**Lab 5. Estimating sample size and allocating sample to strata in stratified sampling**

**MSDS 6370**

**Objective:**

* For the student to learn about estimating the sample size to achieve a specified margin of error in the estimate of the population mean.
* For the student to lean about allocating sample to strata when estimating a proportion.

**Introduction**

The topic of the reading material for Asynchronous Lecture 5 was a discussion of estimating sample size. In this lab, we continue to study methods of allocating sample when using stratified sampling. In addition we consider sample allocations to strata when estimating a proportion.

**Stratified sample when estimating a proportion**

When estimating the proportion of a population that has a characteristic, we define a variable

*Yi* = 1 if population member i has the characteristic

0 if population member *i* does not have the characteristic

The estimate of the proportion with the characteristic from a sample of size n has the form of mean and the estimate of its variance is

The finite population correction may be ignored is the sample size is small relative to the population size, say less than 10%.

The variance of an estimator of a proportion from a stratified sample that uses SRS within strata follows the form of an estimator of the mean for a stratified design and is a weighted average of the variances of the estimators of means for the strata; i.e., take the formula

where , and

substitute for to get

where .

(Use the formula specify in the live session 05 notes)

**Exercises**

1.Suppose you have a population of N = 10,000 accounts receivable, for which you want to estimate the mean audit value. (Note: The object of the audit is to determine the audit value of each account. The “book” or recorded value will be the same as the audit value if the bookkeeper has been accurate on that account, but could be more or less than the recorded value if a mistake was made.) You would like to draw a sample of these accounts to estimate mean audit value.

Suppose you approximate the standard deviation of the audit value in the population to be $75, based on previous audits. How large should the sample size of a SRS be for the auditor to have a 95% chance that the sample mean falls with + or - $2 of the population mean?

Form the estimates of the sample size with and without the population correction, as discussed in Chapter 4 of your text *Applied Survey Sampling*.

2.Suppose that you are estimating the proportions of voters who had to wait more than 10 minutes and whose ID was rejected in Dallas County. Suppose you know there are 1.1 million voters in Dallas County. You are considering three sample designs: a srs of voters, a proportionately allocated stratified sample of voters, and a Neyman allocated stratified sample of voters. In the latter two cases, the strata are 3 geographic regions in Dallas County: stratum 1 = North Dallas, stratum 2 = central Dallas, stratum 3 = South Dallas. The sizes of these three strata are 300K, 400K, and 400K, respectively.

Suppose the proportions of voters in the three strata who had to wait more than 10 minutes is 0.20, 0.40, and 0.60, respectively. (Note that this means that the overall proportion is P = (300K\*0.20+400K\*0.40 + 400K\*0.60)/1100K = 41.8%.)

1. Find the standard error of the estimate of proportion of voters waiting more than 10 minutes for a SRS of 400 voters.
2. What are the stratum sample sizes for a proportionately allocated stratified sample of size 400?
3. What is the standard error of the estimate of proportion of voters waiting more than 10 minutes for a proportionately allocated stratified sample of 400 voters.
4. What are the stratum sample sizes for a Neyman allocated stratified sample of size 400?
5. What is the standard error of the estimate of proportion of voters waiting more than 10 minutes for a Neyman allocated stratified sample of 400 voters.

**Lab 5. Results and Exercises**

1.How large should the sample size of a SRS be for the auditor to have a 95% chance that the sample mean falls with + or - $2 of the population mean? Show your calculations.

(4\*75^2)/4 =5625

5625/(1+5625/10000) = 3600

2.When the proportions of voters in the three strata in Dallas who had to wait more than 10 minutes is 0.20, 0.40, and 0.60, respectively, what are the answers to the 5 questions below.

Set up your calculations in an Excel spreadsheet. Submit your Excel spreadsheet in addition to answering the questions (a) through (e).

(a) What is the standard error of the estimate of proportion of voters waiting more than 10 minutes for a SRS of 400 voters?

(b) What are the stratum sample sizes for a proportionately allocated stratified sample of size 400?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | N\_h | Nh/N | nh | rounded |
| 1 | 300000 | 0.272727273 | 109.0909091 | 110 |
| 2 | 400000 | 0.363636364 | 145.4545455 | 145 |
| 3 | 400000 | 0.363636364 | 145.4545455 | 145 |
| Total | 1100000 |  | 400 | 400 |

(c)What is the standard error of the estimate of proportion of voters waiting more than 10 minutes for a proportionately allocated stratified sample of 400 voters.

|  |  |
| --- | --- |
| SE prop = | 0.02345 |

(d)What are the stratum sample sizes for a Neyman allocated stratified sample of size 400?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Stratum | N\_h | p\_h | S\_h | N\_hS\_h | n\_h | n\_h rounded |
| 1 | 300000 | 0.2 | 0.4 | 120000 | 93.76494379 | 94 |
| 2 | 400000 | 0.4 | 0.489898 | 195959.2 | 153.1175281 | 153 |
| 3 | 400000 | 0.6 | 0.489898 | 195959.2 | 153.1175281 | 153 |
| Total | 1100000 |  |  | 511918.4 | 400 | 400 |

(e) What is the standard error of the estimate of proportion of voters waiting more than 10 minutes for a Neyman allocated stratified sample of 400 voters.

SE Neyman = 0.02335