survey design, we would write `tabulate charcoal`, and to perform the same analysis *accounting* for the survey design, we would write `svy: tabulate charcoal`.

4a. Subpopulation analysis

When using the svy commands to analyze only a portion of the sample (a sub-population), it is important to analyze the entire data set and to use the subpopulation option to identify those observations you want to include in the estimate. This is because Stata needs to have information from every observation in the sample to compute the variance, standard error, and confidence intervals even though only the observations in the sub-sample are needed to compute means, proportions, and regression coefficients.

4b. Subpopulation analysis applied

```
* Kigali
recode v023 (1/3 = 1 "Kigali") ///
    (4/30 = 0 "Not Kigali"), gen(kigali)

* % children charcoal fuel at home, in Kigali
svy, subpop(kigali): tabulate charcoal
```

To use the `subpop` option, we need to generate a variable that has a value of 1 for the observations in the sub-population and a value of 0 for those that should be excluded. Here is an example where we create a variable called Kigali for those people living in one of the three districts that comprise the capital city of Rwanda, Kigali, where 1=Kigali and 0=not Kigali. To estimate the percent of children whose families cook with charcoal in Kigali only, we write `svy, subpop(Kigali): tabulate charcoal`.

Check out the Carolina Population Center website for additional Stata analysis learning material.