**Lab 8. Analysis of cluster samples in PROC SURVEYMEANS**

**MSDS 6370**

**Due 3/12-13/2018 based on your class time**

**Objective:**

* For the student to learn about cluster sampling.
* For the student to continue learning about SAS procedures for forming estimates using cluster samples.

**Introduction**

The topic of the reading material for Asynchronous Lecture 6 was a discussion of cluster sampling and aspects of the design and implementation of cluster samples. In this lab, we continue to study forming estimates with data collected in a cluster sample using SAS.

**Estimating the mean with data from a cluster sample using SAS**

You will learn how to use SAS PROC SURVEYMEANS to analyze data from a cluster sample. The Excel file lab8Dat.xlsx contains an encounters sheet and patients sheet with data from a sample selected for an audit of a health care provider. We used these data sets in Homework 3 where we assumed it was the whole population and selected a PPS of patients. In this Lab 8, we analyze it as a sample. The objective of the audit was to estimate the total overpayment by an insurance company to a medical provider.

Sometimes you will get data from a cluster sample as individual values and sometimes you will get it as sums over the clusters. In the audit case, the individual values are encounters and the clusters are patients.

For our dataset, a cluster design was used to select payments to include in the sample. The clusters were patients, and all the payments made on the sampled patients accounts were included, thus producing a cluster sample of payments. For this lab we assume patients are selected according to an SRS design.

Now, the way to estimate the total overpayment is to estimate the mean of the sample data and then multiply by the total observations in the population, N. The standard error of the total is the standard error of the mean multiplied by N. These values can be expressed with the formulas below:

Total = T = and stdErr(T) = *N\*stderr(*

For this Lab, you will use the cluster option in PROC SURVEYMEANS shown below:

CLUSTER *varname;*

In the case of the audit data, the *varname* = Patient ;

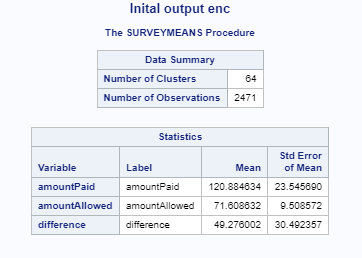
**Analysis 1**

The following SAS code will produce estimates of total overpayment from the audit sample found in the encounters sheet.

libname xl XLSX '/folders/myfolders/smuData/lab8Dat.xlsx';  
  
data encDat; set xl.enc; run;

title "Inital output enc";  
proc surveymeans data = encDat total=11426 mean;  
 var amountPaid amountAllowed difference;  
 cluster patNum;  
run;

Run this code. You should get the following:



**Analysis 2**

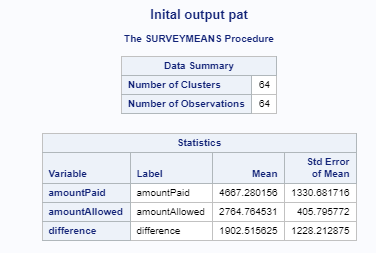
In the patients sheet the sample data set has been “rolled up” to the patient level, and all the variables reported are summed (e.g., sum of paid, allowed, and overpayment amounts). The following SAS code produces estimates of total overpayment from the audit sample found in the patients sheet.

libname xl XLSX '/folders/myfolders/smuData/lab8Dat.xlsx';  
  
data encDat; set xl.patients; run;

title "Inital output pat";  
proc surveymeans data = patDat total=296 mean;  
 var amountPaid amountAllowed difference;  
 cluster patNum;  
run;

Run this code. You should get the following:

**Results Analysis 2:**

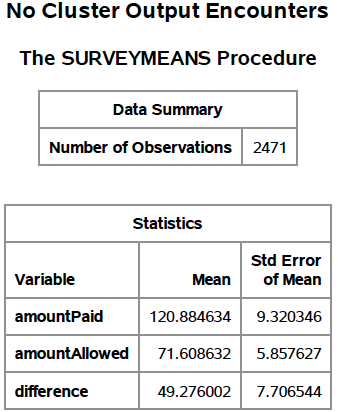
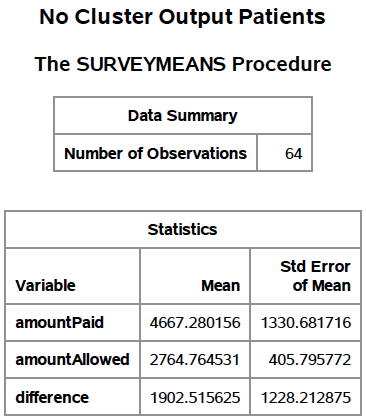


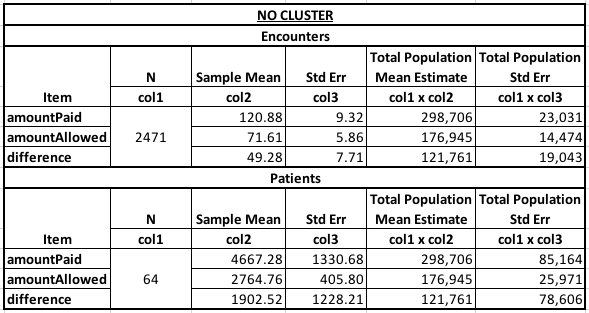
Now, the estimates of the totals of these variables and their standard errors are shown below:



**Exercise 1**

*1. Rerun Analysis 1 and 2, but remove the CLUSTER option from PROC SURVEYMEANS:*



*2. You should get a different standard error for Analysis 1 without the CLUSTER statement. Explain why the standard error is smaller. Which standard error is correct if the sample design is a simple random sample of patients?*

The smaller standard error without the CLUSTER statement has to do with the information that is left out from sampling clusters. In a simple random sample of patients the analysis with the CLUSTER statement is correct. The sampling units are patients, not encounters. Clusters of encounters grouped by patient are randomly sampled.

*3. Do you get a different standard error for Analysis 2? Why or why not?*

No, the results for Analysis 2 do not change whether the CLUSTER statement is included or not. The patients data is already aggregated at the patient level. So, using a clustered sampling design with patients as the variable to cluster with does nothing because you would get clusters of one.

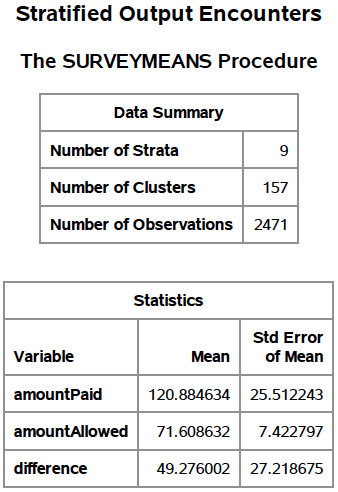
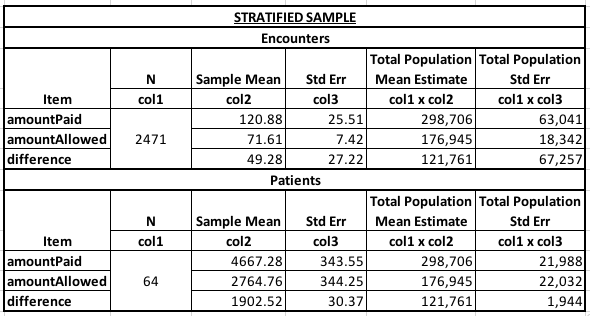
*4. What does this tell you about when you must use the CLUSTER option?*

The CLUSTER option should be used when the tradeoff between the cost of sampling outweighs the desired precision and the CLUSTER option needs to be used on a variable that is not the final sampling unit.

**Exercise 2**

*1. Redo Analysis 1, but this time, take into account the stratification. You need to add the STRATA statement and Total option. Use the stratum sizes below for the Total option:*

|  |  |
| --- | --- |
| **Encounter Strata** | **NumInStrataPop** |
| **1** | 12 |
| **2** | 215 |
| **3** | 1248 |
| **4** | 1607 |
| **5** | 58 |
| **6** | 7116 |
| **7** | 335 |
| **8** | 827 |
| **9** | 8 |
| **Total** | 11426 |

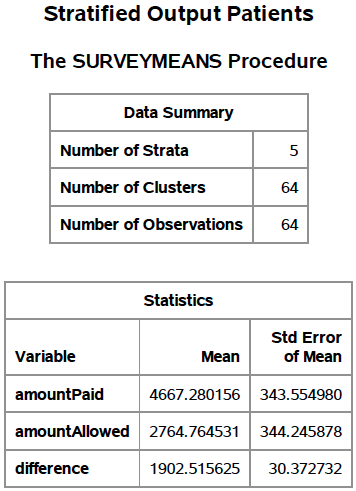
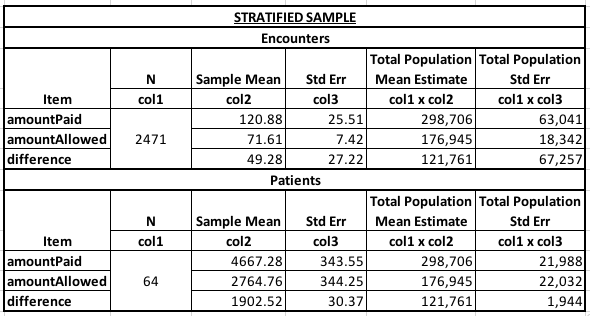
*2. Comment on the results. Did stratification help the precision?*

No, stratification did not help with precision. The standard error estimates without the CLUSTER statement yielded estimates for all three parameters with lower standard error values. If the CLUSTER statement is required, stratification gets slightly better results in terms of precision but not by a wide margin.

**Exercise 3**

*1. Repeat Analysis 2, but this time, take into account the stratification. You need to add the STRATA statement and Total option. Use the stratum sizes below for the Total option:*

|  |  |
| --- | --- |
| **Patient Strata** | **NumInStrataPop** |
| **1** | 245 |
| **2** | 23 |
| **3** | 9 |
| **4** | 14 |
| **5** | 5 |
| **Total** | 296 |

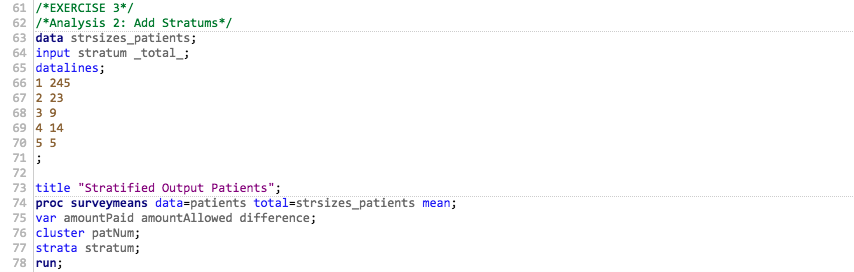
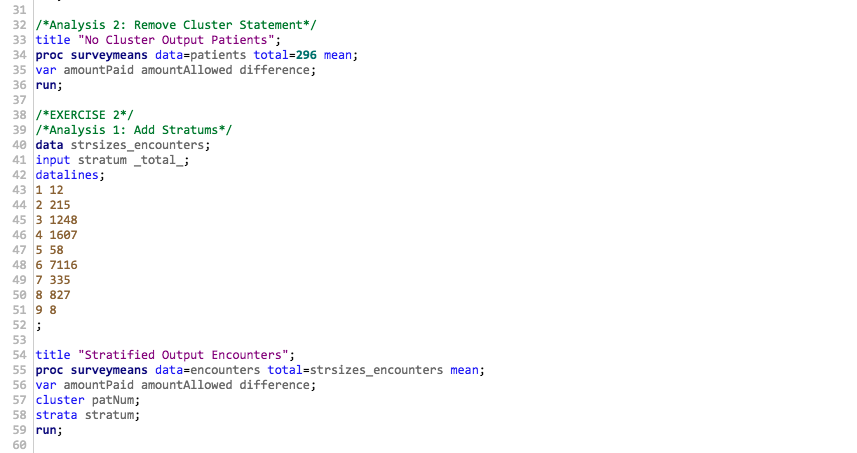
 

*2. Comment on the results. Did stratification help the precision?*

Yes, stratification vastly decreased the standard error estimates for all three parameters.

**APPENDIX:**

*LAB08\_JackNelson.sas*

****