Simple random sampling (SRS) is a method of selection of a sample comprising of \( n \) a number of sampling units out of the population having \( N \) number of sampling units such that every sampling unit has an equal chance of being chosen.

The samples can be drawn in two possible ways.

- The sampling units are chosen without replacement because the units once are chosen are not placed back in the population.
- The sampling units are chosen with replacement because the selected units are placed back in the population.

1. **Simple random sampling without replacement (SRSWOR):**

SRSWOR is a method of selection of \( n \) units out of the \( N \) units one by one such that at any stage of selection, any one of the remaining units have the same chance of being selected, i.e. \( 1/N \).

2. **Simple random sampling with replacement (SRSWR):**

SRSWR is a method of selection of \( n \) units out of the \( N \) units one by one such that at each stage of selection, each unit has an equal chance of being selected, i.e., \( 1/N \).

**Procedure of selection of a random sample:**

The procedure of selection of a random sample follows the following steps:

1. Identify the \( N \) units in the population with the numbers 1 to \( N \).
2. Choose any random number arbitrarily in the random number table and start reading numbers.
3. Choose the sampling unit whose serial number corresponds to the random number drawn from the table of random numbers.
4. In the case of SRSWR, all the random numbers are accepted ever if repeated more than once.

In the case of SRSWOR, if any random number is repeated, then it is ignored, and more numbers are drawn.
Such a process can be implemented through programming and using the discrete uniform distribution. Any number between 1 and \(N\) can be generated from this distribution, and the corresponding unit can be selected in the sample by associating an index with each sampling unit. Many statistical software like R, SAS, etc. have inbuilt functions for drawing a sample using SRSWOR or SRSWR.

**Notations:**

The following notations will be used in further notes:

- \(N\) : Number of sampling units in the population (Population size).
- \(n\) : Number of sampling units in the sample (sample size)
- \(Y\) : The characteristic under consideration
- \(Y_i\) : Value of the characteristic for the \(i^{th}\) unit of the population

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i : \text{sample mean}
\]

\[
\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i : \text{population mean}
\]

\[
S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \frac{1}{N-1} \left( \sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2 \right)
\]

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \frac{1}{N} \left( \sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2 \right)
\]

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n-1} \left( \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \right)
\]

**Probability of drawing a sample:**

1. **SRSWOR:**

If \(n\) units are selected by SRSWOR, the total number of possible samples are \(\binom{N}{n}\).

So the probability of selecting any one of these samples is \(\frac{1}{\binom{N}{n}}\).
Note that a unit can be selected at any one of the \( n \) draws. Let \( u_i \) be the \( i \)th unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, \ldots, or \( n \)th draw.

Let \( P_j(i) \) denotes the probability of selection of \( u_i \) at the \( j \)th draw, \( j = 1,2,\ldots,n \). Then

\[
P_j(i) = P_1(i) + P_2(i) + \ldots + P_n(i)
\]

\[
= \frac{1}{N} + \frac{1}{N} + \ldots + \frac{1}{N} \quad \text{(n times)}
\]

\[
= \frac{n}{N}.
\]

Now if \( u_1, u_2, \ldots, u_n \) are the \( n \) units selected in the sample, then the probability of their selection is

\[
P(u_1, u_2, \ldots, u_n) = P(u_1).P(u_2),\ldots,P(u_n).
\]

Note that when the second unit is to be selected, then there are \((n-1)\) units left to be selected in the sample from the population of \((N-1)\) units. Similarly, when the third unit is to be selected, there are \((n-2)\) units left to be selected in the sample from the population of \((N-2)\) units and so on.

If \( P(u_1) = \frac{n}{N} \), then

\[
P(u_2) = \frac{n-1}{N-1},\ldots,P(u_n) = \frac{1}{N-n+1}.
\]

Thus

\[
P(u_1, u_2, \ldots, u_n) = \frac{n}{N} \cdot \frac{n-1}{N-1} \cdot \frac{n-2}{N-2} \cdots \frac{1}{N-n+1} = \frac{1}{\binom{N}{n}}.
\]

**Alternative approach:**

The probability of drawing a sample in SRSWOR can alternatively be found as follows:

Let \( u_{i(k)} \) denotes the \( i \)th unit drawn at the \( k \)th draw. Note that the \( i \)th unit can be any unit out of the \( N \) units. Then \( s_o = (u_{i(1)}, u_{i(2)}, \ldots, u_{i(n)}) \) is an ordered sample in which the order of the units in which they are drawn, i.e., \( u_{i(1)} \) drawn at the first draw, \( u_{i(2)} \) drawn at the second draw and so on, is also considered. The probability of selection of such an ordered sample is

\[
P(s_o) = P(u_{i(1)})P(u_{i(2)} | u_{i(1)})P(u_{i(3)} | u_{i(1)}u_{i(2)})\ldotsP(u_{i(n)} | u_{i(1)}u_{i(2)}\ldotsu_{i(n-1)}).
\]

*Sampling Theory | Chapter 2 | Simple Random Sampling | Shalabh, IIT Kanpur*
Here \( P(u_{i(k)} | u_{(1)}, u_{(2)}, \ldots, u_{(k-1)}) \) is the probability of drawing \( u_{i(k)} \) at the \( k \)th draw given that
\( u_{(1)}, u_{(2)}, \ldots, u_{(k-1)} \) have already been drawn in the first \((k - 1)\) draws.

Such a probability is obtained as
\[
P(u_{i(k)} | u_{(1)}, u_{(2)}, \ldots, u_{(k-1)}) = \frac{1}{N - k + 1}.
\]

So
\[
P(s_n) = \prod_{k=1}^{n} \frac{1}{N - k + 1} = \frac{(N - n)!}{N!}.
\]

The number of ways in which a sample of size \( n \) can be drawn is \( n! \).

Probability of drawing a sample in a given order is \( \frac{(N - n)!}{N!} \)

So the probability of drawing a sample in which the order of units in which they are drawn is

irrelevant = \( n! \frac{(N - n)!}{N!} \cdot \frac{1}{n!} = \frac{1}{\binom{N}{n}} \).

2. SRSWR

When \( n \) units are selected with SRSWR, the total number of possible samples are \( N^n \).

The Probability of drawing a sample is \( \frac{1}{N^n} \).

Alternatively, let \( u_i \) be the \( i \)th unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, \( \ldots \), or \( n \)th draw. At any stage, there are always \( N \) units in the population in case of SRSWR, so the probability of selection of \( u_i \) at any stage is \( \frac{1}{N} \) for all \( i = 1, 2, \ldots, n \). Then the probability of selection of \( n \) units \( u_1, u_2, \ldots, u_n \) in the sample is

\[
P(u_1, u_2, \ldots, u_n) = P(u_1) \cdot P(u_2) \cdot \ldots \cdot P(u_n)
\]

\[
= \frac{1}{N} \cdot \frac{1}{N} \cdot \ldots \cdot \frac{1}{N}
\]

\[
= \frac{1}{N^n}
\]
Probability of drawing a unit

1. SRSWOR

Let $A_c$ denotes an event that a particular unit $u_j$ is not selected at the $\ell^{th}$ draw. The probability of selecting, say, $j^{th}$ unit at $k^{th}$ draw is

$$P(\text{selection of } u_j \text{ at } k^{th} \text{ draw}) = P(A \cap A_2 \cap \ldots \cap A_{k-1} \cap A_k)$$

$$= P(A_1) P(A_2 | A_1) P(A_3 | A_1, A_2) \ldots P(A_{k-1} | A_1, A_2, \ldots, A_{k-2}) \ P(\bar{A}_k | A_1, A_2, \ldots, A_{k-1})$$

$$= \left(1 - \frac{1}{N}\right) \left(1 - \frac{1}{N-1}\right) \left(1 - \frac{1}{N-2}\right) \ldots \left(1 - \frac{1}{N-k+2}\right) \frac{1}{N-k+1}$$

$$= \frac{1}{N} \cdot \frac{N-1}{N} \cdot \frac{N-2}{N-1} \ldots \frac{N-k+1}{N-k+2} \cdot \frac{1}{N-k+1}$$

$$= \frac{1}{N}$$

2. SRSWR

$$P[ \text{selection of } u_j \text{ at } k^{th} \text{ draw}] = \frac{1}{N}.$$ 

Estimation of population mean and population variance

One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatteredness of the data around the central value. Among various measures of central tendency and dispersion, the popular choices are arithmetic mean and variance. So the population mean and population variability are generally measured by the arithmetic mean (or weighted arithmetic mean) and variance, respectively. There are various popular estimators for estimating the population mean and population variance. Among them, sample arithmetic mean and sample variance is more popular than other estimators. One of the reasons to use these estimators is that they possess nice statistical properties. Moreover, they are also obtained through well-established statistical estimation procedures like maximum likelihood estimation, least squares estimation, method of moments etc., under several standard statistical distributions. One may also consider other measures like median, mode, geometric mean, harmonic mean for measuring the central tendency and mean deviation, absolute deviation, Pitman nearness etc. for measuring the dispersion. Numerical procedures like bootstrapping can study the properties of such estimators.
1. Estimation of population mean

Let us consider the sample arithmetic mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ as an estimator of the population mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ and verify $\bar{y}$ is an unbiased estimator of $\bar{Y}$ under the two cases.

**SRSWOR**

Let $t_i = \sum_{j=1}^{n} y_j$. Then

$$E(\bar{y}) = \frac{1}{n} E(\sum_{i=1}^{n} y_i)$$

$$= \frac{1}{n} E(t_i)$$

$$= \frac{1}{n} \left( \frac{1}{N} \sum_{i=1}^{n} t_i \right)$$

$$= \frac{1}{n} \left( \frac{1}{N} \right) \sum_{i=1}^{n} \left( \sum_{j=1}^{n} y_j \right)$$

When $n$ units are sampled from $N$ units without replacement, each unit of the population can occur with other units selected out of the remaining $(N-1)$ units in the population, and each unit occurs in

$$\left( \begin{array}{c} N-1 \\ n-1 \end{array} \right) \cdot \left( \begin{array}{c} n \\ n \end{array} \right)$$

the $\left( \begin{array}{c} N \\ n \end{array} \right)$ possible samples. So

$$\sum_{i=1}^{n} \left( \sum_{j=1}^{n} y_j \right) = \left( \begin{array}{c} N-1 \\ n-1 \end{array} \right) \sum_{i=1}^{N} y_i$$

So

$$\sum_{i=1}^{n} \left( \sum_{j=1}^{n} y_j \right) = \left( \begin{array}{c} N-1 \\ n-1 \end{array} \right) \sum_{i=1}^{N} y_i$$

Now

$$E(\bar{y}) = \frac{(N-1)!}{(n-1)!(N-n)!} \frac{n!(N-n)!}{nN!} \sum_{i=1}^{N} y_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$= \bar{Y}.$$

Thus $\bar{y}$ is an unbiased estimator of $\bar{Y}$.
Alternatively, the following approach can also be adapted to show the unbiasedness property. Let $P_j(i) = \frac{1}{N}$ denotes the probability of selection of $i^{th}$ unit at $j^{th}$ stage. Then

\[
E(\bar{y}) = \frac{1}{n} \sum_{j=1}^{n} E(y_j)
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \left[ \sum_{i=1}^{N} y_i P_j(i) \right]
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \left[ \sum_{i=1}^{N} y_i \frac{1}{N} \right]
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \bar{Y}
\]

\[
= \bar{Y}
\]

### SRSWR

\[
E(\bar{y}) = \frac{1}{n} E\left( \sum_{i=1}^{n} y_i \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} E(y_i)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( y_i P_1 + y_2 P_2 + ... + y_N P_N \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( y_i \frac{1}{N} + y_2 \frac{1}{N} + ... + y_N \frac{1}{N} \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \bar{Y}
\]

\[
= \bar{Y}
\]

where $P_i = \frac{1}{N}$ for all $i = 1, 2, ..., N$ is the probability of selection of a unit. Thus $\bar{y}$ is an unbiased estimator of the population mean under SRSWR also.
Variance of the estimate

Assume that each observation has some variance $\sigma^2$. Then

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2$$

$$= E\left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{Y})^2 \right]$$

$$= E\left[ \frac{1}{n^2} \sum_{i=1}^{n} (y_i - \bar{Y})^2 + \frac{1}{n^2} \sum_{i} \sum_{j} (y_i - \bar{Y})(y_j - \bar{Y}) \right]$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} E(y_i - \bar{Y})^2 + \frac{1}{n^2} \sum_{i} \sum_{j} E(y_i - \bar{Y})(y_j - \bar{Y})$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 + \frac{K}{n^2}$$

$$= \frac{N-1}{Nn} S^2 + \frac{K}{n^2}$$

where $K = \sum_{i} \sum_{j} E(y_i - \bar{Y})(y_j - \bar{Y})$ assuming that each observation has variance $\sigma^2$. Now we find $K$ under the setups of SRSWR and SRSWOR.

**SRSWOR**

$$K = \sum_{i} \sum_{j} E(y_i - \bar{Y})(y_j - \bar{Y})$$

Consider

$$E(y_i - \bar{Y})(y_j - \bar{Y}) = \frac{1}{N(N-1)} \sum_{k} \sum_{x_l} (y_k - \bar{Y})(y_l - \bar{Y})$$

Since

$$\left[ \sum_{k=1}^{N} (y_k - \bar{Y}) \right]^2 = \sum_{i=1}^{N} (y_k - \bar{Y})^2 + \sum_{k} \sum_{x_l} (y_k - \bar{Y})(y_l - \bar{Y})$$

$$0 = (N-1)S^2 + \sum_{k} \sum_{x_l} (y_k - \bar{Y})(y_l - \bar{Y})$$

$$\Rightarrow \frac{1}{N(N-1)} \sum_{k} \sum_{x_l} (y_k - \bar{Y})(y_l - \bar{Y}) = \frac{1}{N(N-1)}[-(N-1)S^2]$$

$$= -\frac{S^2}{N}.$$
Thus $K = -n(n-1)\frac{S^2}{N}$ and so substituting the value of $K$, the variance of $\bar{y}$ under SRSWOR is
\[
V(\bar{y}_{\text{WOR}}) = \frac{N-1}{Nn} S^2 - \frac{1}{n^2} n(n-1) \frac{S^2}{N}
\]
\[
= \frac{N-n}{Nn} S^2.
\]

**SRSWR**
\[
K = \sum_{i}^{N} \sum_{j \neq i}^{N} E(y_i - \bar{y})(y_j - \bar{y})
\]
\[
= \sum_{i}^{N} \sum_{j \neq i}^{N} E(y_i - \bar{y})E(y_j - \bar{y})
\]
\[
= 0
\]
because the $i^{th}$ and $j^{th}$ draws ($i \neq j$) are independent.

Thus the variance of $\bar{y}$ under SRSWR is
\[
V(\bar{y}_{\text{WR}}) = \frac{N-1}{Nn} S^2.
\]

It is to be noted that if $N$ is infinite (large enough), then
\[
V(\bar{y}) = \frac{S^2}{n}
\]
is both the cases of SRSWOR and SRSWR. So the factor $\frac{N-n}{N}$ is responsible for changing the variance of $\bar{y}$ when the sample is drawn from a finite population in comparison to an infinite population. This is why $\frac{N-n}{N}$ is called a finite population correction (fpc). It may be noted that
\[
\frac{N-n}{N} = 1 - \frac{n}{N}, \quad \text{so} \quad \frac{N-n}{N}
\]
is close to 1 if the ratio of a sample size to population $\frac{n}{N}$, is very small or negligible. The term $\frac{n}{N}$ is called the sampling fraction. In practice, fpc can be ignored whenever $\frac{n}{N} < 5\%$ and for many purposes, even if it is as high as 10%. Ignoring fpc will result in the overestimation of the variance of $\bar{y}$. 

*Sampling Theory| Chapter 2 | Simple Random Sampling | Shalabh, IIT Kanpur*
Efficiency of $\bar{y}$ under SRSWOR over SRSWR

\[
V(\bar{y}_{\text{WOR}}) = \frac{N-n}{Nn} S^2 \\
V(\bar{y}_{\text{WR}}) = \frac{N-1}{Nn} S^2 \\
= \frac{N-n}{Nn} S^2 + \frac{n-1}{Nn} S^2 \\
= V(\bar{y}_{\text{WOR}}) + \text{a positive quantity}
\]

Thus

\[
V(\bar{y}_{\text{WR}}) > V(\bar{y}_{\text{WOR}})
\]

and so, SRSWOR is more efficient than SRSWR.

Estimation of variance from a sample

Since the expressions of variances of the sample mean, involve $S^2$ which is based on population values, so these expressions can not be used in real-life applications. In order to estimate the variance of $\bar{y}$ on the basis of a sample, an estimator of $S^2$ (or equivalently $\sigma^2$) is needed. Consider $s^2$ as an estimator of $S^2$ (or $\sigma^2$) and we investigate its biasedness for $s^2$ in the cases of SRSWOR and SRSWR,

Consider

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
= \frac{1}{n-1} \sum_{i=1}^{n} \left[ (y_i - \bar{Y}) - (\bar{y} - \bar{Y}) \right]^2 \\
= \frac{1}{n-1} \left[ \sum_{i=1}^{n} (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right]
\]

\[
E(s^2) = \frac{1}{n-1} \left[ \sum_{i=1}^{n} E(y_i - \bar{Y})^2 - nE(\bar{y} - \bar{Y})^2 \right] \\
= \frac{1}{n-1} \left[ \sum_{i=1}^{n} Var(y_i) - nVar(\bar{y}) \right] = \frac{1}{n-1} \left[ n\sigma^2 - nVar(\bar{y}) \right]
\]
In the case of SRSWOR

\[ V(\bar{y}_{\text{WOR}}) = \frac{N-n}{Nn} S^2 \]

and so

\[ E(s^2) = \frac{n}{n-1} \left[ \sigma^2 - \frac{N-n}{Nn} S^2 \right] \]

\[ = \frac{n}{n-1} \left[ \frac{N-1}{N} S^2 - \frac{N-n}{Nn} S^2 \right] \]

\[ = S^2 \]

In the case of SRSWR

\[ V(\bar{y}_{\text{WR}}) = \frac{N-1}{Nn} S^2 \]

and so

\[ E(s^2) = \frac{n}{n-1} \left[ \sigma^2 - \frac{N-1}{Nn} S^2 \right] \]

\[ = \frac{n}{n-1} \left[ \frac{N-1}{N} S^2 - \frac{N-1}{Nn} S^2 \right] \]

\[ = \frac{N-1}{N} S^2 \]

\[ = \sigma^2 \]

Hence

\[ E(s^2) = \begin{cases} 
S^2 & \text{in SRSWOR} \\
\sigma^2 & \text{in SRSWR} 
\end{cases} \]

An unbiased estimate of \( Var(\bar{y}) \) is

\[ \hat{V}(\bar{y}_{\text{WOR}}) = \frac{N-n}{Nn} s^2 \] in case of SRSWOR and

\[ \hat{V}(\bar{y}_{\text{WR}}) = \frac{N-1}{Nn} \cdot \frac{N}{N-1} s^2 \]

\[ = \frac{s^2}{n} \] in case of SRSWR.
**Standard errors**

The standard error of $\bar{y}$ is defined as $\sqrt{Var(\bar{y})}$.

In order to estimate the standard error, one simple option is to consider the square root of the estimate of the variance of the sample mean.

- under SRSWOR, a possible estimator is $\hat{\sigma}(\bar{y}) = \sqrt{\frac{N-n}{Nn}}s$.

- under SRSWR, a possible estimator is $\hat{\sigma}(\bar{y}) = \sqrt{\frac{N-1}{Nn}}s$.

It is to be noted that this estimator does not possess the same properties as of $\sqrt{Var(\bar{y})}$.

The reason being if $\hat{\theta}$ is an estimator of $\theta$, then $\sqrt{\hat{\theta}}$ is not necessarily an estimator of $\sqrt{\theta}$. In fact, the $\hat{\sigma}(\bar{y})$ is a negatively biased estimator under SRSWOR.

The approximate expressions for large $N$ case are as follows:

Consider $s$ as an estimator of $S$.

Let $s^2 = S^2 + \varepsilon$ with $E(\varepsilon) = 0, E(\varepsilon^2) = S^2$.

Write

$$s = (S^2 + \varepsilon)^{1/2}$$

$$= S \left(1 + \frac{\varepsilon}{S^2}\right)^{1/2}$$

$$= S \left(1 + \frac{\varepsilon}{2S^2} - \frac{\varepsilon^2}{8S^4} + ...\right)$$

assuming $\varepsilon$ will be small as compared to $S^2$ and as $n$ becomes large, the probability of such an event approaches one. Neglecting the powers of $\varepsilon$ higher than two and taking expectation, we have
\[ E(s) = \left[ 1 - \frac{\text{Var}(s^2)}{8S^4} \right] S \]

where

\[ \text{Var}(s^2) = \frac{2S^4}{(n-1)} \left[ 1 + \left( \frac{n-1}{2n} \right)(\beta_2 - 3) \right] \text{ for large } N. \]

\[ \mu_j = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^j \]

\[ \beta_2 = \frac{\mu_4}{S^4} : \text{coefficient of kurtosis.} \]

Thus

\[ E(s) = S \left[ 1 - \frac{1}{4(n-1)} - \frac{\beta_2 - 3}{8n} \right] \]

\[ \text{Var}(s) = S^2 - S^2 \left[ 1 - \frac{1}{8} \left( \frac{\text{Var}(s^2)}{S^4} \right)^2 \right] = \frac{\text{Var}(s^2)}{4S^2} = \frac{S^2}{2(n-1)} \left[ 1 + \left( \frac{n-1}{2n} \right)(\beta_2 - 3) \right]. \]

Note that for a normal distribution, \( \beta_2 = 3 \) and we obtain

\[ \text{Var}(s) = \frac{S^2}{2(n-1)}. \]

Both \( \text{Var}(s) \) and \( \text{Var}(s^2) \) are inflated due to nonnormality to the same extent, by the inflation factor

\[ \left[ 1 + \left( \frac{n-1}{2n} \right)(\beta_2 - 3) \right] \]

and this does not depend on the coefficient of skewness.

This is an important result to be kept in mind while determining the sample size in which it is assumed that \( S^2 \) is known. If inflation factor is ignored and the population is non-normal, then the reliability on \( s^2 \) may be misleading.
Alternative approach:
The results for the unbiasedness property and the variance of the sample mean can also be proved in an alternative way as follows:

(i) SRSWOR

With the \( i^{th} \) unit of the population, we associate a random variable \( a_i \) defined as follows:

\[
a_i = \begin{cases} 
1, & \text{if the } i^{th} \text{ unit occurs in the sample} \\
0, & \text{if the } i^{th} \text{ unit does not occur in the sample} \quad (i=1,2,...,N)
\end{cases}
\]

Then,

\[
E(a_i) = 1 \times \text{Probability that the } i^{th} \text{ unit is included in the sample} = \frac{n}{N}, \quad i=1,2,...,N.
\]

\[
E(a_i^2) = 1 \times \text{Probability that the } i^{th} \text{ unit is included in the sample} = \frac{n}{N}, \quad i=1,2,...,N.
\]

\[
E(a_ia_j) = 1 \times \text{Probability that the } i^{th} \text{ and } j^{th} \text{ units are included in the sample} = \frac{n(n-1)}{N(N-1)}, \quad i \neq j = 1,2,...,N.
\]

From these results, we can obtain

\[
\text{Var}(a_i) = E(a_i^2) - (E(a_i))^2 = \frac{n(N-n)}{N^2}, \quad i=1,2,...,N
\]

\[
\text{Cov}(a_i,a_j) = E(a_ia_j) - E(a_i)E(a_j) = \frac{n(N-n)}{N^2(N-1)}, \quad i \neq j = 1,2,...,N.
\]

We can rewrite the sample mean as

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{N} a_iy_i
\]

Then

\[
E(\bar{y}) = \frac{1}{n} \sum_{i=1}^{N} E(a_i)y_i = \bar{Y}
\]

and

\[
\text{Var}(\bar{y}) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^{N} a_iy_i \right) = \frac{1}{n^2} \left[ \sum_{i=1}^{N} \text{Var}(a_i)y_i^2 + \sum_{i \neq j} \text{Cov}(a_i,a_j)y_iy_j \right].
\]
Substituting the values of \( \text{Var}(a_i) \) and \( \text{Cov}(a_i,a_j) \) in the expression of \( \text{Var}(\bar{y}) \) and simplifying, we get

\[
\text{Var}(\bar{y}) = \frac{N-n}{Nn} S^2.
\]

To show that \( E(s^2) = S^2 \), consider

\[
s^2 = \frac{1}{(n-1)} \left[ \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \right] = \frac{1}{(n-1)} \left[ \sum_{i=1}^{N} a_i y_i^2 - n\bar{y}^2 \right].
\]

Hence, taking expectation, we get

\[
E(s^2) = \frac{1}{(n-1)} \left[ \sum_{i=1}^{N} E(a_i) y_i^2 - n\left\{\text{Var}(\bar{y}) + \bar{y}^2\right\} \right]
\]

Substituting the values of \( E(a_i) \) and \( \text{Var}(\bar{y}) \) in this expression and simplifying, we get \( E(s^2) = S^2 \).

(ii) \textbf{SRSWR}

Let a random variable \( a_i \) associated with the \( i^{th} \) unit of the population denotes the number of times the \( i^{th} \) unit occurs in the sample \( i = 1, 2, \ldots, N \). So \( a_i \) assumes values 0, 1, 2, \ldots, \( n \). The joint distribution of \( a_1, a_2, \ldots, a_N \) is the multinomial distribution given by

\[
P(a_1,a_2,\ldots,a_N) = \frac{n!}{\prod_{i=1}^{N} a_i!} \cdot \frac{1}{N^n}
\]

where \( \sum_{i=1}^{N} a_i = n \). For this multinomial distribution, we have

\[
E(a_i) = \frac{n}{N},
\]

\[
\text{Var}(a_i) = \frac{n(N-1)}{N^2}, \quad i = 1, 2, \ldots, N.
\]

\[
\text{Cov}(a_i,a_j) = -\frac{n}{N^2}, \quad i \neq j = 1, 2, \ldots, N.
\]

We rewrite the sample mean as

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{N} a_i y_i.
\]

Hence, taking the expectation of \( \bar{y} \) and substituting the value of \( E(a_i) = n/N \) we obtain that

\[
E(\bar{y}) = \bar{Y}.
\]
Further,

\[ \text{Var}(\bar{y}) = \frac{1}{n^2} \left[ \sum_{i=1}^{N} \text{Var}(a_i) y_i^2 + \sum_{i=1}^{N} \text{Cov}(a_i, a_j) y_i y_j \right] \]

Substituting, the values of \( \text{Var}(a_i) = n(N - 1)/N^2 \) and \( \text{Cov}(a_i, a_j) = -n/N^2 \) and simplifying, we get

\[ \text{Var}(\bar{y}) = \frac{N-1}{Nn} S^2. \]

To prove that \( E(s^2) = \frac{N-1}{N} S^2 = \sigma^2 \) in SRSWR, consider

\[ (n-1)s^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{N} a_i y_i^2 - n\bar{y}^2, \]

\[ (n-1)E(s^2) = \sum_{i=1}^{N} E(a_i) y_i^2 - n \{ \text{Var}(\bar{y}) + \bar{y}^2 \} \]

\[ = \frac{n}{N} \sum_{i=1}^{N} y_i^2 - n \frac{(N-1)}{nN} S^2 - n\bar{y}^2 \]

\[ = \frac{(n-1)(N-1)}{N} S^2 \]

\[ E(s^2) = \frac{N-1}{N} S^2 = \sigma^2 \]

**Estimator of population total:**

Sometimes, it is also of interest to estimate the population total, e.g., total household income, total expenditures, etc. Let denotes the population total

\[ Y_r = \sum_{i=1}^{N} Y_i = N\bar{Y} \]

which can be estimated by

\[ \hat{Y}_r = N\hat{Y} = N\bar{y}. \]

Obviously

\[ E(\hat{Y}_r) = NE(\bar{y}) = N\bar{Y} \]
\[ Var(\hat{Y}_r) = N^2 Var(\bar{y}) = \begin{cases} \frac{N^2}{Nn} S^2 = \frac{N(N-n)}{n} S^2 & \text{for SRSWOR} \\ \frac{N^2}{Nn} S^2 = \frac{N(N-1)}{n} S^2 & \text{for SRSWOR} \end{cases} \]

and the estimates of variance of \( \hat{Y}_r \) are

\[ \widehat{Var}(\hat{Y}_r) = \begin{cases} \frac{N(N-n)}{n} s^2 & \text{for SRSWOR} \\ \frac{N}{n} s^2 & \text{for SRSWOR} \end{cases} \]

**Confidence limits for the population mean**

Now we construct the 100\((1 - \alpha)\)% confidence interval for the population mean. Assume that the population is normally distributed \( N(\mu, \sigma^2) \) with mean \( \mu \) and variance \( \sigma^2 \). Then \( \frac{\bar{y} - \mu}{\sqrt{Var(\bar{y})}} \) follows \( N(0,1) \) when \( \sigma^2 \) is known. If \( \sigma^2 \) is unknown and is estimated from the sample, then \( \frac{\bar{y} - \mu}{\sqrt{Var(\bar{y})}} \) follows a \( t \)-distribution with \((n-1)\) degrees of freedom. When \( \sigma^2 \) is known, then the 100\((1 - \alpha)\)% confidence interval is given by

\[
P\left[ -Z_{\alpha, \frac{\alpha}{2}} \leq \frac{\bar{y} - \mu}{\sqrt{Var(\bar{y})}} \leq Z_{\alpha, \frac{\alpha}{2}} \right] = 1 - \alpha
\]

or

\[
P\left[ \bar{y} - Z_{\alpha, \frac{\alpha}{2}} \sqrt{Var(\bar{y})} \leq \bar{Y} \leq \bar{y} + Z_{\alpha, \frac{\alpha}{2}} \sqrt{Var(\bar{y})} \right] = 1 - \alpha
\]

and the confidence limits are

\[
\left( \bar{y} - Z_{\alpha, \frac{\alpha}{2}} \sqrt{Var(\bar{y})}, \bar{y} + Z_{\alpha, \frac{\alpha}{2}} \sqrt{Var(\bar{y})} \right)
\]

where \( Z_{\alpha, \frac{\alpha}{2}} \) denotes the upper \( \frac{\alpha}{2} \)% points on \( N(0,1) \) distribution. Similarly, when \( \sigma^2 \) is unknown, then the 100\((1 - \alpha)\)% confidence interval is
\[ P \left[ -\frac{t_\alpha}{2} \leq \frac{\bar{y} - \bar{Y}}{\sqrt{\text{Var}(\bar{y})}} \leq \frac{t_\alpha}{2} \right] = 1 - \alpha \]

or

\[ P \left[ \bar{y} - t_\alpha \sqrt{\text{Var}(\bar{y})} \leq \bar{Y} \leq \bar{y} + t_\alpha \sqrt{\text{Var}(\bar{y})} \right] = 1 - \alpha \]

and the confidence limits are

\[ \left[ \bar{y} - t_\alpha \sqrt{\text{Var}(\bar{y})}, \bar{y} + t_\alpha \sqrt{\text{Var}(\bar{y})} \right] \]

where \( t_\frac{\alpha}{2} \) denotes the upper \( \frac{\alpha}{2} \) % points on \( t \)-distribution with \((n-1)\) degrees of freedom.

**Determination of sample size**

The size of the sample is needed before the survey starts and goes into operation. One point to be kept in mind is that when the sample size increases, the variance of estimators decreases but the cost of survey increases and vice versa. So there has to be a balance between the two aspects. The sample size can be determined on the basis of prescribed values of the standard error of the sample mean, the error of estimation, the width of the confidence interval, coefficient of variation of the sample mean, the relative error of sample mean or total cost among several others.

An important constraint or need to determine the sample size is that the information regarding the population standard derivation \( S \) should be known for these criteria. The reason and need for this will be clear when we derive the sample size in the next section. A question arises about how to have information about \( S \) beforehand? The possible solutions to this issue are to conduct a pilot survey and collect a preliminary sample of small size, estimate \( S \) and use it as a known value of \( S \) it. Alternatively, such information can also be collected from past data, past experience, the long association of the experimenter with the experiment, prior information, etc.

Now we find the sample size under different criteria assuming that the samples have been drawn using SRSWOR. The case for SRSWR can be derived similarly.
1. Pre-specified variance

The sample size is to be determined such that the variance of $\bar{y}$ should not exceed a given value, say $V$. In this case, find $n$ such that

$$Var(\bar{y}) \leq V$$

or

$$\frac{N-n}{N} \bar{y} \leq V$$

or

$$\frac{N-n}{N} S^2 \leq V$$

or

$$\frac{1}{n} - \frac{1}{N} \leq \frac{V}{S^2}$$

or

$$\frac{1}{n} - \frac{1}{N} \leq \frac{1}{n_e}$$

$$n \geq \frac{n_e}{1 + \frac{n_e}{N}}$$

where $n_e = \frac{S^2}{v}$.

It may be noted here that $n_e$ can be known only when $S^2$ is known. This reason compels to assume that $S$ should be known. The same reasoning will also be seen in other cases.

The smallest sample size needed in this case is

$$n_{\text{smallest}} = \frac{n_e}{1 + \frac{n_e}{N}}$$

It $N$ is large, then the required $n$ is

$$n \geq n_e \text{ and } n_{\text{smallest}} = n_e.$$

2. Pre-specified estimation error

It may be possible to have some prior knowledge of population mean $\bar{Y}$. It may be required that the sample mean $\bar{y}$ should not differ from it by more than a specified amount of absolute estimation error, i.e., which is a small quantity. Such a requirement can be satisfied by associating a probability $(1 - \alpha)$ with it and can be expressed as

$$P[|\bar{y} - \bar{Y}| \leq \varepsilon] = (1 - \alpha).$$
Since $\bar{y}$ follows $N\left(\bar{Y}, \frac{N-n}{Nn} S^2\right)$ assuming the normal distribution for the population, we can write

$$P\left[ \frac{\bar{y} - \bar{Y}}{\sqrt{\text{Var}(\bar{y})}} \leq \frac{e}{\sqrt{\text{Var}(\bar{y})}} \right] = 1 - \alpha$$

which implies that

$$\frac{e}{\sqrt{\text{Var}(\bar{y})}} = Z_{\frac{\alpha}{2}}$$

or

$$Z_{\frac{\alpha}{2}} \text{Var}(\bar{y}) = e^2$$

or

$$Z_{\frac{\alpha}{2}} \frac{N-n}{Nn} S^2 = e^2$$

or

$$n = \frac{\left(\frac{Z_{\frac{\alpha}{2}} S}{\frac{e}{\sqrt{\text{Var}(\bar{y})}}}\right)^2}{1 + \frac{1}{N} \left(\frac{Z_{\frac{\alpha}{2}} S}{\frac{e}{\sqrt{\text{Var}(\bar{y})}}}\right)^2}$$

which is the required sample size. If $N$ is large then

$$n = \left(\frac{Z_{\frac{\alpha}{2}} S}{\frac{e}{\sqrt{\text{Var}(\bar{y})}}}\right)^2$$

### 3. Pre-specified width of the confidence interval

If the requirement is that the width of the confidence interval of $\bar{y}$ with confidence coefficient $(1 - \alpha)$ should not exceed a pre-specified amount $W$, then the sample size $n$ is determined such that

$$2 Z_{\frac{\alpha}{2}} \sqrt{\text{Var}(\bar{y})} \leq W$$

assuming $\sigma^2$ is known and population is normally distributed. This can be expressed as

$$2 Z_{\frac{\alpha}{2}} \sqrt{\frac{N-n}{Nn}} S \leq W$$

or

$$4 Z_{\frac{\alpha}{2}} \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \leq W^2$$
or \( \frac{1}{n} \leq \frac{1}{N} + \frac{W^2}{4Z_a^2 S^2} \)

or \( n \geq \frac{4Z_a^2 S^2}{W^2} \cdot \frac{1}{4Z_a^2 S^2} \cdot \frac{1}{NW^2} \)

The minimum sample size required is

\[
n_{\text{smallest}} = \frac{4Z_a^2 S^2}{W^2} \cdot \frac{1}{4Z_a^2 S^2} \cdot \frac{1}{NW^2} \]

If \( N \) is large then

\[
n \geq \frac{4Z_a^2 S^2}{W^2} \]

and the minimum sample size needed is

\[
n_{\text{smallest}} = \frac{4Z_a^2 S^2}{W^2} . \]

### 4. Pre-specified coefficient of variation

The coefficient of variation (CV) is defined as the ratio of standard error (or standard deviation) and mean. The knowledge of the coefficient of variation has played an important role in the sampling theory as this information has helped in deriving efficient estimators.

If it is desired that the coefficient of variation of \( \bar{y} \) should not exceed a given or pre-specified value of the coefficient of variation, say \( C_0 \), then the required sample size \( n \) is to be determined such that

\[
CV(\bar{y}) \leq C_0
\]

or \[
\frac{\sqrt{Var(\bar{y})}}{\bar{y}} \leq C_0
\]
\[ \frac{N-n}{N} \frac{S^2}{\bar{Y}^2} \leq C_0^2 \]

or

\[ \frac{1}{n} - \frac{1}{N} \leq \frac{C_0^2}{C^2} \]

or

\[ n \geq \frac{C^2}{C_0^2} \left( 1 + \frac{C^2}{NC_0^2} \right) \]

is the required sample size where \( C = \frac{S}{\bar{Y}} \) is the population coefficient of variation.

The smallest sample size needed in this case is

\[ n_{\text{smallest}} = \frac{C^2}{C_0^2} \left( 1 + \frac{C^2}{NC_0^2} \right) \]

If \( N \) is large, then

\[ n \geq \frac{C^2}{C_0^2} \]

and

\[ n_{\text{smallest}} = \frac{C^2}{C_0^2} \]

5. Pre-specified relative error

When \( \bar{y} \) is used for estimating the population mean \( \bar{Y} \), then the relative estimation error is defined as \( \frac{\bar{y} - \bar{Y}}{\bar{Y}} \). If it is required that such relative estimation error should not exceed a pre-specified value \( R \) with probability \( (1 - \alpha) \), then such requirement can be satisfied by expressing it like such requirement can be satisfied by expressing it like

\[ P \left[ \frac{|\bar{y} - \bar{Y}|}{\sqrt{Var(\bar{y})}} \leq \frac{R\bar{Y}}{\sqrt{Var(\bar{y})}} \right] = 1 - \alpha. \]

Assuming the population to be normally distributed, \( \bar{y} \) follows \( N \left( \bar{Y}, \frac{N-n}{Nn} S^2 \right) \).
So it can be written that
\[
R \bar{Y} \sqrt{Var(\bar{Y})} = Z_{\alpha/2}
\]
or
\[
Z_{\alpha/2} \left( \frac{N - n}{Nn} \right) S^2 = R^2 \bar{Y}^2
\]
or
\[
\left( \frac{1}{n} - \frac{1}{N} \right) = \frac{R^2}{C^2 Z_{\alpha/2}^2}
\]
\[
\left( \frac{Z_{\alpha} C}{\bar{Y}} \right)^2
\]
or
\[
n = \frac{\left( \frac{Z_{\alpha} C}{\bar{Y}} \right)^2}{1 + \left( \frac{Z_{\alpha} C}{\bar{Y}} \right)^2}
\]
where \( C = \frac{S}{\bar{Y}} \) is the population coefficient of variation and should be known.

If \( N \) is large, then
\[
n = \left( \frac{Z_{\alpha} C}{R} \right)^2
\]

6. **Pre-specified cost**

Let an amount of money \( C \) be designated for sample survey to called \( n \) observations, \( C_0 \) be the overhead cost and \( C_1 \) be the cost of collection of one unit in the sample. Then the total cost \( C \) can be expressed as
\[
C = C_0 + nC_1
\]
Or
\[
n = \frac{C - C_0}{C_1}
\]
is the required sample size.