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

**MSDS 6370**

**Submitted by Hari Narayan Sanadhya**

**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 will look as determining sample size and will continue our study of methods of allocating a sample to strata and estimating means. 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

*Yi* = 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 if 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 .

**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 estimate average real value of the accounts. 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 the mean real value.

Suppose you approximate the standard deviation of the real 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*.

**Solution**

To find the sample size for SRS, the formula to be used is:-

N = 10000

S = 75

I95% = MOE = 2

Sample size for SRS without finite population correction factor:

= 5402.25 = 5402

Sample size for SRS with finite population correction factor:

= 3507.33 = 3507

We get the same answer for the sample size when instead of using the above formula, we use the formula from the text which is

Sample size for SRS with finite population correction factor and correction for sampling without replacement:

The estimates of the sample size without the population correction = 5402

The estimates of the sample size with the population correction = 3507

The estimates of the sample size with the population correction and and correction for sampling without replacement = 3508

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: An 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 in the population 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% in the population.)

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 the proportion of voters waiting more than 10 minutes for a proportionately allocated stratified sample of 400 voters. State any assumptions you need to make to calculate the estimate.
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 the proportion of voters waiting more than 10 minutes for a Neyman allocated stratified sample of 400 voters.

State any assumptions you need to make to calculate estimates.

**Solution**

Excel document containing the work done in Excel



1. Standard error = = =

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Strata | N | Nh | n | nh | stratum sample sizes |
| 1 - North Dallas | 1100000 | 300000 | 400 | 109.0909 | 109 |
| 2 - Central Dallas | 1100000 | 400000 | 400 | 145.4545 | 145 |
| 3 - South Dallas | 1100000 | 400000 | 400 | 145.4545 | 146 |

1. The formula that would be used to get standard error of the estimate of proportion of voters waiting for more than 10 minutes

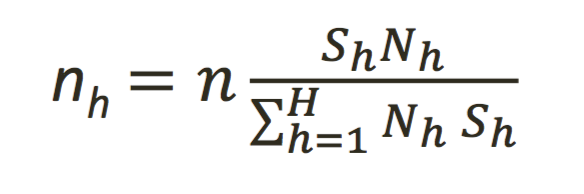
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Strata | N | Nh | n | nh | stratum sample sizes | ph | |  | | --- | |  | | StdErr |
| 1 - North Dallas | 1100000 | 300000 | 400 | 109.0909091 | 109 | 0.2 | 0.000110153 | 0.023436 |
| 2 - Central Dallas | 1100000 | 400000 | 400 | 145.4545455 | 145 | 0.4 | 0.000220306 |  |
| 3 - South Dallas | 1100000 | 400000 | 400 | 145.4545455 | 146 | 0.6 | 0.000218786 |  |

Since the stratum population size is large, we can ignore the finite population correction factor. The standard error without FPE is:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Strata | N | Nh | n | nh | stratum sample sizes | ph | |  | | --- | |  | | StdErr |
| 1 - North Dallas | 1100000 | 300000 | 400 | 109 | 109 | 0.2 | 0.000110193 | 0.023440228 |
| 2 - Central Dallas | 1100000 | 400000 | 400 | 145 | 145 | 0.4 | 0.000220386 |  |
| 3 - South Dallas | 1100000 | 400000 | 400 | 145 | 146 | 0.6 | 0.000218866 |  |

So, the standard error comes out to be 0.02344

1. Formula for Neyman allocation is



The standard deviation of the strata is not known but the proportion of voters in each strata is known. Assuming that the variance within the strata is proportional to the number of voters in the strata as it's very likely that as the number of voters increase, the wait time of the voters will increase. Hence instead of standard deviation, we can use the proportion of voters in each strata as a proxy variable and compute the stratum sample sizes as defined by Neymans allocation.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Strata | N | Nh | n | ph | phNh | |  | | --- | |  | | |  | | --- | |  | | Stratum Sample Size |
| 1 - North Dallas | 1100000 | 300000 | 400 | 0.2 | 60000 | 0.130434783 | 52.17391304 | 52 |
| 2 - Central Dallas | 1100000 | 400000 | 400 | 0.4 | 160000 | 0.347826087 | 139.1304348 | 139 |
| 3 - South Dallas | 1100000 | 400000 | 400 | 0.6 | 240000 | 0.52173913 | 208.6956522 | 209 |

1. Standard Error Calculation

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Strata | N | Nh | n | ph | phNh | |  | | --- | |  | | |  | | --- | |  | | Stratum Sample Size | |  | | --- | |  | | StdErr |
| 1 - North Dallas | 1100000 | 300000 | 400 | 0.2 | 60000 | 0.130434783 | 52.17391304 | 52 | 0.000233309 | 0.024813 |
| 2 - Central Dallas | 1100000 | 400000 | 400 | 0.4 | 160000 | 0.347826087 | 139.1304348 | 139 | 0.000229888 |  |
| 3 - South Dallas | 1100000 | 400000 | 400 | 0.6 | 240000 | 0.52173913 | 208.6956522 | 209 | 0.000152495 |  |

Since the stratum population size is large, we can ignore the finite population correction factor. The standard error without FPE is:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Strata | N | Nh | n | ph | phNh | |  | | --- | |  | | |  | | --- | |  | | Stratum Sample Size | |  | | --- | |  | | StdErr |
| 1 - North Dallas | 1100000 | 300000 | 400 | 0.2 | 60000 | 0.130434783 | 52.17391304 | 52 | 0.00023335 | 0.024817169 |
| 2 - Central Dallas | 1100000 | 400000 | 400 | 0.4 | 160000 | 0.347826087 | 139.1304348 | 139 | 0.000229968 |  |
| 3 - South Dallas | 1100000 | 400000 | 400 | 0.6 | 240000 | 0.52173913 | 208.6956522 | 209 | 0.000152575 |  |

The standard error comes out to be 0.024817.