Introduction to Sampling Theory

Lecture 12
Stratified Random Sampling

Shalabh

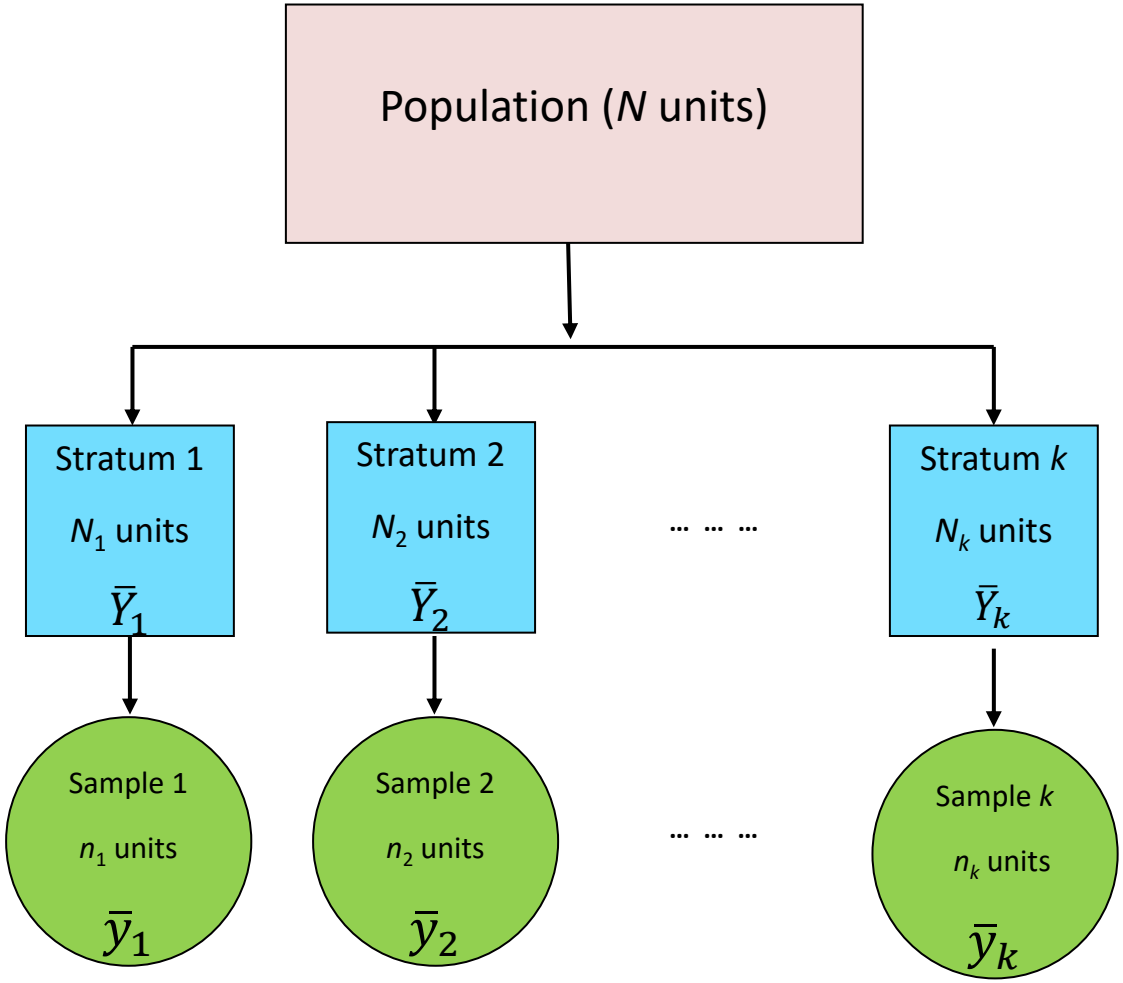
Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Procedure of Selection of a Random Sample:



$$N = \sum_{i=1}^{k} N_i$$

$$N = \sum_{i=1}^{k} N_i$$

Stratified Random Sampling: Example

- **Find the average height of the students in a school of class 1 to 12.**

- **Height varies. Students in class 1 are around 6 years old and students in class 10 are around 16 years old.**

- **Divide the students into different subpopulations or strata such as**
  - ✓ **Students of class 1, 2 and 3: Stratum 1**
  - ✓ **Students of class 4, 5 and 6: Stratum 2**
  - ✓ **Students of class 7, 8 and 9: Stratum 3**
  - ✓ **Students of class 10, 11 and 12: Stratum 4**

Stratified Random Sampling: Example

- **Draw the samples by SRS from each of the strata 1, 2, 3 and 4.**

- **All the drawn samples combined together will constitute the final stratified sample.**

Advantages of Stratified Sampling :

1. Data of known precision may be required for certain parts of the population. This can be accomplished with a more careful investigation to few strata.

<u>Example:</u> In order to know the direct impact of hike in petrol prices, the population can be divided into strata like lower income group, middle income group and higher income group. Obviously, the higher income group is more affected than the lower income group. So more careful investigation can be made in the higher income group strata.

Advantages of Stratified Sampling :

2. **Sampling problems may differ in different parts of the population.**

<u>Example:</u> To study the consumption pattern of households, the people living in houses, hotels, hospitals, prison etc. are to be treated differently.

Advantages of Stratified Sampling :

3. **Administrative convenience can be exercised in stratified sampling.**

<u>Example:</u> In taking a sample of villages from a big state, it is more administratively convenient to consider the districts as strata so that the administrative setup at district level may be used for this purpose. Such administrative convenience and the convenience in organization of field work are important aspects in national level surveys.

Advantages of Stratified Sampling :

4. Full cross-section of population can be obtained through stratified sampling. It may be possible in SRS that some large part of the population may remain unrepresented. Stratified sampling enables one to draw a sample representing different segments of the population to any desired extent. The desired degree of representation of some specified parts of population is also possible.

5. Substantial gain in the efficiency is achieved if the strata are formed intelligently.

Advantages of Stratified Sampling :

6. In case of skewed population, use of stratification is of importance since larger weight may have to be given for the few extremely large units which in turn reduces the sampling variability.

7. When estimates are required not only for the population but also for the subpopulations, then stratified sampling is helpful.

8. When the sampling frame for subpopulations is more easily available than the sampling frame for whole population, then the stratified sampling is helpful.

Advantages of Stratified Sampling :

9. If population is large, then it is convenient to sample separately from the strata rather than the entire population.

10. The population mean or population total can be estimated with higher precision by suitably providing the weights to the estimates obtained from each stratum.

Stratified Random Sampling:

We use the following symbols and notations:

N **: Population size**

k **: Number of strata**

N_i **: Number of sampling units in** i^{th} **strata**

$$N = \sum_{i=1}^{k} N_i$$ **Total population size**

n_i **: Numbers of sampling units to be drawn from** i^{th} **stratum.**

$$n = \sum_{i=1}^{k} n_i$$ **: Total sample size**

Stratified Random Sampling:

Let

Y **: characteristic under study,**

y_{ij} **: value of** j^{th} **unit in** i^{th} **stratum** $j = 1,2,\ldots,n_i$, $i = 1,2,\ldots,k$,

$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$ **: population mean of** i^{th} **stratum,** $j = 1,2,\ldots,n_i, i = 1,2,\ldots,k,$

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ **: sample mean of units from** i^{th} **stratum**

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{Y}_i = \sum_{i=1}^{k} w_i\bar{Y}_i$ **: population mean where** $w_i = \frac{N_i}{N}$.

Stratified Random Sampling:

There are *k* independent samples drawn through SRS from each of the stratum.

There will be *k* estimators of parameter but the ultimate goal is to have a single estimator.

How to combine the different sample information together into one estimator which is good enough to provide the information about the parameter.

Estimation of Population Mean:

In case of stratified sampling, the population mean is defined as the weighted arithmetic mean of stratum means where the weights are provided in terms of strata sizes.

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i,$$

Find the sample mean of the units drawn from each startum.

Find their weighted mean, called as <u>stratified mean</u>.

Use stratified mean as to estimate the population mean as

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i.$$

Unbiased Estimator of Population Mean:

Since the sample in each stratum is drawn by SRS, so

$$E(\bar{y}_i) = \bar{Y}_i,$$

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i.$$

$$\begin{aligned} E(\bar{y}_{st}) &= \frac{1}{N}\sum_{i=1}^{k} N_i E(\bar{y}_i) \\ &= \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i \\ &= \bar{Y}. \end{aligned}$$

Biased Estimator of Population Mean:

Since the sample in each stratum is drawn by SRS, so

$$E(\bar{y}_i) = \bar{Y}_i,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i$$

$$\begin{aligned} E(\bar{y}) &= \frac{1}{n}\sum_{i=1}^{k} n_i E(\bar{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{k} n_i \bar{Y}_i \\ &\neq \frac{1}{n}\sum_{i=1}^{k} n_i \bar{Y}_i \\ &\neq \bar{Y} \end{aligned}$$

Variance of Stratum Mean:

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \, Var(\bar{y}_i) + \sum_{i(\neq j)=1}^{k} \sum_{j=1}^{n_i} w_i \, w_j \, Cov(\bar{y}_i, \bar{y}_j)$$

Since all the samples have been drawn independently from each of the strata by SRSWOR, so

$$Cov(\bar{y}_i, \bar{y}_j) = 0, \quad i \neq j$$

$$Var(\bar{y}_i) = \frac{N_i - n_i}{N_i n_i} S_i^2$$

where

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2.$$

Variance of Stratum Mean:

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \, Var(\bar{y}_i) + \sum_{i(\neq j)=1}^{k} \sum_{j=1}^{n_i} w_i \, w_j \, Cov(\bar{y}_i, \bar{y}_j)$$

Thus

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \frac{N_i - n_i}{N_i n_i} S_i^2$$

$$= \sum_{i=1}^{k} w_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_i^2}{n_i}.$$

How to Construct Strata:

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_i^2}{n_i}.$$

Variance is small when S_i^2 is small.

This suggests how to construct the strata.

If S_i^2 is small for all i = 1,2,...,k, then variance will also be small.

That is why the strata are constructed such that they are within homogeneous, i.e., S_i^2 is small and among heterogeneous.