**Project. Designing a stratified sampling plan**

**MSDS 6370**

**Objective:**

* For the student to learn how to design a stratified sampling plan
* For the student to continue learning about SAS procedures for selecting stratified samples and forming estimates using the selected samples.

**Introduction**

The course materials have covered different aspects of implementing stratified sampling using SAS. In this project, students have the opportunity to put together all the aspects of implementing a stratified sample from beginning to end in the design of a sampling plan.

**Population data**

In this project, you will use a simulated population of one month’s data for an industry[[1]](#footnote-1). You have Sales and Inventory data for the complete population of companies for one month. Your task is to design a sampling plan for a sample of size 500 to be selected for data collection each month for the next 2 years. The purpose of the data collection is to estimate the total Sales and total Inventory for the entire population each month. The quality of the estimate of total Sales has priority over the estimate of total Inventory although both are important.

The data is in the file projectData.csv. The variables in the file are shown in the table below.

|  |  |
| --- | --- |
| coID | Unique identifier |
| Sales | Current month sales for a company |
| Inventory | Current month inventory for a company |

In designing the sampling plan, the class has to consider that estimates will be formed for two variables, Sales and Inventory. The methods for determining strata and sample allocation to strata are designed to use one variable. Although the correlation coefficient for sales and inventory is high at 0.82, a design based on one probably will not be perfect for the other. It is not economical to collect two separate stratified samples, one for sales and one for inventory.

Students will be divided into groups of two. One student will work on the sales variable and one student will work on the inventory variable. Each student will create a sampling design using sales or inventory and form estimates. By the twelfth week of class, students in each group will share their results and determine if sales or inventory is the better variable on which to stratify the data. Each student will present their results to class and each group will make a joint presentation describing how it selected the stratification variable for the entire data.

Part of creating a sampling design is deciding on the number of strata and how to divide the population into the strata. One popular method for deciding how to divide the population into strata is the *cum* method. Two other methods, the Cumulative Method and Equal *WhSh* per Stratum Method, also are useful in dividing the population into strata. The paper methodsToFormStrataCutoffs.docx describes the three methods. The paper “Empirical Tests of Stratum Boundry Methods in Tax Populations,” McGrath (2003) also may be helpful since it gives some more background about the methods and compares them to each other. The Equal *WhSh* per stratum method is the most complex of the methods. You will find this method most difficult to implement.

**Project Steps:**

1. Provide descriptive statistics for your dataset.

2. Design a sampling plan using the Sales(Inventory) variable.

1. Form sampling strata for the population using the Sales(Inventory) variable.
   * A certainty stratum is required. However, the assignment includes deciding on the number of units in the certainty stratum.
   * The total number of strata must be at least 6 but no larger than 11. Since 1 stratum has to be a certainty stratum, the number of non-certainty strata must range between 5 and 10.
   * Choose an approach for dividing the population units into strata and implement it.
2. Allocate the sample of size 500 to the strata using Neyman allocation.
3. Use Proc Surveyselect to select 5 samples using the sample sizes allocated to the strata by the Neyman allocation. Submit your code and seed(s).
4. Use SAS Procedure Surveymeans and the samples selected with Neyman allocation to form the following estimates for Sales(Inventory for each sample: The population mean, the standard error of the mean, the 95% confidence interval for the mean, the population total, the standard error, and the 95% confidence interval for the total.
5. Determine how many of the 5 samples produce a 95% confidence interval that includes the population total. Examine the mean of the estimates of total and mean of the estimated standard error of the total.

3. After analyzing the data for each variable, each group will decide whether to use the sampling plan based on the variable Sales or Inventory for monthly collection of both variables for the next 2 years.

4. Each student should discuss what they have learned about how to stratify a population?

# Submitting Results

## **Basic Requirements**

Each student is expected to give a 15 minute presentation concerning their results and conclusions in Live Session 13. Each group will present results concerning which variable is best for stratification. Each student will submit written results and conclusions. Part of each student’s submission should include a discussion of their group’s findings concerning selection of the stratification variable. Each graphic or table should be clearly labeled and discussed in the text. Power point slides are acceptable for submission, but each student should submit a 3-7 page description, not including any appendices, of their work and findings. Your SAS code should be included in the appendix. Also, please include comments in your SAS code indicating the purpose of each procedure.

## **Evaluation**

In general, expect the written results and conclusion to be weighed about 80% on content and 20% on organization and presentation. Keep in mind that projects are generally much more sensitive to sincere and obvious effort, over and above actual results. Your Lab 13 grade will be based on your presentation in Live Session 13.

## **Presentation Structure**

Your presentation will include the following:

* Description of your stratification based on the Sales(Inventory) variable, how you arrived at it, and the performance of estimates of Sales(Inventory).
* The stratification chosen for the entire data and why.

It will not be possible to show all the details in your presentation. Therefore, emphasize your more interesting results and choices.

**Project Results and Conclusions**

### Section 1.Descriptive Statistics

Provide a summary of your data set that includes:

* The number of units in the population
* For the variable analyzed, the population total, standard deviation, the population mean, and a histogram.
* Include your SAS code and log in the appendix.

### Section 2. Stratification Using Sales(Inventory)

In this section, report the results for using Sales(Inventory) as the stratification variable.

2.1 Describe the process you followed in forming sampling strata for the population using the Sales(Inventory) variable:

* + Describe the number and range of values you chose for the certainty stratum and your criteria for the units in the certainty stratum.
  + Describe your process for deciding on the number of non-certainty strata. Include your calculations such as a table. If you used SAS, include your code and your log and put it in an Appendix.
  + Provide you strata and support your choices. Include a table that has a row for each stratum with the stratum number, number of units, and range of values.

2.2 Allocate the sample of size 500 to the strata using Neyman allocation. Provide a table with appropriate headings showing your calculations and the allocation of the sample to strata. If you use SAS to calculate table entries, include that SAS code in the appendix.

2.3 Use Proc Surveyselect to select 5 samples using the sample sizes allocated to the strata by the Neyman allocation. Place your SAS code, seeds, and log in an appendix.

2.4 Use SAS Procedure Surveymeans and the 5 samples selected using Neyman allocation to form the following estimates for Sales(Inventory) for each sample: the population mean, the standard error of the mean, the 95% confidence interval for the mean, the population total, the standard error for the total, and the 95% confidence interval for the total. Show the estimates for the 5 samples in a table as well as the averages and ranges of the output variables. Place your SAS code in an appendix.

2.5. Using your estimates for the mean and total and compare them to the true mean and total. Do the confidence intervals contain the true values?

2.6 Run the analysis removing the stratification. Comment on the results of removing the stratification by comparing the results with and without stratification.

#### Section 4. Conclusion

Each group should discuss which stratification variable they would choose and support their decision. The discussion should include any trade-offs and compromises that were made in forming the decision. Remember that the quality of the estimate of total Sales has priority over the estimate of total Inventory although both are important. Summarize what you learned about how to stratify a population and design a sampling plan.

Note: references are not required, but if you use material from a book, an article, or a website, you must cite the source, and use quotation marks for any paragraph (or part thereof) that is quoted word-for-word from the source(s).

1. Developed by the Complex Survey Methods and Analysis Group in the Economic Statistical Methods Division at the U.S. Census Bureau. [↑](#footnote-ref-1)