SIMPLE RANDOM SAMPLING

BY DR RAJIV SAKSENA
DEPARTMENT OF STATISTICS
UNIVERSITY OF LUCKNOW

1. INTRODUCTION

Population and Sample: A **population** is an aggregate of individual about whom information is sought and a **sample** is a small part of that aggregate. In everyday life as well as in scientific research or any type of enquiry about a population, our attitudes and actions are often based on samples. For e.g., when information about the proportion of good items in a lot of manufactured product, total no. of graduates in a state, average consumer expenditure in a city, total turnover of sales of an item, total area under a crop, total number of fish in a lake is required, often the estimates are observed on the basis of suitably selected sample.

It is assumed that the population is well defined and consists of a finite member N (usually quite large) of individuals, called units, U_1, U_2, …, U_n. The definition of a population may be easy, as in case of electric bulbs manufactured in a factory, but not so as in case of forms or fields in villages where clear specification is necessary to take care of borderline or doubtful cases. Also, the units constituting population must be defined properly depending upon the type of information sought. For example, in case of a population of people the unit may be individual persons or a family of persons or a group of families residing in a locality etc. It is very useful to have a complete list of units in a population which is called a frame or a sampling frame. Sometimes it is impossible to control a frame as in case of the fish population in a lake.

It is also supposed that every unit in the population can be measured, quantitatively or qualitatively, with respect to some characteristic under study, say Y, who for U_i is Y_i (I= 1, 2, -- , N). For example, the unit may be a person and Y may be his age or income; the unit may be a family and Y may be the total consumer expenditure in a particular month; or the unit may be a field and Y may be its area or yield of a crop sown in it. Any function of the values of all the population units is known as a population parameter which is to be calculated. Two common parameter are 'total' and 'mean' and our basic problem is to estimate them. These are:

Total Y $= y_1 + y_2 + \text{----}\ y_n$

Mean $\bar{Y}$ $=$ Y / N

There are two ways of finding the value of Y or $\overline{Y}$, either by measuring every unit in the population, called complete enumeration or census, or estimating them on the basis of sampling or sample survey which consist of selecting a sample suitably and using the sample values.

2. ADVANTAGES OF SAMPLING OVER COMPLETE ENUMERATION

(a) Greater speed: Since a sample consist of only a small number of units, sampling requires less time than complete enumeration.

(b) Reduced Cost: Again, because of small size, sampling requires less cost than complete enumeration. The total cost of complete enumeration is considerable for it involves a large administrative machinery for the purpose.

(c) Greater accuracy: The sampling method involves a small number of workers who can be given efficient training and facilities to gather information more correctly than in case of a large body of enumeration for the entire population.

(d) Greater Scope: The smaller scope of sampling procedure and the better training to which the investigators may be exposed also enable the enquirer to called data on a greater number of items than it is possible for a complete enumeration.

(e) Greater applicability: Sometime it is not easily possibly to provide proper facilities to workers for complete enumeration and than the only methods available is sampling. Also, when enumeration is a destructive process, sampling is a must.

(f) Provision of measure of accuracy: The complete enumeration involves a lot of errors of measurement and no estimate of accuracy of result is possible, while sampling, done statistically provide us with a measure of the accuracy of the estimate obtained.

3. SAMPLING AND NON-SAMPLING ERRORS

The errors arising in the stages of measurement and processing of data are termed Non-sampling errors These are common to both complete enumeration and sampling. The main types of such errors are the following:

(a) Errors of measurement: When the units are measurement by observation e.g. eye-estimation of a crops area or yield, the measurement will depend on the personal judgment of the enumerator and will be subject to errors. Usually, it has been observed that eye-estimation of yield is under-estimation while that of average is over–estimation.

Such errors may also arise due to response biases. For example, persons interviewed may give wrong answers regarding one’s education or age, make under-statement of income or over statement of expenses.

Sometimes interviewer biases may crop in due to answers given by suggestions from the interviewer or due to influence of interviewer's belief and prejudices in interpreting some questions.

(b) Errors of non-response: These may arise if the respondent is not found at home or office even after repeated calls or if he refuses or fails to furnish the required information. Non–response leads to a section of the population being totally excluded and results in biased estimation.

(c) Errors of tabulation: Such errors arise due to inadequate scrutiny of the data, errors in coding, computation and tabulation of data and errors committed during presentation and printing of results.

On the other hand, "sampling error" arises solely due to sampling fluctuation. Since a sample is just a fraction of the entire population, there will generally be a difference between the population parameter and its sampling estimate. Complete enumeration is free from such errors although the non- sampling errors for it is usually much greater than in sample surrey. The sampling error usually decreases with increase with sample size.

4. PROBABILITY (OR RANDOM) SAMPLING

A sample should be a true representative of the population of the estimate based on it has to be near the true value of the parameter which is being estimated. In practice we are not very much worried if the difference between the true & estimated value is large for a specific sample as long as such differences are small (or negligible) in repeated samples. Hence the sampling procedure is important. Instead of purposive or haphazard selection, method of probability or random sampling is adopted. This has the following steps:

(i) We are able to define the set of destined samples S_1, S_2, S_ν which the procedure is capable of selecting from the population,

(ii) each possible sample Si has assigned to it a known probability of selection,

(iii) we select one of the sample by a mechanism such that the ith–sample Si receives its appropriate probability of being selected, For example, we may assign equal probability to all possible Sample.

(iv) The method of computing the estimate from the sample must be stated and must lead to a unique estimate for any particular sample, For example, the average of sample value be an estimate of the population mean.

For any sampling procedure, there is a frequency distribution of the estimator which it generates when the procedure is repeatedly applied to the same population. This is known as the sampling distribution of the estimate.

5. BIAS OF AN ESTIMATOR

Suppose the parameter of the population is θ(Y or $\overline{Y}$). Let Z be an estimate for which Z = Zi when the sample selected is Si (i= 1, 2,. , ν) with probability π_i. The bias of the estimator is given by

$$B(Z) = E(Z) - \theta$$
$$= \sum_{i=1}^{\nu} \pi_i Zi - \theta$$

The estimator is unbiased if B(Z) = 0 or E(Z) = θ .

6. MEASURES OF SAMPLING ERROR

Since probability sampling gives rise to different samples, the estimates based on the sample observations will differ from sample to sample and, also, deviate from the value of the parameter. The difference between the estimate Zi, based on the ith sample Si, and the parameter θ i.e.(Z i - θ), may be called the error of the estimate and this error varies from sample to sample. An average measure of the divergence of the estimator from the true value is given by

$$M(Z) = \sum_{i=1}^{\nu} (Zi - \theta)^2 \pi_i$$

which is called the mean square error (m.s.e) of the estimator. The m.s.e. may be considered to be a measure of accuracy with which the estimator Z estimates the parameter. The sq.root of the m.s.e. is termed as “root mean square error.” The accuracy is inversely proportional to m.s.e. The sampling variance of the estimator is defined by

$$V(Z) = \sum_{i=1}^{\nu} [Zi - E(z)]^2 \pi_i$$

$$\text{or} \quad M(z) = V(z) + [B(z)]^2$$

If the estimator is unbiased, M (z) = V (z) . The positive square root of the variance of an estimator z is called the Standard error of Z. The sampling variance (or Standard error) may have been considered as a measure of precision of the estimator. The precision is inversely proportional to sampling variance.

SIMPLE RANDOM SAMPLING

The simplest type of probability sampling where the probabilities associated with different possible samples are equal is called simple random sampling procedure. In this procedure, the sample is drawn unit by unit with equal probability of selection for every unit in each draw.

Suppose a simple random sample of n units is to be drawn from a population of N units U_1,U_2,---,U_n, for which the values of the characteristic Y under study are y_1, y_2, --, yn. The sample may be drawn in two different ways.

7. SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT (SRSWOR)

In this case, the n units of the sample are drawn from the population one by one, the unit obtained at any drawn not being replaced in the population, in such a way the probability of any unit in the first draw is 1/N, that of any unit in the second drawn is 1/(N-1),……….., that of any unit in the rth draw is 1/(N-r+1) and so on. Therefore, the probability of drawing a sample of n units in SRSWOR is $\frac{n!}{N(N-1)......(N-n+1)} = \frac{1}{\binom{N}{n}}$

This means that there are $\binom{N}{n}$ possible samples; the probability of drawing each of these is the same.

The probability that a specified unit is selected at the rth draw in SRSWOR is

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

The probability that a specified unit is included in the sample is

$$\sum_{r=1}^{n} \text{Probability of selecting the unit in the rth draw} = n/N$$

8. SIMPLE RANDOM SAMPLING WITH REPLACEMENT (SRSWR)

In this case, the n units of the sample are drawn from the population one by one, the units obtained at any draw being replaced in the population, in such a way that the probability of drawing any unit in any draw is 1/N

The probability of drawing a Sample of n units in SRSWR is

$$\frac{1}{N^n}$$

This means that are N^n possible samples, the probability of drawing each of these is the sample. The probability that a specified unit is selected at any draw is 1/N and the probability that a specified unit is included in the sample is n/N.

9. PROCEDURE OF SELECTING A RANDOM SAMPLE

The first step is to prepare a list of all (N) units in the population and number them serially from 1 to N. Then a sample of n units is taken by are the following methods.

(a) Lottery method : Taken N cards or tickets or counters bearing numbers from 1 to N, these are thoroughly mixed and n tickets are drawn, one by one, from this lot (either without replacement or with replacement)& these numbers noted, mixing the tickets thoroughly after each draw. Subsequently, the sample of n units is selected from the population which bears the numbers on these tickets. This method is cumbersome &does not guarantee that units will be selected with equal probability. Human bias & prejudice may also creep in the method.

(b) <u>Use of random numbers tables</u>: A random number tables is an arrangement of digits 0 to 9, in either a linear \ rectangular pattern, such that all the ten numbers appear, independently of one another, with the same frequency. By combining the numbers in pairs or triplets or quadruplets, we have the numbers 00 to 99, 000 to 999 and 0000 to 9999 which ocean with approximately the same frequency. Some random members tables in common use are.

(i) Tippet's random numbers tables
(ii) Fisher & Yates tables
(iii) Rand corporation series
(iv) Cordell & smith Series
(v) I.S.I. Series (in Tables of Ras – Mibe & Matter)

The simplest way of selecting a sample of n units from a population of N units is to look at any row (or column) of the tables and select n random numbers between 1 and N and, then, taking the population units bearing those numbers. The procedure involves a number of rejections since all numbers greater than N appearing in the table is rejected for consideration. The use of random numbers is, therefore modified by using either remainder approach or quotient approach as follows:

<u>Remainder approach:</u> Let N be a r-digit number and let its r-digit highest multiple be N. A random number k is chore from I to N and the unit equal to the remainder obtained on dividing k by N is selected. If the remainder is zero, the last unit is selected. For example, let N = 123, the highest these digit multiple is 984. For selecting a unit, are random number from 001 to 984 is selected, say 287 which on division by 123 gives 41 as the remainder. Hence the unit with serial number 41 is selected in the sample.

<u>Quotient approach:</u> Let N be a r-digit number and let its r-digit highest multiple be N′ such that N′/N =q, A random number k is chosen from 0 to (N′-1) and the quotient r is obtained on dividing k by q. Then the unit bearing the serial number (r-1) is selected in the sample. For example, let N = 123 and N= 984 or that q = 8. If the random number selected is 287 than r=35. Hence the unit with serial number 34 is included in the sample.

10. ESTIMATION OF POPULATION MEAN OR TOTAL

Let us suppose that the value of Y, the character under study, is Yi for the ith- population unit Ui (i=1, 2, …., N). We define

Population mean $\overline{Y} = \sum_{i=1}^{N} y_i / N$

Population total Y = $\sum_{i=1}^{N} y_i \;=\; N\overline{Y}$

Population variance $\sigma^2 \;=\; \sum_{i=1}^{N} (y_i - \overline{Y})^2 / N$

$$\text{and } S^2 \;=\; \sum_{i=1}^{N} (y_i - \overline{Y})^2 / (N - 1) = \frac{N\sigma^2}{N - 1}$$

Also, for a sample of size n let the sample values be y_1, y_2 ,…., y_n. It doesn't mean that only first n population units are selection in the sample but, without any loss of generally, these may be considered to be values of the units selected in that order. Evidently rth sample unit value Y_r may be any of the population values y_1, y_2 ,…., y_N, We define,

Sample mean $\overline{y} \;=\; \sum_{i=1}^{n} y_i / n$

Sample total y $\;=\; \sum_{i=1}^{n} y_i \;=\; n\overline{y}$

Sample variance $s^2 \;=\; \frac{1}{n - 1} \sum_{i=1}^{n} (y_i - \overline{y})^2$

<u>Case I</u>: <u>Simple random sampling without replacement (SRSWOR):</u>

<u>Theorem I</u>: $\overline{y}$ is an unbiased estimator of $\overline{Y}$ and its variance is given by

$$V(\overline{y}) \;=\; \frac{N - n}{N}\,\frac{S^2}{n} \;=\; (1 - f)\,\frac{S^2}{n}$$

Where, f = n/N, is the sampling fraction

<u>Proof:</u> We have

$$E(\overline{y}) \;=\; E\left[\sum_{1}^{n} y / n\right]$$

$$= \frac{1}{n} \sum_{r=1}^{n} E(y_r)$$

$$= \frac{1}{n} \sum_{r=1}^{h} \left(\sum_{r=1}^{N} y_r / N \right)$$

$$= \frac{1}{h} \sum_{r=1}^{h} \left(\frac{1}{N} \sum_{i=1}^{N} y_r \right)$$

Because y_r takes the value y_i (i = 1, .., N) with probability 1/N.

therefore, $E(\bar{y}) = \frac{1}{n}\sum_{r=1}^{n}\bar{Y}$

$$= \bar{Y}$$

so that $\bar{y}$ is unbiased for $\bar{Y}$

We here

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

$$= E[\{1/n\sum_{r=1}^{n}(y_i - \bar{Y})\}^2]$$

$$= \frac{1}{n^2}E\{\sum_{r=1}^{n}(y_r - \bar{Y})\}^2$$

$$= \frac{1}{n^2}E\{\sum_{r=1}^{n}(y_r - \bar{Y})^2 + \underset{r \neq l}{\sum^{n}\sum^{n}}(y_r - \bar{Y})(y_l - \bar{Y})\}$$

Now,

$$E\{\sum_{r=1}^{n}(y_r - \bar{Y})^2\} = \sum_{r=1}^{n}E(y_r - \bar{Y})^2$$

$$= \sum_{r=1}^{n}1/N\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

$$= n\sigma^2$$

Also,

$$E\{\underset{r \neq l}{\sum\sum}(y_r - \bar{Y})(y_l - \bar{Y})\} = \underset{r \neq l}{\sum^{n}\sum^{n}}E(y_r - \bar{Y})(y_l - \bar{Y})$$

$$= \underset{r \neq l}{\sum\sum}\left(\frac{1}{N(N-1)}\underset{i \neq j}{\sum^{N}\sum^{N}}(y_i - \bar{Y})(y_j - \bar{Y})\right)$$

$$= \frac{1}{N(N-1)}\underset{r \neq l}{\sum^{n}\sum^{n}}\left(\{\sum_{i=1}^{N}(y_i - \bar{Y})\}^2 - \sum_{i=1}^{N}(y_i - \bar{Y})^2\right)$$

$$= \frac{-1}{N(N-1)}\underset{r \neq l}{\sum^{n}\sum^{n}}\left(\sum_{i=1}^{N}(y_i - \bar{Y})\right)^2$$

$$= -n(n-1)\sigma^2$$

$$\overline{}\\ \text{N-1}$$

Therefore,

$$\begin{aligned} V(\bar{y}) &= \frac{1}{n^2}\left(n\sigma^2 - \frac{n(n-1)\sigma^2}{N-1} \right) \\ &= \frac{N-n}{N-1}\,\frac{\sigma^2}{n} \\ &= \left(\frac{N-n}{N}\right)\frac{S^2}{n} \end{aligned}$$

The standard error of $\bar{y}$ is $\sigma_{\bar{y}} = \dfrac{N-n}{N}\,\dfrac{S}{\sqrt{n}}$

Corollary : The unbiased estimate of population total $\bar{Y}$ is Ny having variance

$$V(N\bar{y}) = N^2V(\bar{y}) = (N\text{-}n)N\frac{S^2}{n}$$

Theorem 2: s^2 is an unbiased estimator of S^2

Proof : We have

$$\begin{aligned} E(s^2) &= E\left(\frac{1}{n-1} \sum_{r=1}^{n} (y_r - \bar{y})^2 \right) \\ &= \frac{1}{n-1}\, E\left(\sum_{r=1}^{n} \{(y_r - \bar{Y}) - (\bar{y} - \bar{Y})\}^2 \right) \\ &= \frac{1}{n-1}\, E\left(\sum_{r=1}^{n} (y_r - \bar{Y})^2 - 2(\bar{y} - \bar{Y})\}^2 \sum_{r=1}^{n} (y_r - \bar{Y}) + n(\bar{y} - \bar{Y})^2 \right) \\ &= \frac{1}{n-1}\, E\left(\sum_{r=1}^{n} (y_r - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{r=1}^{n} E(y_r - \bar{Y})^2 - nE(\bar{y} - \bar{Y})^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{r=1}^{n} \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2 - nV(\bar{y}) \right) \\ &= \frac{1}{n-1}\left(n\,\frac{(N-1)}{N}\,S^2 - n\,\frac{(N-n)}{N}\,\frac{S^2}{n} \right) \\ &= 1 \left(S^2 \;\; (n^2 (N-1) - n (N-n)) \right. \end{aligned}$$

$$\overline{n-1} \quad \overline{nN}$$

$$= \frac{1}{n-1}\ \frac{S^2}{nN}\ nN(n-1)$$

$$= S^2$$

<u>Example</u> :

In a population with N = 5, the values of y are 7,1,10, 3, 9. Calculate the sample mean $\bar{y}$ and the sample variance s^2 for all simple random samples (SRSWOR) of size 2 and verify

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

(ii) $V(\bar{y}) = \frac{N-h}{N}\frac{S2}{n}$

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

<u>Sol:</u> We have $\binom{5}{2}$ = 10 samples for which,

<u>Sample</u>	$\bar{y}$	<u>s^2</u>
(7, 1)	4.0	18.00
(7, 10)	8.0	4.50
(7, 3)	8.5	8.00
(7, 9)	8.0	2.00
(7, 10)	5.5	40.50
(1, 3)	2.0	2.00
(1, 9)	5.0	32.00
(10, 3)	6.5	24.00
(10, 9)	9.5	0.50
(3, 9)	6.0	18.00
Sum	60.0	150.00
Explanation	6.0	15.0

We have $\bar{Y} = \frac{30}{5} = 6.0$

$$S^2 = \frac{1}{4}$$

$$\frac{N-n}{N}\ \frac{S2}{n} = \frac{5-2}{5}\ \frac{15}{2} = \frac{45}{10} = 4.5$$

$$V(\bar{y}) = \frac{1}{10}$$

<u>Case II:</u> **<u>Simple random sampling with replacement (SRSWR)</u>**

<u>Theorem</u> 3: (i) $E(\bar{y}) = \bar{Y}$

(ii) $V(\bar{y}) = \frac{\sigma^2}{n} = \frac{N-1}{N}\frac{S^2}{n}$

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

We have,

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{1}^{n} y_r\right) = \frac{1}{n}\sum_{r=1}^{n} E(y_r) = \frac{1}{n}\sum_{r=1}^{n}\bar{Y} = \bar{Y}$$

Also we have,

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

$$= \frac{1}{n^2} E\left(\sum_{r=1}^{n}(y_r - \bar{Y})\right)^2$$

$$= \frac{1}{n^2}\sum_{r=1}^{n} E(y_r - \bar{Y})^2$$

Since y_r are independent

$$V(\bar{y}) = \frac{1}{n^2}\sum_{r=1}^{n}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2\right)$$

$$= \frac{1}{n^2} n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

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

Again,

$$E(s^2) = \frac{1}{n-1} E\left(\sum_{r=1}^{n}(y_r - \bar{y})^2\right)$$

$$= \frac{1}{n-1}\sum_{r=1}^{n} E(y_r - \bar{y})^2$$

$$= \frac{1}{n-1}\sum_{r=1}^{n} E\{(y_r - \bar{Y}) - (\bar{y} - \bar{Y})\}^2$$

$$= \frac{1}{n-1}\sum_{r=1}^{n}\{E(y_r - \bar{Y})^2 - E(\bar{y} - \bar{Y})^2\}$$

$$= 1 \left(\sum^{n} \underline{1} \sum^{N} (y_i - \bar{Y})^2 - nV(\bar{y})\right)$$

$$\overline{\text{n-1}} \qquad {}^{r=1}\, N \; {}^{i=1}$$

$$= \frac{1}{n-1}\left(n\,\sigma^2 \; - \; n\,\frac{\sigma^2}{n} \right)$$

$$= \sigma^2$$

The Standard error of $\overline{y}$ is $\sigma_{\overline{y}} = \sigma/\sqrt{n}$ where unbiased estimator is $s/\sqrt{n}$

Corollary : The unbiased estimator of population total $\bar{Y}$ is Ny and its variance is (N-1)N $\frac{S^2}{n}$

11. ESTIMATION OF POPULATION PROPORTION

Sometimes, the characteristic under study is qualitative or an attribute and every unit in the population are classified into one of two classes viz, (a) those possessing the attribute and (b) those not possessing the attribute. Suppose M out of N units in the population possess the attribute, So that the population proportion is P = M/N. We and interested in estimating P on the basis of a simple random sample of size n.

Let the number of units in the sample which posses the attribute be m and, consequently, the sample proportion is p = m/n.

<u>Theorem</u> 4: For SRSWOR, p is the unbiased estimator of P and its variance is given by

$$V\,(p) \;=\; \frac{N-n}{N-1} \qquad \frac{PQ}{n} \qquad \text{where } Q = 1 - P$$

<u>Proof</u> : We know that m has the hypergeometric distribution given by p.m.f.

$$f(x) = P(m{=}x) = \frac{\left(\frac{M}{N}\right)\;\left(\frac{N-M}{n-x}\right)}{\left(\frac{N}{x}\right)} \quad ; \; x = 0, 1, \ldots, n$$

Whose mean and variance are known to be

$$E\,(m) \;=\; n\,\left(\frac{M}{N}\right) \;=\; n\,P$$

$$V\,(m) \;=\; n\,(\frac{N-n)}{N-1} \qquad \frac{M}{N} \quad (1 - \frac{M}{N})$$

$$= \frac{n(N-n)}{N-1} \quad PQ$$

Therefore,

$$E\,(p) = E\,\left(\frac{m}{n}\right) = P$$

$$V\,(p) = V\,\left(\frac{m}{n}\right) = \frac{N-n}{N-1} \quad \frac{DQ}{n}$$

<u>Alternative Proof</u>: Suppose we associate a measurement y to each population unit such that

$$y_i = \begin{cases} 1 & \text{if the unit } U_i \text{ possesses the attribute} \\ 0 & \text{otherwise} \end{cases}$$

Then we have the population values $y_1, \ldots, y_N$ and sample values $y_1, \ldots, y_n$ such that

$$P = \sum_1^N y_i / N \; (= M/N)$$

$$= \bar{Y}$$

$$p = \sum_1^n y_r / n \; (= m/n)$$

$$= \bar{y}$$

$$S^2 = \sum_1^N \frac{(y_i - \bar{Y})^2}{N-1} = \sum_1^N \frac{y_i^2 - N(\bar{Y})^2}{N-1}$$

$$= \frac{NP - NP^2}{N-1} = \frac{NPQ}{N-1}$$

$$s^2 = \frac{npq}{n-1}$$

Therefore, by Theorem 1 we get

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

$$V\,(p) = V\,(\bar{y}) = \frac{N-n}{N} \quad \frac{S}{n} = \frac{N-n}{N-1} \quad \frac{PQ}{n}$$

<u>Theorem</u> 5: An unbiased estimator of V (p) is

$$v\ (p)\ =\ \frac{N-n}{(n-1)N}\ pq$$

Proof : Since E (s^2) = S^2, therefore

$$E\ \left(\frac{npq}{n-1}\right)\ =\ \frac{NPQ}{N-1}$$

$$\text{or}\ \ E\ (pq)\ =\ \frac{n-1}{N-1}\ .\ \frac{N}{n}\ PQ$$

So that

$$E\ (v(p))\ =\ \frac{N-n}{(n-1)N}\ E\ (pq)$$

$$=\ \frac{N-n}{N-1}\ \frac{PQ}{n}$$

$$=\ V\ (p)$$

Theorem 6: For SRSWOR, p is an unbiased estimator of P and its variance is given by

$$V\ (p)\ =\ \frac{PQ}{n}$$

Its unbiased estimator is $v(p)\ =\ \dfrac{pq}{n-1}$

Proof : In this case, p has binomial distribution given by p.m.f.

$$f\ (x)\ =\ P\ (m=x)\ =\ \binom{N}{x}\ \left(\frac{M}{N}\right)^{x}\ \left(\frac{1-M}{N}\right)^{n-x}$$

$$=\ \binom{n}{x}\ P^x\ (1-P)^{n-x}$$

whose mean and variance are given by

$$E\ (m) = nP$$
$$V\ (m) = nPQ$$

Therefore,

$$E\ (p) = E\ \left(\frac{m}{n}\right)\ =\ P$$

$$\text{and}\quad V\ (p) = V\ \left(\frac{m}{n}\right)\ =\ \frac{PQ}{n}$$

Alternative proof: It follows from Theorem 3 that

$$E(p) = E(\overline{y}) = \overline{Y} = P$$

$$V(p) = V(\overline{y}) = \sigma^2 / n$$

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

$$= \frac{N-1}{N} \frac{NPQ}{(N-1)n}$$

$$= \frac{PQ}{n}$$

Since $E(s^2) = \sigma^2$, therefore

$$E\left(\frac{npq}{n-1}\right) = PQ$$

or $$E(pq) = PQ\left(\frac{n-1}{n}\right)$$

So that

$$E(v(p)) = E\left(\frac{pq}{n-1}\right) = \frac{PQ}{n} = V(p)$$

CONFIDENCE LIMITS FOR POPULATION MEAN AND PROPORTION

If the sample size is not too small and N is passage, the sample mean will be approximately normally distributed i.e.

$$\bar{y} \sim \mathrm{N}\left(\bar{Y}\ ,\ \sqrt{\frac{N-n}{N}}\ .\ \frac{S}{\sqrt{n}}\right)$$

If $Z_{\alpha/2}$ is the $\alpha/2$ % point of the standard normal distribution, then

$$\mathrm{P}\left(|\ \bar{y} - \bar{Y}\ | > Z_{\alpha/2}\sqrt{\frac{N-n}{N}}.\ \frac{S}{\sqrt{n}}\right) = \alpha$$

or, $\mathrm{P}\left(\bar{y} - Z_{\alpha/2}\sqrt{\frac{N-n}{N}}.\ \frac{S}{\sqrt{n}} \le \bar{Y} \le \bar{y} + Z_{\alpha/2}\sqrt{\frac{N-n}{N}}.\ \frac{S}{\sqrt{n}}\right) =$ 1 - α

which gives the (1-α)% confidence interval for the population mean $\bar{Y}$ as

$$\left(\bar{y} - Z_{\alpha/2}\sqrt{\frac{N-n}{N}}.\ \frac{S}{\sqrt{n}}\ ,\ \bar{y} + Z_{\alpha/2}\sqrt{\frac{N-n}{N}}.\ \frac{S}{\sqrt{n}}\right)$$

when S^2 is not known we use its estimate s^2 and note that $\dfrac{\bar{y} - \bar{Y}}{\sqrt{\frac{N-n}{N}}\ \frac{s}{\sqrt{n}}}$ has t-dist on (n-1)d.f. Therefore, the (1-α)% confidence interval for $\bar{Y}$ will be

$$\left(\bar{y} - t_{n-1,\alpha/2}\sqrt{\frac{N-n}{N}}.\ \frac{s}{\sqrt{n}}\ ,\ \bar{y} + t_{n-1,\alpha/2}\sqrt{\frac{N-n}{N}}.\ \frac{s}{\sqrt{n}}\right)$$

where, $t_{n-1,\alpha/2}$ is the $\alpha/2$ % point of $t_{n-1.}$

To get the confidence interval for population proportion P we note that, for large samples

$$p \sim N\left(P,\ \sqrt{\frac{N-n}{N}\frac{PQ}{n}}\right)$$

$\alpha/2$ % Therefore, $(1-\alpha)$% confidence interval for P is

$$\left(p\ -\ Z_{\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n}}\ ,\ p+\ Z_{\alpha/2}\sqrt{\frac{N-n}{N}\frac{pq}{n}}\right)$$