06

Simple Random and Stratified Sampling

Mini K. G.

Simple Random Sampling

Simple random sampling is a method of selecting a sample from a finite population in such a way that every unit of the population is given an equal chance of being selected. In practice, you can draw a simple random sample unit by unit through the following steps:

- Define the population
- Make a list of all the units in the population and number them from 1 to N.
- Decide the size of the sample, or the number of units to be included in the sample.
- Use either the 'lottery method' or 'random number tables' to pick the units to be included in the sample.

For example, you may use the lottery method to draw a random sample by using a set of 'N' tickets, with numbers ' 1 to N' if there are 'N' units in the population. After shuffling the tickets thoroughly, the sample of a required size, say n, is selected by picking the required n number of tickets. The units which have the serial numbers occurring on these tickets will be considered selected. The assumption underlying this method is that the tickets are shuffled so that the population can be regarded as arranged randomly. When the size of population is large, this procedure of numbering units on tickets and selecting one after reshuffling becomes cumbersome. Human bias and prejudice may creep in this method. Therefore, this method is generally discouraged.

The best method of drawing a simple random sample is to use a table of random numbers. These random number tables have been prepared by Fisher and Yates (1967). After assigning consecutive numbers to the units of population, the researcher starts at any point on the table of random numbers and reads the consecutive numbers in any direction horizontally, vertically or diagonally. If the read out number corresponds with the one written on a unit card, then that unit is chosen for the sample.

Simple random sampling can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. There are two methods of simple random sampling namely simple random sampling with replacement (SRSWR) and simple random sampling without replacement

(SRSWOR). Sampling with replacement means that each unit selected in the sample is returned to the population before the next is drawn.

In SRSWR, one unit of element is randomly selected from population is the first sampled unit. Then the sampled unit is replaced in the population. The second sample is drawn with equal probability. The procedure is repeated until the requisite sample units *n* are drawn. The probability of selection of an element remains unchanged after each draw. The same units could be selected more than once. Let N denote the population size and n is sample size. As the population size remains the same after each draw, not only the probability of each unit being selected in the sample is $\frac{1}{N}$ at each draw, it remains same even when included in the sample more than once. In case of SRSWR, the number of all possible samples is N^n. The probability of drawing any of these N^n is $\frac{1}{N^n}$. For example, if there is three landing centres in a Tehsil, denoted by A, B, C (The population size N= 3). Then 3^2 = 9 possible samples of size 2 can be drawn through SRSWR. They are (A, A) (A B) (A C) (B A) (B B) (B C) (C A) (C B) (C C). Each sample can be selected with equal probability of $\frac{1}{9}$.

In SRSWOR, unlike SRSWR, once an element is selected as a sample unit, will not be replaced in the population pool. The selected sample units are distinct. SRSWOR is a method of selecting n units out of N such that every one of the ${}_N C_n$ distinct samples has an equal change of being drawn. The simple random samples are drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw, the process must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. With reference to the same example above in SRSWR, the possible samples with SRSWOR without giving any importance to the ordering of the units are (A B) (A C) (B C). The total number of samples is ${}_3 C_2 = 3$. Any of these samples have equal probability of selection and the probability of selection of each of the samples is $\frac{1}{3}$. So, in SRSWOR, the probability of selection of each unit at any draw is $\frac{1}{N}$ and the probability of inclusion of any unit in the sample is $\frac{n}{N}$.

Estimating the Population Mean

Let Y be the character of interest and $Y_1, Y_2, \ldots, Y_i, \ldots, Y_N$ be the values of the character on N units of the population. Let, $y_1, y_2, \ldots, y_i, \ldots, y_n$ be the sample of size n selected by SRSWOR.

The estimator for the population mean is given by

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

An unbiased estimator of variance of the population mean is given by

$$\hat{V}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \text{ , where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Bound on the error of estimation (B)

$$B_y = t\,SE(\hat{\bar{Y}}) = t\sqrt{\hat{V}(\hat{\bar{Y}})}$$

Where *t* is the Student's t value for *n-1* degrees of freedom at the $1 - \alpha/2$ level of significance. For $\alpha = 0.05$, t = 1.96., Confidence Interval is given by C.I = $\hat{\bar{Y}} \pm B_y$

Estimating the Population Total

The estimator for the population total is given by

$$\hat{Y} = N\hat{\bar{Y}} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

An unbiased estimator of the variance of the population total is given by

$$V(\hat{Y}) = N^2 V\left(\hat{\bar{Y}}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

Where s^2 is an unbiased estimator of the population mean square S^2, $S^2 = \frac{1}{N-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Advantages of Simple Random Sampling

One of the best things about simple random sampling is the ease of assembling the sample. It is also considered as a fair way of selecting a sample from a given population since every member is given equal opportunities of being selected. Another key feature of simple random sampling is that the sample will be a representative of the population. If the sample is not representative of the population, the random variation called sampling error will be large. An unbiased random selection and a representative sample are important in drawing conclusions from the results of a study.

Disadvantages of Simple Random Sampling

One of the most obvious limitations of simple random sampling method is its need of a complete list of all the members of the population. However, from a practical point of view, a list of all the units of a population is not possible to obtain for large populations. Even if it is possible, it may involve a very high cost which a researcher or an organisation may not be able to afford. Therefore, simple random sampling is difficult to realize. Also, in case of a highly heterogeneous population, a simple random sample may not necessarily represent the characteristics of the total population, even though all selected units participate in the investigation. In those cases, it is wiser to use other sampling techniques.

Stratified Random Sampling

When the population is heterogeneous, the whole population can be divided into sub-populations, called strata, to increase the precision of the estimates. In stratified sampling the population of N units is first divided into disjoint groups of N_1, N_2,.., N_h,.., N_L units, respectively. These subgroups, called strata, together they compromise the whole population, so that $N_1 + N_2 + ... + N_h + .. + N_L = N$. The strata should not overlap and each stratum should be sampled following some sampling design. The strata are sampled separately and the estimates from each stratum combined into one estimate for the whole population. If a simple random sample selection scheme is used in each stratum then the corresponding sample is called a stratified random sample.

Reasons for stratification

- To obtain estimates of known precision for certain subdivisions of the population by treating each subdivision as a stratum. Since sampling is done independently in each stratum, separate stratum estimates and their precision can be obtained by treating each stratum as a ``population'' in its own right. For example, in fishery surveys estimates may be required by state, district, month, landing centre, craft, species etc.
- For administrative convenience; for example stratification can provide survey organization to control the distribution of fieldwork among its regional offices.

- Sometimes different parts of the population may call for different sampling procedures.
- Stratification may often produce a gain in precision of the estimates of characteristics of the whole population. The amount in the gain depends on the type of stratification. If the population is heterogeneous and if it can be divided, using prior information about the population, into subpopulations (strata), each of which is internally homogeneous. If each stratum is homogeneous, that is characteristic under consideration vary little from one unit to another, a precise estimate (an estimate with smaller variance) of any stratum parameter can be obtained from a small sample in that stratum. These estimates can then be combined to obtain a precise estimate for the whole population.

Notations

The suffix h (h =1,2,...,L) denotes the stratum and i the unit within the stratum.

N_h - Total number of population units in stratum h.

n_h - Total number of sample units in stratum h.

$w_h = N_h / N$ - The h^{th} stratum weight.

Y_{hi} - Value of the characteristic for the i^{th} unit in stratum h.

$Y_h = \sum_{i=1}^{N_h} Y_{hi}$ - Population total of Y – values for units belonging to stratum

$\overline{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi} = \frac{Y_h}{N_h}$ - Population mean of Y-values for units belonging to stratum h.

$\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}\left(Y_{hi} - \overline{Y}_h\right)^2$ -Population variance of Y-values for units belonging to stratum h.

$\overline{Y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi}}{\sum_{h=1}^{L} N_h} = \sum_{h=1}^{L} W_h \overline{Y}_h$ - Population mean of Y-values.

$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} = \frac{y_h}{n_h}$ - Sample mean of Y-values for units belonging to stratum h.

$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$ - Sample variance of Y-values for units belonging to stratum h.

$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} Y_{hi}}{\sum_{h=1}^{L} n_h}$ - Sample mean of Y-values.

Estimation of Population Mean and Variance

An unbiased estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}} = \sum_{h=1}^{L} W_h \bar{y}_h$$

Also, since sampling is done independently within each stratum

$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1}\sigma_h^2$$

Note that $V\left(\hat{\bar{Y}}\right)$ can not be computed since it involves Y-values for all the units in the population. However, based on Y-values for the sampled units we can estimate $V\left(\hat{\bar{Y}}\right)$ by using the following formula.

$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h n_h} s_h^2$ which estimates $V\left(\hat{\bar{Y}}\right)$ unbiasedly.

Allocation in Stratified Random Sampling

In planning a study, requiring stratification of the population, an important consideration is how to allocate a total sample size n among the L identified strata. There are three types of allocation.

Allocation of a Sample to Strata

1. Equal: If the strata are presumed to be of roughly equal size, and there is no additional information regarding the variability or distribution of the response in the strata, equal allocation to the strata is probably the best choice. $n_h = n/L$

2. Proportional: If the strata differ in size, allocation of sample sizes to strata might be performed proportional to these stratum sizes. $n_h = \left(\frac{N_h}{N}\right) n$

3. Optimum (Neyman): The allocation which minimizes the variance of the estimator of the mean (and total) is given by $n_h = \frac{nN_h\sigma_h}{\sum_{h=1}^{L} N_h\sigma_h}$

Whenever the strata are heterogeneous among themselves and the variance of each stratum is small, the sampling variance of the mean or total value estimators obtained by stratified sampling will always be smaller than the simple random sampling. The relative sizes of strata must be known to obtain the full benefits of the stratification technique. Each stratum should be internally homogeneous. If information about heterogeneity is not available then consider all strata equally variable. A short stratified pilot survey can sometimes provide useful information about internal dispersion within strata. A small sized sample could be taken from a stratum if the variability among their units is small. A larger sample from a stratum should be taken if the stratum is larger, the stratum is more heterogeneous and the cost of sampling the stratum is low.

References

Cochran, W. G. 1977. *Sampling Techniques*, third edition. John Wiley & Sons, Inc., New York, 428pp.

Lohr, S. L. 1999. *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA, 494pp.

Sukhatme, P.V. and B.V. Sukhatme. 1970. *Sampling Theory of Surveys with Applications*. Iowa State University Press, Ames, IA. 452pp.

Thompson, M. 1997. *Theory of Sample Surveys*. Chapman & Hall, 312pp.

06

Simple Random and Stratified Sampling

Mini K. G.

Simple Random Sampling

Simple random sampling is a method of selecting a sample from a finite population in such a way that every unit of the population is given an equal chance of being selected. In practice, you can draw a simple random sample unit by unit through the following steps:

- Define the population
- Make a list of all the units in the population and number them from 1 to N.
- Decide the size of the sample, or the number of units to be included in the sample.
- Use either the 'lottery method' or 'random number tables' to pick the units to be included in the sample.

For example, you may use the lottery method to draw a random sample by using a set of 'N' tickets, with numbers ' 1 to N' if there are 'N' units in the population. After shuffling the tickets thoroughly, the sample of a required size, say n, is selected by picking the required n number of tickets. The units which have the serial numbers occurring on these tickets will be considered selected. The assumption underlying this method is that the tickets are shuffled so that the population can be regarded as arranged randomly. When the size of population is large, this procedure of numbering units on tickets and selecting one after reshuffling becomes cumbersome. Human bias and prejudice may creep in this method. Therefore, this method is generally discouraged.

The best method of drawing a simple random sample is to use a table of random numbers. These random number tables have been prepared by Fisher and Yates (1967). After assigning consecutive numbers to the units of population, the researcher starts at any point on the table of random numbers and reads the consecutive numbers in any direction horizontally, vertically or diagonally. If the read out number corresponds with the one written on a unit card, then that unit is chosen for the sample.

Simple random sampling can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. There are two methods of simple random sampling namely simple random sampling with replacement (SRSWR) and simple random sampling without replacement

(SRSWOR). Sampling with replacement means that each unit selected in the sample is returned to the population before the next is drawn.

In SRSWR, one unit of element is randomly selected from population is the first sampled unit. Then the sampled unit is replaced in the population. The second sample is drawn with equal probability. The procedure is repeated until the requisite sample units *n* are drawn. The probability of selection of an element remains unchanged after each draw. The same units could be selected more than once. Let N denote the population size and n is sample size. As the population size remains the same after each draw, not only the probability of each unit being selected in the sample is $\frac{1}{N}$ at each draw, it remains same even when included in the sample more than once. In case of SRSWR, the number of all possible samples is N^n. The probability of drawing any of these N^n is $\frac{1}{N^n}$. For example, if there is three landing centres in a Tehsil, denoted by A, B, C (The population size N= 3). Then 3^2 = 9 possible samples of size 2 can be drawn through SRSWR. They are (A, A) (A B) (A C) (B A) (B B) (B C) (C A) (C B) (C C). Each sample can be selected with equal probability of $\frac{1}{9}$.

In SRSWOR, unlike SRSWR, once an element is selected as a sample unit, will not be replaced in the population pool. The selected sample units are distinct. SRSWOR is a method of selecting n units out of N such that every one of the ${}_N C_n$ distinct samples has an equal change of being drawn. The simple random samples are drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw, the process must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. With reference to the same example above in SRSWR, the possible samples with SRSWOR without giving any importance to the ordering of the units are (A B) (A C) (B C). The total number of samples is ${}_3 C_2 = 3$. Any of these samples have equal probability of selection and the probability of selection of each of the samples is $\frac{1}{3}$. So, in SRSWOR, the probability of selection of each unit at any draw is $\frac{1}{N}$ and the probability of inclusion of any unit in the sample is $\frac{n}{N}$.

Estimating the Population Mean

Let Y be the character of interest and $Y_1, Y_2, \ldots, Y_i, \ldots, Y_N$ be the values of the character on N units of the population. Let, $y_1, y_2, \ldots, y_i, \ldots, y_n$ be the sample of size n selected by SRSWOR.

The estimator for the population mean is given by

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

An unbiased estimator of variance of the population mean is given by

$$\hat{V}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \text{ , where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Bound on the error of estimation (B)

$$B_y = t\,SE(\hat{\bar{Y}}) = t\sqrt{\hat{V}(\hat{\bar{Y}})}$$

Where *t* is the Student's t value for *n-1* degrees of freedom at the $1 - \alpha/2$ level of significance. For $\alpha = 0.05$, t = 1.96., Confidence Interval is given by C.I = $\hat{\bar{Y}} \pm B_y$

Estimating the Population Total

The estimator for the population total is given by

$$\hat{Y} = N\hat{\bar{Y}} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

An unbiased estimator of the variance of the population total is given by

$$V(\hat{Y}) = N^2 V\left(\hat{\bar{Y}}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

Where s^2 is an unbiased estimator of the population mean square S^2, $S^2 = \frac{1}{N-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Advantages of Simple Random Sampling

One of the best things about simple random sampling is the ease of assembling the sample. It is also considered as a fair way of selecting a sample from a given population since every member is given equal opportunities of being selected. Another key feature of simple random sampling is that the sample will be a representative of the population. If the sample is not representative of the population, the random variation called sampling error will be large. An unbiased random selection and a representative sample are important in drawing conclusions from the results of a study.

Disadvantages of Simple Random Sampling

One of the most obvious limitations of simple random sampling method is its need of a complete list of all the members of the population. However, from a practical point of view, a list of all the units of a population is not possible to obtain for large populations. Even if it is possible, it may involve a very high cost which a researcher or an organisation may not be able to afford. Therefore, simple random sampling is difficult to realize. Also, in case of a highly heterogeneous population, a simple random sample may not necessarily represent the characteristics of the total population, even though all selected units participate in the investigation. In those cases, it is wiser to use other sampling techniques.

Stratified Random Sampling

When the population is heterogeneous, the whole population can be divided into sub-populations, called strata, to increase the precision of the estimates. In stratified sampling the population of N units is first divided into disjoint groups of N_1, N_2,.., N_h,.., N_L units, respectively. These subgroups, called strata, together they compromise the whole population, so that $N_1 + N_2 + ... + N_h + .. + N_L = N$. The strata should not overlap and each stratum should be sampled following some sampling design. The strata are sampled separately and the estimates from each stratum combined into one estimate for the whole population. If a simple random sample selection scheme is used in each stratum then the corresponding sample is called a stratified random sample.

Reasons for stratification

- To obtain estimates of known precision for certain subdivisions of the population by treating each subdivision as a stratum. Since sampling is done independently in each stratum, separate stratum estimates and their precision can be obtained by treating each stratum as a ``population'' in its own right. For example, in fishery surveys estimates may be required by state, district, month, landing centre, craft, species etc.
- For administrative convenience; for example stratification can provide survey organization to control the distribution of fieldwork among its regional offices.

- Sometimes different parts of the population may call for different sampling procedures.
- Stratification may often produce a gain in precision of the estimates of characteristics of the whole population. The amount in the gain depends on the type of stratification. If the population is heterogeneous and if it can be divided, using prior information about the population, into subpopulations (strata), each of which is internally homogeneous. If each stratum is homogeneous, that is characteristic under consideration vary little from one unit to another, a precise estimate (an estimate with smaller variance) of any stratum parameter can be obtained from a small sample in that stratum. These estimates can then be combined to obtain a precise estimate for the whole population.

Notations

The suffix h (h =1,2,...,L) denotes the stratum and i the unit within the stratum.

N_h - Total number of population units in stratum h.

n_h - Total number of sample units in stratum h.

$w_h = N_h / N$ - The h^{th} stratum weight.

Y_{hi} - Value of the characteristic for the i^{th} unit in stratum h.

$Y_h = \sum_{i=1}^{N_h} Y_{hi}$ - Population total of Y – values for units belonging to stratum

$\overline{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi} = \frac{Y_h}{N_h}$ - Population mean of Y-values for units belonging to stratum h.

$\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}\left(Y_{hi} - \overline{Y}_h\right)^2$ -Population variance of Y-values for units belonging to stratum h.

$\overline{Y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi}}{\sum_{h=1}^{L} N_h} = \sum_{h=1}^{L} W_h \overline{Y}_h$ - Population mean of Y-values.

$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} = \frac{y_h}{n_h}$ - Sample mean of Y-values for units belonging to stratum h.

$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$ - Sample variance of Y-values for units belonging to stratum h.

$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} Y_{hi}}{\sum_{h=1}^{L} n_h}$ - Sample mean of Y-values.

Estimation of Population Mean and Variance

An unbiased estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}} = \sum_{h=1}^{L} W_h \bar{y}_h$$

Also, since sampling is done independently within each stratum

$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1}\sigma_h^2$$

Note that $V\left(\hat{\bar{Y}}\right)$ can not be computed since it involves Y-values for all the units in the population. However, based on Y-values for the sampled units we can estimate $V\left(\hat{\bar{Y}}\right)$ by using the following formula.

$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h n_h} s_h^2$ which estimates $V\left(\hat{\bar{Y}}\right)$ unbiasedly.

Allocation in Stratified Random Sampling

In planning a study, requiring stratification of the population, an important consideration is how to allocate a total sample size n among the L identified strata. There are three types of allocation.

Allocation of a Sample to Strata

1. Equal: If the strata are presumed to be of roughly equal size, and there is no additional information regarding the variability or distribution of the response in the strata, equal allocation to the strata is probably the best choice. $n_h = n/L$

2. Proportional: If the strata differ in size, allocation of sample sizes to strata might be performed proportional to these stratum sizes. $n_h = \left(\frac{N_h}{N}\right) n$

3. Optimum (Neyman): The allocation which minimizes the variance of the estimator of the mean (and total) is given by $n_h = \frac{nN_h\sigma_h}{\sum_{h=1}^{L} N_h\sigma_h}$

Whenever the strata are heterogeneous among themselves and the variance of each stratum is small, the sampling variance of the mean or total value estimators obtained by stratified sampling will always be smaller than the simple random sampling. The relative sizes of strata must be known to obtain the full benefits of the stratification technique. Each stratum should be internally homogeneous. If information about heterogeneity is not available then consider all strata equally variable. A short stratified pilot survey can sometimes provide useful information about internal dispersion within strata. A small sized sample could be taken from a stratum if the variability among their units is small. A larger sample from a stratum should be taken if the stratum is larger, the stratum is more heterogeneous and the cost of sampling the stratum is low.

References

Cochran, W. G. 1977. *Sampling Techniques*, third edition. John Wiley & Sons, Inc., New York, 428pp.

Lohr, S. L. 1999. *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA, 494pp.

Sukhatme, P.V. and B.V. Sukhatme. 1970. *Sampling Theory of Surveys with Applications*. Iowa State University Press, Ames, IA. 452pp.

Thompson, M. 1997. *Theory of Sample Surveys*. Chapman & Hall, 312pp.

06

Simple Random and Stratified Sampling

Mini K. G.

Simple Random Sampling

Simple random sampling is a method of selecting a sample from a finite population in such a way that every unit of the population is given an equal chance of being selected. In practice, you can draw a simple random sample unit by unit through the following steps:

- Define the population
- Make a list of all the units in the population and number them from 1 to N.
- Decide the size of the sample, or the number of units to be included in the sample.
- Use either the 'lottery method' or 'random number tables' to pick the units to be included in the sample.

For example, you may use the lottery method to draw a random sample by using a set of 'N' tickets, with numbers ' 1 to N' if there are 'N' units in the population. After shuffling the tickets thoroughly, the sample of a required size, say n, is selected by picking the required n number of tickets. The units which have the serial numbers occurring on these tickets will be considered selected. The assumption underlying this method is that the tickets are shuffled so that the population can be regarded as arranged randomly. When the size of population is large, this procedure of numbering units on tickets and selecting one after reshuffling becomes cumbersome. Human bias and prejudice may creep in this method. Therefore, this method is generally discouraged.

The best method of drawing a simple random sample is to use a table of random numbers. These random number tables have been prepared by Fisher and Yates (1967). After assigning consecutive numbers to the units of population, the researcher starts at any point on the table of random numbers and reads the consecutive numbers in any direction horizontally, vertically or diagonally. If the read out number corresponds with the one written on a unit card, then that unit is chosen for the sample.

Simple random sampling can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. There are two methods of simple random sampling namely simple random sampling with replacement (SRSWR) and simple random sampling without replacement

(SRSWOR). Sampling with replacement means that each unit selected in the sample is returned to the population before the next is drawn.

In SRSWR, one unit of element is randomly selected from population is the first sampled unit. Then the sampled unit is replaced in the population. The second sample is drawn with equal probability. The procedure is repeated until the requisite sample units *n* are drawn. The probability of selection of an element remains unchanged after each draw. The same units could be selected more than once. Let N denote the population size and n is sample size. As the population size remains the same after each draw, not only the probability of each unit being selected in the sample is $\frac{1}{N}$ at each draw, it remains same even when included in the sample more than once. In case of SRSWR, the number of all possible samples is N^n. The probability of drawing any of these N^n is $\frac{1}{N^n}$. For example, if there is three landing centres in a Tehsil, denoted by A, B, C (The population size N= 3). Then 3^2 = 9 possible samples of size 2 can be drawn through SRSWR. They are (A, A) (A B) (A C) (B A) (B B) (B C) (C A) (C B) (C C). Each sample can be selected with equal probability of $\frac{1}{9}$.

In SRSWOR, unlike SRSWR, once an element is selected as a sample unit, will not be replaced in the population pool. The selected sample units are distinct. SRSWOR is a method of selecting n units out of N such that every one of the ${}_N C_n$ distinct samples has an equal change of being drawn. The simple random samples are drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw, the process must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. With reference to the same example above in SRSWR, the possible samples with SRSWOR without giving any importance to the ordering of the units are (A B) (A C) (B C). The total number of samples is ${}_3 C_2 = 3$. Any of these samples have equal probability of selection and the probability of selection of each of the samples is $\frac{1}{3}$. So, in SRSWOR, the probability of selection of each unit at any draw is $\frac{1}{N}$ and the probability of inclusion of any unit in the sample is $\frac{n}{N}$.

Estimating the Population Mean

Let Y be the character of interest and $Y_1, Y_2, \ldots, Y_i, \ldots, Y_N$ be the values of the character on N units of the population. Let, $y_1, y_2, \ldots, y_i, \ldots, y_n$ be the sample of size n selected by SRSWOR.

The estimator for the population mean is given by

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

An unbiased estimator of variance of the population mean is given by

$$\hat{V}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \text{ , where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Bound on the error of estimation (B)

$$B_y = t\,SE(\hat{\bar{Y}}) = t\sqrt{\hat{V}(\hat{\bar{Y}})}$$

Where *t* is the Student's t value for *n-1* degrees of freedom at the $1 - \alpha/2$ level of significance. For $\alpha = 0.05$, t = 1.96., Confidence Interval is given by C.I = $\hat{\bar{Y}} \pm B_y$

Estimating the Population Total

The estimator for the population total is given by

$$\hat{Y} = N\hat{\bar{Y}} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

An unbiased estimator of the variance of the population total is given by

$$V(\hat{Y}) = N^2 V\left(\hat{\bar{Y}}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

Where s^2 is an unbiased estimator of the population mean square S^2, $S^2 = \frac{1}{N-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Advantages of Simple Random Sampling

One of the best things about simple random sampling is the ease of assembling the sample. It is also considered as a fair way of selecting a sample from a given population since every member is given equal opportunities of being selected. Another key feature of simple random sampling is that the sample will be a representative of the population. If the sample is not representative of the population, the random variation called sampling error will be large. An unbiased random selection and a representative sample are important in drawing conclusions from the results of a study.

Disadvantages of Simple Random Sampling

One of the most obvious limitations of simple random sampling method is its need of a complete list of all the members of the population. However, from a practical point of view, a list of all the units of a population is not possible to obtain for large populations. Even if it is possible, it may involve a very high cost which a researcher or an organisation may not be able to afford. Therefore, simple random sampling is difficult to realize. Also, in case of a highly heterogeneous population, a simple random sample may not necessarily represent the characteristics of the total population, even though all selected units participate in the investigation. In those cases, it is wiser to use other sampling techniques.

Stratified Random Sampling

When the population is heterogeneous, the whole population can be divided into sub-populations, called strata, to increase the precision of the estimates. In stratified sampling the population of N units is first divided into disjoint groups of N_1, N_2,.., N_h,.., N_L units, respectively. These subgroups, called strata, together they compromise the whole population, so that $N_1 + N_2 + ... + N_h + .. + N_L = N$. The strata should not overlap and each stratum should be sampled following some sampling design. The strata are sampled separately and the estimates from each stratum combined into one estimate for the whole population. If a simple random sample selection scheme is used in each stratum then the corresponding sample is called a stratified random sample.

Reasons for stratification

- To obtain estimates of known precision for certain subdivisions of the population by treating each subdivision as a stratum. Since sampling is done independently in each stratum, separate stratum estimates and their precision can be obtained by treating each stratum as a ``population'' in its own right. For example, in fishery surveys estimates may be required by state, district, month, landing centre, craft, species etc.
- For administrative convenience; for example stratification can provide survey organization to control the distribution of fieldwork among its regional offices.

- Sometimes different parts of the population may call for different sampling procedures.
- Stratification may often produce a gain in precision of the estimates of characteristics of the whole population. The amount in the gain depends on the type of stratification. If the population is heterogeneous and if it can be divided, using prior information about the population, into subpopulations (strata), each of which is internally homogeneous. If each stratum is homogeneous, that is characteristic under consideration vary little from one unit to another, a precise estimate (an estimate with smaller variance) of any stratum parameter can be obtained from a small sample in that stratum. These estimates can then be combined to obtain a precise estimate for the whole population.

Notations

The suffix h (h =1,2,...,L) denotes the stratum and i the unit within the stratum.

N_h - Total number of population units in stratum h.

n_h - Total number of sample units in stratum h.

$w_h = N_h / N$ - The h^{th} stratum weight.

Y_{hi} - Value of the characteristic for the i^{th} unit in stratum h.

$Y_h = \sum_{i=1}^{N_h} Y_{hi}$ - Population total of Y – values for units belonging to stratum

$\overline{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi} = \frac{Y_h}{N_h}$ - Population mean of Y-values for units belonging to stratum h.

$\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}\left(Y_{hi} - \overline{Y}_h\right)^2$ -Population variance of Y-values for units belonging to stratum h.

$\overline{Y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi}}{\sum_{h=1}^{L} N_h} = \sum_{h=1}^{L} W_h \overline{Y}_h$ - Population mean of Y-values.

$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} = \frac{y_h}{n_h}$ - Sample mean of Y-values for units belonging to stratum h.

$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$ - Sample variance of Y-values for units belonging to stratum h.

$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} Y_{hi}}{\sum_{h=1}^{L} n_h}$ - Sample mean of Y-values.

Estimation of Population Mean and Variance

An unbiased estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}} = \sum_{h=1}^{L} W_h \bar{y}_h$$

Also, since sampling is done independently within each stratum

$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1}\sigma_h^2$$

Note that $V\left(\hat{\bar{Y}}\right)$ can not be computed since it involves Y-values for all the units in the population. However, based on Y-values for the sampled units we can estimate $V\left(\hat{\bar{Y}}\right)$ by using the following formula.

$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h n_h} s_h^2$ which estimates $V\left(\hat{\bar{Y}}\right)$ unbiasedly.

Allocation in Stratified Random Sampling

In planning a study, requiring stratification of the population, an important consideration is how to allocate a total sample size n among the L identified strata. There are three types of allocation.

Allocation of a Sample to Strata

1. Equal: If the strata are presumed to be of roughly equal size, and there is no additional information regarding the variability or distribution of the response in the strata, equal allocation to the strata is probably the best choice. $n_h = n/L$

2. Proportional: If the strata differ in size, allocation of sample sizes to strata might be performed proportional to these stratum sizes. $n_h = \left(\frac{N_h}{N}\right) n$

3. Optimum (Neyman): The allocation which minimizes the variance of the estimator of the mean (and total) is given by $n_h = \frac{nN_h\sigma_h}{\sum_{h=1}^{L} N_h\sigma_h}$

Whenever the strata are heterogeneous among themselves and the variance of each stratum is small, the sampling variance of the mean or total value estimators obtained by stratified sampling will always be smaller than the simple random sampling. The relative sizes of strata must be known to obtain the full benefits of the stratification technique. Each stratum should be internally homogeneous. If information about heterogeneity is not available then consider all strata equally variable. A short stratified pilot survey can sometimes provide useful information about internal dispersion within strata. A small sized sample could be taken from a stratum if the variability among their units is small. A larger sample from a stratum should be taken if the stratum is larger, the stratum is more heterogeneous and the cost of sampling the stratum is low.

References

Cochran, W. G. 1977. *Sampling Techniques*, third edition. John Wiley & Sons, Inc., New York, 428pp.

Lohr, S. L. 1999. *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA, 494pp.

Sukhatme, P.V. and B.V. Sukhatme. 1970. *Sampling Theory of Surveys with Applications*. Iowa State University Press, Ames, IA. 452pp.

Thompson, M. 1997. *Theory of Sample Surveys*. Chapman & Hall, 312pp.