**Lab 4. Allocating sample to strata in stratified sampling**

**MSDS 6370**

**Objective:**

* For the student to learn about allocating sample to strata using a measure of size.
* For the student to continue learning about SAS procedures for selecting stratified samples and forming estimates using the selected samples.

**Introduction**

The topic of the reading material for Asynchronous Lecture 4 was a discussion of implementing stratified sampling using SAS. In this lab, we continue to study methods of allocating sample when using stratified sampling. In addition we consider particular approaches to sample allocation.

**Proportional allocation and Neyman allocation**

Suppose we have a population of size *N* population *yi* of size *N* with mean and standard deviation Also suppose that we have divided the population into *H* strata of size *Nh, h = 1, 2,…, H* and standard deviation .

When using proportional allocation of a sample of size *n* to the strata, the number of units drawn from stratum *h* is

When using Neyman allocation of a sample of size *n* to the strata, the number of units units drawn from stratum *h* is

However, often we do not know the standard deviation of the population of interest overall or within each stratum. In this case, we may be able to find a measure of size of each unit from an auxiliary source that we can use in place of standard deviation because size tends to be highly correlated with variance. Possibly we will have only the sum of the measure of size for some subgroups, which we could use as strata as they are or collapse into larger groupings. Using to be the sum of the measure of size for the units in stratum *h*, the allocation to stratum *h* is

**Population data**

In this assignment, you will use a simulated population of monthly retail sales data for an industry we will call Industry 2. This simulated population was created for a student paper contest where the prize was an expense-paid trip to an international conference. Usually this type of population has a few very large units that are included in samples with certainty, but we exclude these units from our population for this assignment. Our population has 373 units, a size that corresponds to a large state or a group of smaller states for Industry 2. The data is in the file MSDS\_6370\_Lab\_4.xls and includes the variables shown in the table below.

|  |  |
| --- | --- |
| UnitID | Unique identifier |
| MOS | Measure of size from administrative data |
| Strata | Sampling Strata indexed by 1 - 5 |
| Sales | 1st month sales for unit |

In this lab, we are interested in estimating the population total sales each month for the population over the next two years and want to use the same sample design in each month. To decide how to allocate a sample of size across the strata in each month for the next two years, we must explore proportional allocation and Neyman allocation. In determining the sample allocation for each, we use a measure of size from administrative data for the entire Industry 2 population that has a measure of size for each unit. Since we are in the unusual position of having sales for the entire population for the 1st month of data collection, we can see what our estimates would have been with each of the sample allocations.

**Exercise**

You may use SAS proc means or Excel to calculate the means, sums, and standard deviations required for the allocations in No. 1, 2, 3.

1.For a sample of size 50 from the population for Industry 2, use proportional allocation to determine the sample size for each of the 5 strata. Enter these in Table 1 on the Results page.

2. For a sample of size 50 from the population for Industry 2, determine the sample size for each of the 5 strata using Neyman allocation where the administrative measure of size of each unit is used in calculating standard deviation for each stratum. Enter these in Table 2 on the Results page

3. For a sample of size 50 from the population for Industry 2, determine the sample size for each of the 5 strata using the sum of the measure of size in stratum *h* in place of in a Neyman allocation. Enter these in Table 3 on the Results page.

4.

(a) Use SAS Procedure Surveyselect to select a sample of size 50 using the proportional allocation to strata as determined in No. 1 above. Output the selected units to a file named propsample.

(b) Use SAS Procedure Surveyselect to select a sample of size 50 using the Neyman allocation to strata as determined in No. 2 above. Output the selected units to a file named neysample.

(c) Use SAS Procedure Surveyselect to select a sample of size 50 using the Neyman allocation to strata as determined in No. 3 above. Output the selected units to a file named mossample.

You may use the SAS code below. Be sure to use the random seed indicated in **bold**.

proc import datafile="c:\myfiles\MSDS\_6370\_lab\_4.xls"

out = industry2;

run;

proc surveyselect data=industry2 method = srs out = propsample

sampsize = (n1,n2,n3,n4,n5) seed=**91118**;

strata stratum;

title "Proportional allocation";

proc surveyselect data=industry2 method = srs out = neysample

sampsize = (n1,n2,n3,n4,n5) seed=**91119**;

strata stratum;

title "Neyman allocation";

proc surveyselect data=industry2 method = srs out = mossample

sampsize = (n1,n2,n3,n4,n5) seed=**91120**;

strata stratum;

title "Neyman with mos for std dev allocation";

run;

4.(a) Use SAS Procedure Surveymeans and the sample selected using proportional allocation to form these estimates for the 1st month sales: the population mean, the standard error of the mean, the 95% confidence interval for the mean, the population total, the standard deviation, and the 95% confidence interval for the total. (b) Use SAS Procedure Surveymeans and the sample selected using Neyman allocation to form the same estimates in 4 (a). Enter the estimates from both samples on the Results page.

You may use the following SAS code.

data strsizes;

input stratum \_total\_;

datalines;

1 N1

2 N2

3 N3

4 N4

5 N5

;

run;

proc surveymeans data = propsample sum clsum total = strsizes

mean sum CLSUM;

var sales;

weight SamplingWeight;

strata stratum;

title "Proportional allocation";

run;

data strsizes;

input stratum \_total\_;

datalines;

1 N1

2 N2

3 N3

4 N4

5 N5

;

run;

proc surveymeans data = neysample sum clsum total = strsizes

mean sum CLSUM;

var sales;

weight SamplingWeight;

strata stratum;

title "Neyman allocation";

run;

proc surveymeans data = mossample sum clsum total = strsizes

mean sum CLSUM;

var sales;

weight SamplingWeight;

strata stratum;

title "Neyman substituting MOS allocation";

run;

**Lab 4. Results and Exercises**

**Table 1. Proportional allocation of sample of size 50 from Industry 2 population**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| stratum | Number of units | Nh/N | 50\*Nh/N | sample size |
| 1 | 141 | 0.378016 | 18.900804 | 19 |
| 2 | 100 | 0.268097 | 13.404826 | 13 |
| 3 | 63 | 0.168901 | 8.445040 | 9 |
| 4 | 38 | 0.101877 | 5.093834 | 5 |
| 5 | 31 | 0.083110 | 4.155496 | 4 |

**Table 2. Neyman allocation of sample of size 50 from Industry 2 population**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| stratum |  |  | 50\* | sample size |
| 1 | 2458416 | 0.140231 | 7.01153 | 7 |
| 2 | 3772748 | 0.215201 | 10.76007 | 11 |
| 3 | 4734638 | 0.270069 | 13.50343 | 14 |
| 4 | 3287098 | 0.187499 | 9.37497 | 9 |
| 5 | 3278338 | 0.187000 | 9.34999 | 9 |

**Table 3. Neyman allocation of sample of size 50 from Industry 2 population when substituting the sum of measure of size for product of stratum size and std. deviation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| stratum | MOSh | MOSh/total MOS | 50\* MOSh/total MOS | sample size |
| 1 | 4599050 | 0.063172 | 3.158601 | 3 |
| 2 | 10650570 | 0.146295 | 7.314749 | 7 |
| 3 | 15590567 | 0.214150 | 10.707510 | 11 |
| 4 | 17290210 | 0.237496 | 11.874815 | 12 |
| 5 | 24671620 | 0.338886 | 16.944324 | 17 |

2. With the proportional allocation in Table 1, calculate these estimates: the population mean, the standard error of the mean, the 95% confidence interval for the mean, the population total, the standard deviation, and the 95% confidence interval for the total.

180819, 213492]

3. With the Neyman allocation in Table 2, calculate these estimates: the population mean, the standard error of the mean, the 95% confidence interval for the mean, the population total, the standard deviation, and the 95% confidence interval for the total.

183970, 205451]

4. With the Neyman allocation in Table 3, calculate these estimates: the population mean, the standard error of the mean, the 95% confidence interval for the mean, the population total, the standard deviation, and the 95% confidence interval for the total.

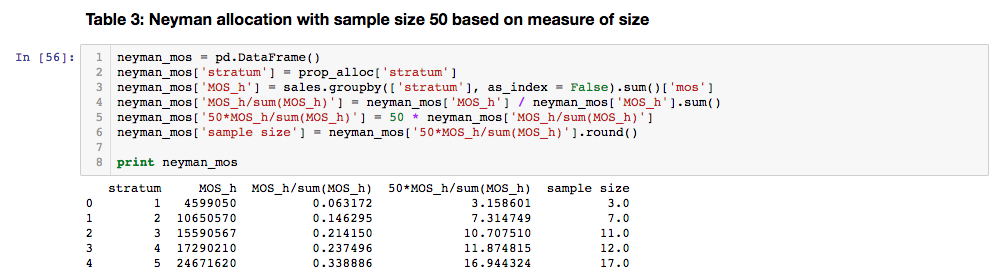
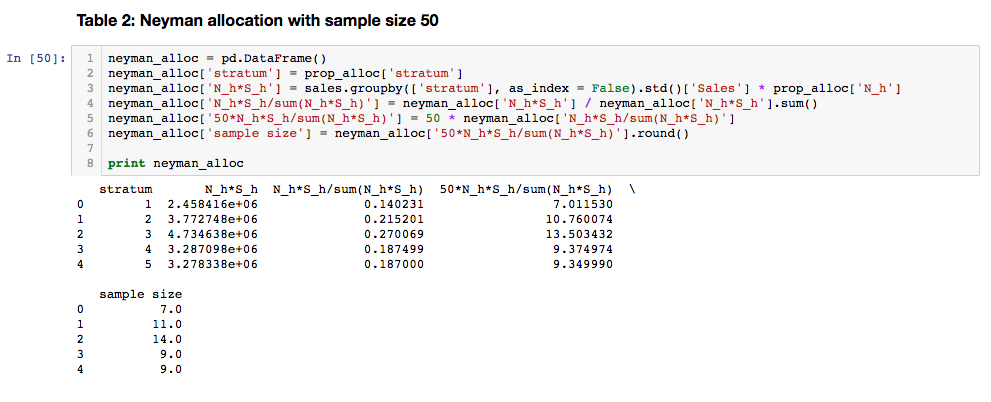
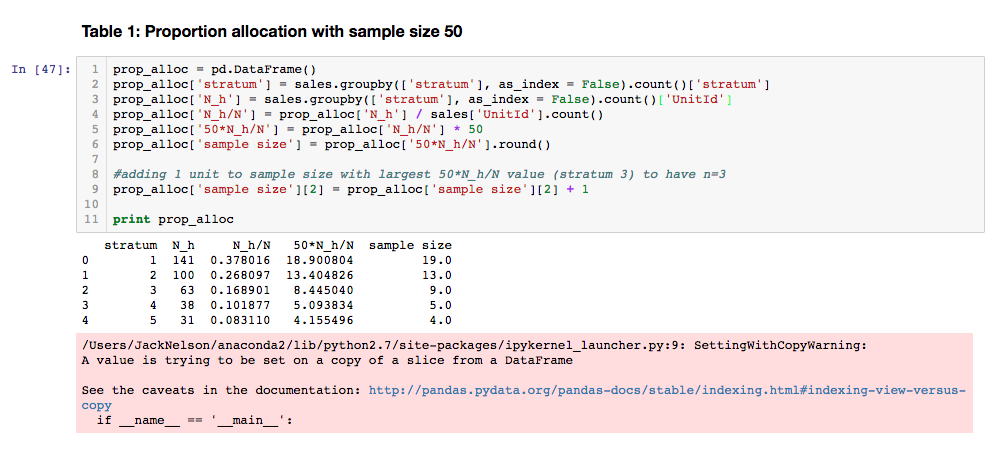
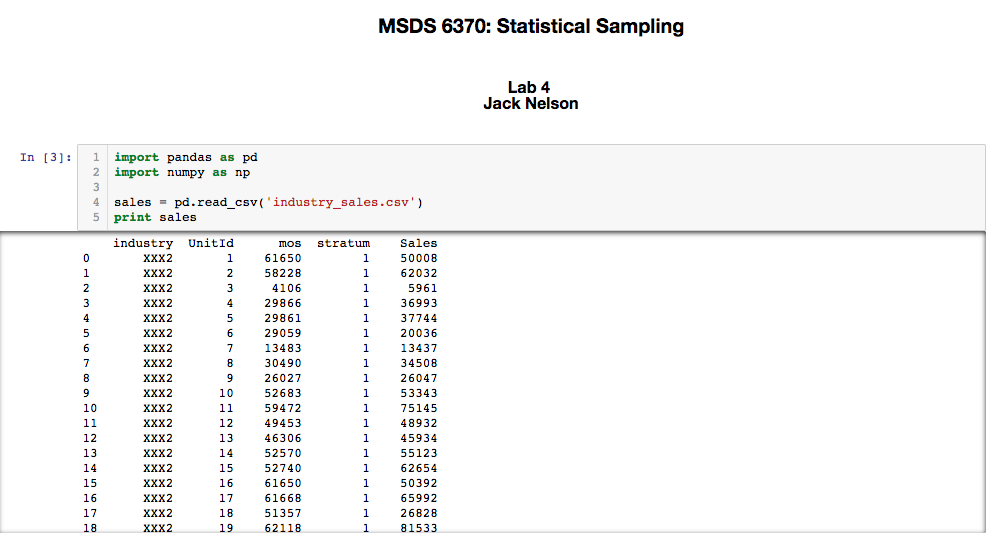
174826, 201270]

5. How do the estimates using the proportional allocation compare to the estimates from the 2 versions of Neyman allocation?

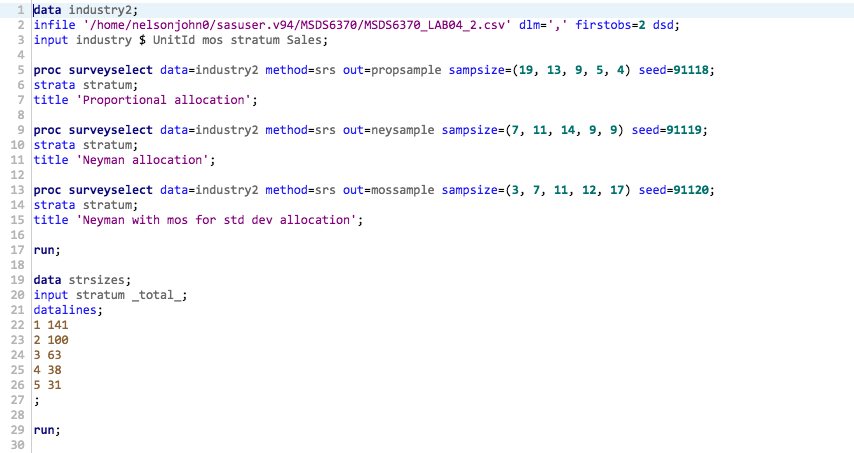
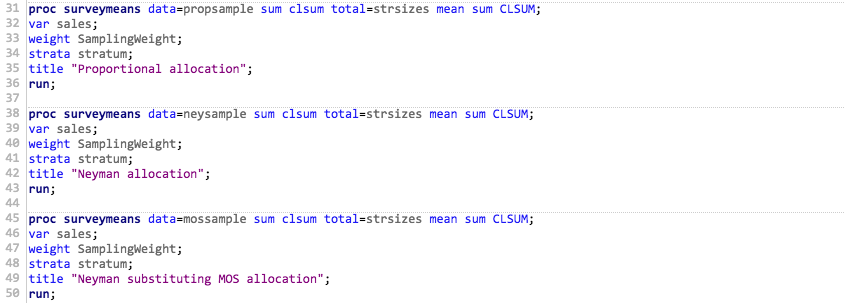
All of the mean and total estimates hover over relatively the same range however there are stark differences in the standard errors. Both Neyman allocations had considerably lower standard errors meaning there had tighter confidence widths. This allows us to narrow the range where we are 95% confident the estimates lie in that range while taking the same total sample size as proportion allocation. The normal Neyman allocation performed best which is expected, Neyman based on MOS had a slight larger standard error but still performed better than proportion allocation.

**Appendix:**

LAB04\_JackNelson.ipynb



LAB04\_JackNelson.sas

LAB04\_JackNelson.sas: OUTPUT

