Lab 3: Stratified Random Sample Selection and Analysis Using SAS

Introduction

Most surveys, especially large ones, involve stratification of the population, so it is important to know how to select a stratified random sample and how to analyze the data under a stratified design. In this material we will be working with the Agricultural Census sampling frame. We will draw a stratified random sample using PROC SURVEYSELECT from the sampling frame in two ways:

1. We will use a STRATA statement in PROC SURVEYSELECT.
2. We will partition the frame into strata, then use SURVEYSELECT to draw an SRS from each stratum.

We will obtain population mean and total estimates using PROC SURVEYMEANS with STRATA statements and WEIGHTS statements. In the first selection method mentioned above, we will use the sampling weights provided by SURVEYSELECT. In the second selection case, we will use a DATA step to assign sampling weights to the observations.

Goals:

1. To learn how to use PROC SURVEYSELECT to draw a stratified random sample from a sampling frame.
2. To understand that stratified random sampling is the same as selecting independent samples from each stratum.
3. To learn how to use PROC SURVEYMEANS to estimate population means and totals with STRATA and WEIGHT statements.
4. To learn how to use PROC SURVEYMEANS to estimate stratum means and totals with BY statements.
5. To understand the connection between stratified population estimates and stratum estimates.
6. To recognize what self-weighting implies about the stratified mean estimate.

Setting Up the Sampling Frame:

We are working with all 3078 of the US counties included in the Agricultural Census. (agpop.csv). The variables we will be using are the acres92 variable and the region variable. We will be using region to create our strata.

1. Submit the OPTIONS, FILENAME, and DATA step to read the sampling frame into SAS.
2. Knowing the stratum sizes will be useful to us when we perform analyses. Submit the PROC FREQ statement to obtain the stratum sizes.

Selecting a Stratified Random Sample Using PROC SURVEYSELECT:

Before we can select a stratified random sample, we must sort the data by the stratification variable, in our case region. The reason for this requirement is that SAS expects units to be grouped together by stratum so that SURVEYSELECT knows how to partition the sampling frame.

1. Submit the PROC SORT statement to sort the sampling frame by region. It is important to note that region is in alphanumeric format, so SAS sorts alphabetically. We need to know the order of region so that we can correctly match sample sizes to strata in the SURVEYSELECT statement.
2. Submit the PROC SURVEYSELECT statement to draw the stratified random sample. The SURVEYSELECT syntax is very similar to the syntax we used when selecting an SRS, except now we have a SAMPSIZE= statement and a STRATA statement. The SAMPSIZE= statement is where we provide the sample sizes for each stratum. They are listed in the order that the values of our stratification variable occur in the sorted data set. In our case, the data set is sorted so that the North Central counties are first, then the Northeast, then the South, and last are the West counties. So the sample sizes in sorted order are 103 from the North Central, 21 from the Northeast, 135 from the South, and 41 from the West.

   The STRATA statement is where we put our stratification variable, or in other words, where we tell SAS what variable that identifies what stratum an observation belongs to. To generate a STS with SURVEYSELECT, we still choose the METHOD for selection. This is how we could specify a different method for selecting units within a stratum other than SRS.
3. Submit the PROC PRINT statement to see the sample we selected. Notice that SAS has added a variable called "samplingweight". For an SRS within strata design, the sampling weight is the inverse of the sampling fraction for the stratum from which the observation was drawn. Unlike in an SRS setting, each unit in the population often does not have the same inclusion probability. In cases where the inclusion probabilities differ, sampling weights will play an integral role in estimation.

Estimating Means and Totals from a Stratified Random Sample:

With our sample and sampling weights in hand, we can now use PROC SURVEYMEANS to obtain estimates of means and totals. There are two levels at which we can do the estimation:

1. We can estimate population means and totals.
2. We can estimate stratum means and totals.

In either case, we will need to tell SAS what the stratum sizes are. SAS expects us to place these values in a data set that pairs the stratum sizes with the correct values of our stratification variable, region.

1. Submit the DATA step for to create the statdef data set.

   SAS requires this data set to have the stratification variable and a variable called _TOTAL_ that has the stratum sizes. Now we can perform our estimation. First we will obtain estimates for the population mean number of acres of farms per county in 1992 and the population total number of acres of farms in 1992.

2. Submit the first PROC SURVEYMEANS statement.

   Here we have requested a SUM as well as the MEAN, which we have been obtaining in the past. If there is a WEIGHT statement, the SUM statement will cause SAS to produce an estimate for the total. If there is no WEIGHT statement, SUM will cause SAS to merely sum the observations, which will not provide an estimate of the total.

   The WEIGHT statement is where we specify the variable with the sampling weights for SAS to use in its calculations. A WEIGHT statement is necessary when the inclusion probabilities differ from unit to unit. If there is no WEIGHT statement, SAS will assume that all observations have equal weight.

   The STRATA statement is where we specify the stratification variable to inform SAS what the stratum groupings are. The LIST option that accompanies the STRATA statement makes SAS produce a summary table containing sampling rates, population sizes and sample sizes for each stratum. This information is useful for checking whether the strata were defined correctly in SAS.

   Even though we include a STRATA statement, SAS will not automatically provide estimates for each stratum. In order to get these estimates, we need to include a BY statement where we asks for separate analyses for each level of our stratification variable. Recall that we must sort the data by the BY variable prior to requesting analyses with a BY statement. The data were sorted earlier in the program in preparation for selecting a stratified random sample via SURVEYSELECT.

3. Submit the next PROC SURVEYMEANS statement to get the individual stratum estimates.

A special case of stratified random sampling occurs when the sampling rate is the same for all strata. In such cases, we call the sample self-weighting. In our situation, we can see by the information provided by the LIST statement that the sampling rates are very close across strata. For self-weighting samples, the population estimate for the mean is simply the sample mean of all of the observations, that is the average of stratum means.