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

**MSDS 6370**

**Objective:**

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

Today you will learn how to use SAS PROC SURVEYMEANS to analyze data from a cluster sample. The Excel files MSDS\_6370\_Lab8\_encounter.csv and MSDS\_6370\_Lab8\_patients.csv contain data from sample selected for an audit of a health care provider. We used the encounter.xls file 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, which it really is.

The objective of the audit was to estimate the total overpayment by an insurance company to a medical provider. 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. The patients were actually selected using a stratified design, but in our first analysis we will ignore the stratification, and act as though the patients were a simple random sample of all the provider’s patients.

You will learn one new feature of PROC SURVEYMEANS.

The option

CLUSTER *varname ;*

is used to specify the cluster definitions on the PROC SURVEYMEANS statement. 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 encounter.csv.

**data** treatments;

infile 'C:\MSDS 6370 SPRING2017\Live session 08\Lab8\MSDS 6370 lab8\_encounter.csv' dlm=',' firstobs=**2** ;

input patient paid allowed diff str ;

**run**;

/\**Note that the path to your data will replace this generic path to the data in this line of code*\*/

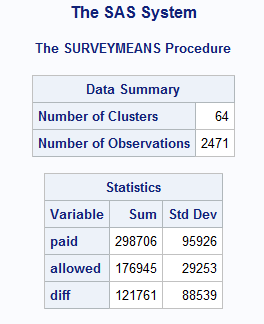
**proc** **surveymeans** data = treatments total=**11426** sum;

var paid allowed diff;

cluster Patient;

**run**;

Run this code. You should get the following:



**Analysis 2**

In the file Lab8\_patients.csv, the encounters have 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 will produce estimates of total overpayment from the audit sample found in patient.csv.

**data** patients;

infile 'C:\MSDS 6370 SPRING2017\Live session 08\Lab8\MSDS 6370 Lab 8\_patients.csv' dlm=',' firstobs=**2** ;

input patient str sumoverpaid sumallowed sumpaid ;

**run**;

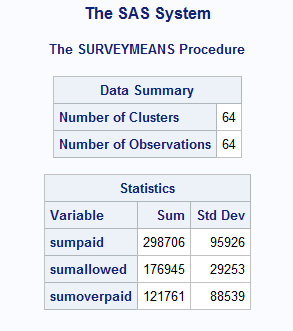
**proc** **surveymeans** data = patients total=**11426** sum;

var sumpaid sumallowed sumoverpaid;

cluster Patient;

**run**;

Run this code. Your results should be



The results are identical. Sometimes you will get data from a cluster sample as individual values and sometimes you will get it as sums over the clusters. SAS PROC SURVEYMEANS can handle it either way. But BE CAREFUL, as the following analysis will show.

**Exercise 1**

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

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Observations** | 2471 |

| **Statistics** | | |
| --- | --- | --- |
| **Variable** | **Sum** | **Std Error of Sum** |
| **paid** | 298706 | 23031 |
| **allowed** | 176945 | 14474 |
| **diff** | 121761 | 19043 |

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Observations** | 64 |

| **Statistics** | | |
| --- | --- | --- |
| **Variable** | **Sum** | **Std Error of Sum** |
| **sumpaid** | 298706 | 95926 |
| **sumallowed** | 176945 | 29253 |
| **sumoverpaid** | 121761 | 88539 |

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 (that is, if this sample had really been a simple random sample)?

The standard error is smaller because it is now using individual encounters rather than patients for estimation. Individual encounters are less variable than patients. Some patients may be unhealthy and have many more expensive encounters while other may have just a few and less expensive.

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

Analysis 2 does not change because it is already in effect clustered by being aggregated to the patient level.

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

The cluster option should be used when within cluster units are independent.

**Exercise 2**

1. Redo Analysis 1, but this time, take into account the stratification. You need to add the STRATA option. Also use the following stratum sizes.

**data** strsizes;

input str \_total\_;

datalines;

1 12

2 215

3 1248

4 1607

5 58

6 7116

7 335

8 827

9 8

;

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Strata** | 9 |
| **Number of Clusters** | 64 |
| **Number of Observations** | 2471 |

| **Statistics** | | |
| --- | --- | --- |
| **Variable** | **Sum** | **Std Error of Sum** |
| **paid** | 298706 | 79399 |
| **allowed** | 176945 | 22784 |
| **diff** | 121761 | 73184 |

**The SURVEYMEANS Procedure**

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Strata** | 9 |
| **Number of Clusters** | 64 |
| **Number of Observations** | 64 |

| **Statistics** | | |
| --- | --- | --- |
| **Variable** | **Sum** | **Std Error of Sum** |
| **sumpaid** | 298706 | 79399 |
| **sumallowed** | 176945 | 22784 |
| **sumoverpaid** | 121761 | 73184 |

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

Strata performed better than the clustered but not as well as the simple design. The strata deal with the problems introduced by the clustering but not perfectly.

**Exercise 3**

1. Repeat Analysis 2, but this time take into account the stratification. You need to add the STRATA option

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Strata** | 9 |
| **Number of Observations** | 2471 |

| **Statistics** | | |
| --- | --- | --- |
| **Variable** | **Sum** | **Std Error of Sum** |
| **paid** | 298706 | 25475 |
| **allowed** | 176945 | 16117 |
| **diff** | 121761 | 20837 |

**The SURVEYMEANS Procedure**

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Strata** | 9 |
| **Number of Observations** | 64 |

| **Statistics** | | |
| --- | --- | --- |
| **Variable** | **Sum** | **Std Error of Sum** |
| **sumpaid** | 298706 | 96871 |
| **sumallowed** | 176945 | 27005 |
| **sumoverpaid** | 121761 | 89310 |

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

Stratification did help the precision but grouping similarly situated patients together.