Simple Random Sampling for Attributes:

A qualitative characteristic which cannot be measured numerically is known as an **Attribute** , i.e. honesty, intelligence, beauty, etc. In many situations, it is not possible to measure the characteristic under study but possible to classify the population into various classes according to the attributes under study. For example: we can divide a population of a given city into two classes only, say, literate and illiterate with respect to the attribute ' **literacy**'. Hence the units in the population can be distributed in these two classes accordingly as it possesses or does not possess the given attribute. After taking a sample of size n, we may be interested in estimating the total number or the proportion of the defined attribute.

Notations:

Let us suppose that a population having **N** units U_1, U_2, U_3,…..,U_N is classified into 2 mutually disjoint and exhaustive classes. First class of possessing the attribute and the second class of not possessing the attribute. Then let,

A = Number of population units possessing the given attribute

N – A = Number of population units **not** possessing the given attribute

P = the **proportion** of population units possessing the given attribute = A/N

Q = the **proportion** of population units **not** possessing the given attribute = (N – A)/N = N/N – A/N = 1 – A/N = 1 – P

SRSWOR:

Let us consider SRSWOR sample of size n from this population of size N. If '**a'** is the number of units in a sample possessing the given attribute, then ,

p = **proportion** of sampled units possessing the given attribute

= a/n

q = **proportion** of sampled units **not** possessing the given attribute

= $a^{/}/n$ (where $a^{/} = n - a$)

Let Y_i be the value associated with the i^{th} unit of the population , where i = 1, 2, ……,N.

Y_i = 1, if i^{th} unit possesses the given attribute

= 0, if it does not possess the given attribute.

So, $\Sigma Y_i = A$, (i = 1, 2,….,N), is the number of units in the population possessing the given attribute.

Similarly, let us define y_i to be the value associated with i^{th} unit of the sample, where i = 1, 2,…..,n.

y_i = 1, if i^{th} sampled unit **possess** the given attribute

y_i = 0, if i^{th} sampled unit **does not possess** the given attribute.

Then, $\Sigma y_i = a$, is the number of sampled units possessing the given attribute.

Thus,

$\bar{Y} = (1/N)\Sigma Y_i = A/N = P$ and

$\bar{y} = (1/n)\Sigma y_i = a/n = p$

Now, $\Sigma Y_i^2 = \Sigma Y_i = A = NP$ i=1, 2,…..N and

$\Sigma y_i^2 = \Sigma y_i = a = np$, i=1, 2,…..,n

$S^2 = (1/N-1)\Sigma (Y_i - \bar{Y})^2$ i=1, 2,……,N

$= (1/N-1)[\Sigma Y_i^2 - N\bar{Y}^2]$

$= (1/N-1)[NP - NP^2]$

$=(NP/N-1)(1 - P)$

= NPQ/N-1 [Q=1-P]

Similarly , we get s^2 = npq/n-1

(a) **SRSWOR:**

From the population consisting of N units a sample of size n is selected by SRSWOR.

Estimation of population proportion and its variance.

Theorem 6:

(a) Sample proportion is an unbiased estimate of population proportion.

(b) It's variance is (N – n/N – 1)(PQ/n)

Proof: We know that in simple random sampling , sample mean is an unbiased estimate of population mean, i.e.

$E(\bar{y}) = \bar{Y}$ and

$\bar{Y} = P$, $\bar{y} = p$

Expression for variance of Sample Proportion :

We have,

$V(\bar{y})_{WOR} = (N - n/Nn) S^2$ and $S^2 = NPQ/N-1$

$V(p)_{WOR} = (N – n/Nn)* (NPQ/N-1)$

$= (N – n/n) * (PQ/N-1)$

$=(N – n/N – 1) * (PQ/n)$

$=PQ/n$, if N is large

Estimate of variance:

$V(p)_{WOR} = (N – n/N – 1)* (PQ/n)$

$= (N - n/ nN) * (NPQ/N - 1)$

Also, we have the result,

$E(s^2) = S^2$

$\Rightarrow E (npq/n-1) = (NPQ/N-1)$

$V(p)_{WOR} = (N - n/Nn) * (npq/n-1)$

$= (N - n/N) * (pq/n-1)$

$= pq/n-1,$ if N is large

$\approx pq/n-1,$ if N is large.

Confidence Interval for population proportion:

When from the population consisting of N units a sample of size n is selected by SRSWOR, we have the results

$E(p) = P , \quad V(p) = (N - n/N - 1)* PQ/n$

Estimate of $V (p) = v(p) = (N - n/N)* pq/n-1$

Assuming the Normal Distribution for the population, sample proportion p follows $N(P, V(p)_{WOR})$. From Standard Normal tables for given value α, we can read the ordinate $Z_{\alpha/2}$. Now we can write

$P (|p - P|/\sqrt{V(p)} < Z_{\alpha/2}) = 1 - \alpha$

$\Rightarrow P (-Z_{\alpha/2} < (p - P)/ \sqrt{V (p)} < Z_{\alpha/2}) = 1 - \alpha$

$\mathbf{\Rightarrow P (p - \sqrt{V(p)}Z_{\alpha/2} < P < p + \sqrt{V(p)}Z_{\alpha/2}) = 1 - \alpha}$

$\Rightarrow P (p - \sqrt{\hat{V}(p)}\, Z_{\alpha/2} < P < p + \sqrt{\hat{V}(p)}\, Z_{\alpha/2}) = 1 - \alpha$

$100(1 - \alpha)\%$ confidence interval for population proportion P is

$\{ p - Z_{\alpha/2}\sqrt{\hat{V}(p)}, p + Z_{\alpha/2}\sqrt{\hat{V}(p)}\}$

$= \{ p - Z_{\alpha/2}\sqrt{} (N-n/N)*pq/n-1 , p+ Z_{\alpha/2}\sqrt{} (N-n/N)*pq/n-1 \}$

$= \{ p - Z_{\alpha/2}\sqrt{} pq/n-1, p+ Z_{\alpha/2}\sqrt{} pq/n-1\}$ if N is large.

Estimation of total number of units possessing attributes and its variance:

$E(Np) = NP = A$

$\Rightarrow \hat{A} = Np$

$V(\hat{A}) = \hat{V}(Np)$

$= N^2\hat{V}(p)$

$= N^2\{(N-n)/N-1\}^* PQ/n$

$\hat{V}(\hat{A}) = \hat{V}(Np)$

$= N^2\hat{V}(p)$

$=N^2\{(N-n)/n-1\}^*pq/n-1$

$=N^2\{(N-n)/N\}^*pq/n-1$ [Since N-1=N when N is large]

$=N(N-n)/n-1^*pq$

Confidence Interval:

$100(1-\alpha)\%$ confidence interval for total number of units possessing attribute A is

$\{ Np - Z_{\alpha/2}\sqrt{\hat{V}(Np)}, Np + Z_{\alpha/2}\sqrt{\hat{V}(Np)} \}$

$\{ Np - Z_{\alpha/2}\sqrt{\{N(N-1)/(n-1)^*pq}, Np + Z_{\alpha/2}\sqrt{\{N(N-1)/(n-1)^*pq} \}$

(B) SRSWR:

Estimation of population proportion and its variance:

Theorem 7: Sample proportion is an unbiased estimate of population proportion and its variance is PQ/n.

Proof: We know that in SRSWR, sample mean is an unbiased estimate of population mean.

$\Rightarrow E(\bar{y}) = \bar{Y}$ and $\bar{Y} = P$, $\bar{y} = p$

$\Rightarrow E(p) = P$

Expression for variance of sample proportion:

We have,

$V(\bar{y})_{WR} = (N-n/Nn)^*S^2$ and $S^2 = NPQ/N-1$

$V(p)_{WR} = (N-n/Nn)^*(NPQ/N-1)$

$= PQ/n$ when N is large.

And we have ,

$E(s^2) = \sigma^2$

$\Rightarrow E(npq/n-1) = PQ/n$

$\hat{V}(p)_{WR} = npq/n(n-1) = pq/n-1$

Confidence Interval for population proportion:

100(1-α)% confidence interval for population proportion P is

$\{ p - Z_{\alpha/2}\sqrt{\hat{V}(p)} , p + Z_{\alpha/2}\sqrt{\hat{V}(p)} \}$

$\{ p - Z_{\alpha/2}\sqrt{(pq/n-1)} , p + Z_{\alpha/2}\sqrt{(pq/n-1)} \}$

Estimation of total number of units possessing attribute and its variance:

$E(Np) = NP = A$

$\Rightarrow \hat{A} = Np$

$V(\hat{A}) = V(Np)$

$= N^2 V(p)$

$= N^2 * PQ/n$

$\hat{V}(\hat{A}) = \hat{V}(Np)$

$= N^2 \hat{V}(p)$

$= N^2 * (pq/n-1)$

Confidence Interval:

100(1 – α)% confidence interval for total number of units possessing attribute A is

$\{ Np - Z_{\alpha/2}\sqrt{\hat{V}(Np)} , Np + Z_{\alpha/2}\sqrt{\hat{V}(Np)} \}$

$\{ Np - Z_{\alpha/2}\sqrt{\hat{V}(pq/n-1} , Np + Z_{\alpha/2}\sqrt{\hat{V}(pq/n-1} \}$

Estimation of Sample Size for Specific Precision:

The size of the sample is needed before the survey starts and goes into operation. We know 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 standard error of sample mean, error of estimation, width of the C.I, co-efficient of

variation of sample mean, relative error of sample mean or total cost among several others.

The most important problem faced by a statistician in any sample survey is to determine the sample size so that the population parameters may be estimated with a specified precision. The degree of precision can be determined in terms of:

(i) The level of significance in the estimate and

(ii) The CI with which this estimate lie with respect to the given level of significance.

(A) Data for variables:

It may be possible to have some prior knowledge of the population mean and it may be required that the sample mean should not differ from it by more than a specific amount of absolute estimation error Ɛ, which is a small quantity. Such requirement can be satisfied by associating probability with it and can also be expressed as

$$P(|\bar{y} - \bar{Y}| < Ɛ) = 1 - \alpha$$

(i) Sampling Without Replacement(SRSWOR):

We know,

$$E(\bar{y}) = \bar{Y} \quad \text{and} \quad V(\bar{y}) = (N - n/Nn)S^2$$

$$P(| \bar{y} - \bar{Y}| < Ɛ) = 1 - \alpha$$

$$\Rightarrow P(| \bar{y} - \bar{Y} |/\sqrt{V(\bar{y})} < Ɛ/\sqrt{V(\bar{y})}) = 1 - \alpha$$

As $\bar{y}$ follows $N(\bar{Y}, (N-n/Nn)S^2)$ assuming the Normal Distribution for the population, from standard normal tables for given value of α and we can read the ordinate $Z_{\alpha/2.}$ So we can write

$$P(| \bar{y} - \bar{Y} |/\sqrt{V(\bar{y})} < Ɛ/\sqrt{V(\bar{y})}) = 1 - \alpha$$

$\Rightarrow \mathcal{E}/\sqrt{V(\bar{y})} = Z_{\alpha/2}$

We have, $V(\bar{y}) = (N-n/Nn)S^2 = (1/n - 1/N)S^2$

$\Rightarrow \mathcal{E} / S\sqrt{(1/n - 1/N)} = Z_{\alpha/2}$

Squaring both sides, we get

$\mathcal{E}^2/S^2 (1/n - 1/N) = Z^2_{\alpha/2}$

$\Rightarrow (1/n - 1/N) = \mathcal{E}^2/S^2 Z^2_{\alpha/2}$

$\Rightarrow 1/n = \mathcal{E}^2/S^2 Z^2_{\alpha/2} + 1/N$

$\Rightarrow 1/n = (N\mathcal{E}^2 + S^2 Z^2_{\alpha/2})/N S^2 Z^2_{\alpha/2}$

$\Rightarrow n = N S^2 Z^2_{\alpha/2}/ (N\mathcal{E}^2 + S^2 Z^2_{\alpha/2})$

$\Rightarrow n = N S^2 Z^2_{\alpha/2}/\{ \mathcal{E}^2 (N\mathcal{E}^2 + S^2 Z^2_{\alpha/2}/ \mathcal{E}^2)\}$

$\Rightarrow n = \{N S^2 Z^2_{\alpha/2}/ \mathcal{E}^2\}/ N(1 + S^2 Z^2_{\alpha/2}/ \mathcal{E}^2 N)$

$\Rightarrow n = (SZ_{\alpha/2}/\mathcal{E})^2 / \{ 1 + 1/N(SZ_{\alpha/2}/\mathcal{E})^2\}$

$\Rightarrow n = n_0/(1 + n_0/N_0)$

Where $N = N_0$, $n_0 = (SZ_{\alpha/2}/\mathcal{E})^2$

Thus , we get,

$n_0 = (SZ_{\alpha/2}/\mathcal{E})^2$

and $n = n_0$, where N is large

$= n_0/(1 + n_0/N_0)$, otherwise

(ii) Sampling With Replacement(SRSWR):

Since $P(| \bar{y} - \bar{Y}| < \mathcal{E}) = 1 - \alpha$

$\Rightarrow P(| \bar{y} - \bar{Y} |/\sqrt{V(\bar{y})} < \mathcal{E}/\sqrt{V(\bar{y})}) = 1 - \alpha$

We have the result

$E(\bar{y}) = \bar{Y}$ and $V(\bar{y}) = \sigma^2/n$

So $\bar{y}$ follows N($\bar{Y}$, σ^2/n) assuming the Normal Distribution for the population . From Standard Normal table for given α, we can read the ordinate $Z_{\alpha/2}$

$P (| \bar{y} - \bar{Y} |/\sqrt{V(\bar{y})} < Z_{\alpha/2} = 1 - \alpha$

So that , $\varepsilon/\sqrt{V(\bar{y})} = Z_{\alpha/2}$

We have, $V(\bar{y}) = \sigma^2/n$

$\Rightarrow \varepsilon/\sqrt{\sigma^2/n} = Z_{\alpha/2}$

$\Rightarrow \varepsilon/(\sigma/\sqrt{n}) = Z_{\alpha/2}$

$\Rightarrow \sqrt{n} = \sigma Z_{\alpha/2}/ \varepsilon$

$\Rightarrow n = (\sigma Z_{\alpha/2}/ \varepsilon)^2$

(B) Data for Attributes:

(i) Sampling Without Replacement(SRSWOR):

When from the population consisting of N units a sample of size n is selected by SRSWOR, we have the results,

$E(p) = P$ and $V(p) = (N - n/N - 1)* PQ/n$

Estimate of $V(p) = v(p) = (N - n/N)*pq/n-1)$

We have,

$P (|p - P| < \varepsilon) = 1 - \alpha$

$\Rightarrow P \{|p - P|/\sqrt{V(p)} < \varepsilon/\sqrt{V(p)}\} = 1 - \alpha$

Since p follows N (P, (N-n/N-1)*PQ/n) from Standard Normal tables for given value α, we can read the ordinate

$Z_{\alpha/2}$ as $P \{|p - P|/\sqrt{V(p)} < Z_{\alpha/2}) = 1 - \alpha$

$\Rightarrow \varepsilon/\sqrt{V(p)} = Z_{\alpha/2}$

We have $V(p) = (N-n/N-1)*PQ/n$

$\Rightarrow \varepsilon/\sqrt{V(p)} = Z_{\alpha/2}$

$\Rightarrow \mathcal{E}/\sqrt{(N-n/N-1)*PQ/n} = Z_{\alpha/2}$

Squaring both sides, we get

$\mathcal{E}^2/(N-n/N-1)*PQ/n = Z^2_{\alpha/2}$

$\mathcal{E}^2 = Z^2_{\alpha/2}(N-n/N-1)*(PQ/n)$

$\mathcal{E}^2 = Z^2_{\alpha/2}(N-n/N)*(PQ/)n$

$nN = 1/\mathcal{E}^2 [Z^2_{\alpha/2}(N-n)*(PQ)]$

$nN = 1/\mathcal{E}^2 [Z^2_{\alpha/2} NPQ - Z^2_{\alpha/2} nPQ]$

$nN + Z^2_{\alpha/2} nPQ/\mathcal{E}^2 = Z^2_{\alpha/2} NPQ/\mathcal{E}^2$

$nN\mathcal{E}^2 + Z^2_{\alpha/2} nPQ = Z^2_{\alpha/2} NPQ$

$n(N\mathcal{E}^2 + Z^2_{\alpha/2} PQ) = Z^2_{\alpha/2} NPQ$

$n = Z^2_{\alpha/2} NPQ/(N\mathcal{E}^2 + Z^2_{\alpha/2} PQ)$

$n = Z^2_{\alpha/2} NPQ/N(\mathcal{E}^2 + Z^2_{\alpha/2} PQ/N)$

$n = Z^2_{\alpha/2} PQ/\mathcal{E}^2 + Z^2_{\alpha/2}(PQ/N)$

$n = Z^2_{\alpha/2} PQ/\mathcal{E}^2(1 + Z^2_{\alpha/2}(PQ/\mathcal{E}^2N)$

$n = (Z^2_{\alpha/2} PQ/\mathcal{E}^2)/[1 + \{Z^2_{\alpha/2}(PQ/\mathcal{E}^2)\}/N]$

$n = n_0/(1 + n_0/N_0)$

where $n_0 = Z^2_{\alpha/2}(PQ/\mathcal{E}^2)$

$N_0 = N$

(ii)Sampling With Replacement (SRSWR):

When from the population consisting of N units a sample size n is selected by SRSWR, we have the results

$E(p) = P$, $V(p) = PQ/n$

We have, $P(|p - P| < \mathcal{E}) = 1 - \alpha$

$\Rightarrow P\{|p - P|/\sqrt{V(p)} < \varepsilon/\sqrt{V(p)}\} = 1 - \alpha$

Assuming the Normal Distribution for the population, sample population p follows $N(P, V(p)_{WR})$, i.e,p follows $N(p, PQ/n)$

From Standard Normal tables for a given value α, we can read the ordinate $Z_{\alpha/2}$

$P(|p - P|/\sqrt{V(p)} < Z_{\alpha/2}) = 1 - \alpha$

So that $\varepsilon/\sqrt{V(p)} = Z_{\alpha/2}$

We have $V(p) = PQ/n$

$\varepsilon/\sqrt{PQ/n} = Z_{\alpha/2}$

Squaring both sides, we get,

$\varepsilon^2/(PQ/n) = Z^2_{\alpha/2}$

$n\,\varepsilon^2/PQ = Z^2_{\alpha/2}$

$n = Z^2_{\alpha/2} PQ/\varepsilon^2$

Advantages and Disadvantages of Simple Random Sampling(SRS):

Advantages:

(i) It needs only a minimum knowledge of the study group of population in advance.
(ii) It is free from errors in classification
(iii) This is suitable for data analysis which includes the use of inferential statistics.
(iv) Simple random sampling is representative of the population.
(v) It is totally free from bias and prejudice.
(vi) The method is simple to use.
(vii) It is very easy to assess the sampling error in this method.

Disadvantages:

(i) This method carries larger error from the same sample size than that are found in stratified sampling.
(ii) In SRS, the selection of sample becomes impossible if the units or items are widely dispersed.
(iii) It cannot be used where the units of the population are heterogeneous in nature.
(iv) This method lacks the use of available knowledge concerning the population.
(v) Sometimes, it is difficult to have a completely catalogued universe.
(vi) It may be impossible to contact the cases which are very widely dispersed.