*CHAPTER 2*

SIMPLE RANDOM SAMPLING

**2.1 INTRODUCTION**

Random Sampling or more precisely, Simple Random Sampling is a term covering two of the most straightforward selection procedures used in probability sampling. In both these selection procedures population units are drawn with equal probabilities. If a unit selected once is not allowed to be selected again, the procedure is known as simple random sampling without replacement (srswor). If the selection at each draw is from the whole population, the procedure is known as simple random sampling with replacement (srswr). The selection of units using simple random sampling with replacement is independent from draw to draw and the selection of units using simple random sampling without replacement is dependent. This is because in simple random sampling without replacement probability of selection of a population unit at any given draw depends on whether or not another unit has been selected at some previous draw. **The selection procedure of simple random sampling with replacement is also known as unrestricted random sampling.**

**THEOREM (2.1)**

In simple random sampling with and without replacement the probability that any given population unit is selected at any particular draw is 1/N, where N denotes the number of units in the population. **◊**

**PROOF**

For simple random sampling with replacement the theorem follows immediately from the independence and identical nature of a draw. If first unit is selected, its probability of selection is . The unit is replaced before second unit is drawn. In the second draw, the total population units is again N. The probability of selection of any unit at second draw is again . Thus for each draw the probability of selection is .

For simple random sampling without replacement it is necessary to consider each draw individually.

Clearly each population unit has the probability 1/N of being selected at the first draw. To select a unit at the second draw, however, it is necessary that unit is not selected at the first draw. The probability that the unit is not selected at the first draw is . The probability of selection of aat the second draw conditional on its non-selection at the first draw is 1/(N-1), so the unconditional probability of selection at the second draw is the product of probability that it is not selected at the first draw and is selected at the second draw . This argument can be easily generalized to the kth draw, for which the unconditional selection probability is:

.**◊**

**ALTERNATIVE PROOF**

This may be proved by another way. Suppose that the population unit say ith is selected at the first draw, the probability of selection of this unit is 1/N. The conditional probability that the second selection will produce jth (j ≠ i) unit is 1/(N-1) as there will be N-1 units in the population. The chance of getting the jth unit at the first selection is 1/N. Hence the absolute probability that ith unit is selected in the sample at the second draw is “the sum of the product that the jth unit is selected at the first and ith unit is selected at the second draw” i.e.

. **◊**

This result may be illustrated by a simple example. Consider a population of three units; A, B and C. Ignoring the order of selection, there are three possible samples of two distinct units. These sample are {A B}, {B C} and {C A}. Each of these sample can be selected in two ways. Thus {A B} may be achieved by selecting A at the first draw and B at the second, or B at the first draw and A at the second. If sampling is without replacement, the probability of selecting A at the first draw is 1/3, and the conditional probability of selecting B at the second draw given that A was selected at the first is 1/2. Thus the probability of selecting A first and B second is 1/6. Similarly the probability of selecting B first and A second is 1/6. Hence the probability of selecting the sample {A B}is:   
. The same is true for {B C} and {A C}.

A number of corollaries may be drawn from Theorem (2.1).

**COROLLARY (2.1.1)**

Suppose simple random sample of size n is drawn without replacement from a population of N units. The probability that any specified unit is in the sample is n/N. This may be proved by simple argument. The probability that the ith unit may either is in the sample at the first draw, or at the second draw, or at the third draw, and so on is 1/N. There are n possible draws, so that probability of inclusion of the ith unit to be in the sample is the sum of the probabilities that it is selected at the first or at the second or at the third draw and so on i.e. **.**◊**

Note: The ratio, n/N, plays an important role in simple random sampling theory. It is known as the *sampling fraction* and denoted by f.

This probability is generally called the probability of inclusion of the ith unit to be in the sample

**COROLLARY (2.1.2)**

Suppose a Simple Random Sampling of size n is drawn without replacement from a population of N. The member of possible samples using simple random sampling without replacement is . The proof is trivial.

**COROLLARY (2.1.3)**

Suppose a Simple Random Sampling of size n is drawn without replacement from a population of N. The probability of selection of a sample of size n is. Each of these  possible samples occurs with the same probability i.e. . This may be proved as:

**PROOF**

The probability that a specified unit is in the sample is n/N, and the probability that another specified unit from the remaining N-1 units is in the sample is (n-1)/(N-1). Hence the probability that any particular sample  of size n is selected from a population of N units is

 =**◊**

**COROLLARY (2.1.4)**

The number of times the ith (or any other) population unit appears in the set of all these possible samples is .

The last three of these corollaries do not have any direct analogues in simple random sampling with replacement, because of the possibility of multiple selections. Consider for instance the nine possible samples of two units selected with replacement from the population of three units A, B, C. These are

AA, BA, CA, AB, BB, CB, AC, BC, CC,

where, for instance, CA denotes the sample ‘C selected first, then A’. With simple random sampling with replacement each of these nine samples are equally probable, but the sample **{A B} can be formed in two ways and thus has twice the probability of occurrence of A A (the only sample in which the unit A is selected twice).**

It is nevertheless possible to formulate corollaries for simple random sampling with replacement in terms of (ordered) outcomes rather than in terms of (unordered) samples.

**COROLLARY (2.1.5)**

There are possible samples using simple random sampling with replacement. Each of these samples occur with probability and every units appears times in these samples.

There are more possible samples using simple random sampling with replacement than there are using simple random sampling without replacement. Suppose, for instance, that we have a population of three units, A, B and C, and we select a sample of two units from the population. Using simple random sampling without replacement there are six possible samples, which are equally likely:

A selected at the first draw, B at the second draw

A selected at the first draw, C at the second draw

B selected at the first draw, A at the second draw

B selected at the first draw, C at the second draw

C selected at the first draw, A at the second draw

C selected at the first draw, B at the second draw

We therefore, say there are six possible *ordered samples* AB, AC, BA, BC, CA and CB. But if we ignore the order in which the sample units are selected, the ordered samples AB and BA can be put together as the *unordered sample* AB, and so on. There are, then, three equally likely unordered samples, all, namely AB, AC and BC. It is a common practice to ignore orderings and to describe the unordered sample simply as *sample*. Since the sample AB, AC and BC are equally likely; each same probability of selection.

With simple random sampling with replacement (srswr), however, there are nine possible equally likely unordered samples, namely AA, AB, AC, BA, BB, BC, CA, CB and CC. If ordering is ignored, we have six possible unordered samples, AA, AB, AC, BB, BC and CC, but now the samples containing repeated selections (AA, BB and CC) are made up of only a single ordered sample while the others (AB, AC and BC) are each made up of two ordered samples, and we have twice the probability of selection. AA, BB and CC have 1/9 probability of selection , while AB, BC, and CA each has 2/9 probability of selection. Further, it is a common practice to ignore repeats (as well as to ignore orderings so the six samples possible under simple random sampling with replacement would be written as AB, AC and BC (each with probability 2/9) and AA, BB and CC (each with probability1/9).

More generally, when n sample units are selected from N population units, simple random sampling without replacement yields



equally likely ordered samples and



equally likely unordered samples. There are n! ordered samples for each unordered sample.

Simple random sampling with replacement, however yields Nn equally likely unordered samples of which only N!/(N-n)! are of size n and the remainder have smaller sample sizes. However, when n is small compared to N, the average sample size will be smaller than n.

The reason for the common practice of ignoring repeats is that when a population unit appears in sample more than once it supplies no additional information about the population than it does when it appears only once. Orderings are similarly ignored on the grounds that the order in which a sample is selected supplies no information about the population in addition to that supplied by the values of the sample variables. The numbers of distinct population units in sample is commonly called the *sample size*, so the samples AB, AC and BC are called samples of size two units while the samples AA, BB and CC (or A, B and C) are called samples of size one unit.

There are two good reasons for preferring this sample selection procedure, which always give samples of the maximum possible size. The first is the obvious one that large samples contain more information about the population than small samples, and this affects the precision of the sample estimates. The other relates to the allocation of resources; it is nice to know before sample selection, exactly how many sample units need to be collected from! Against this, however, samples selected with replacement have certain advantages in simplicity which are sometimes important enough to outweigh the two objectives just described. Because this simplicity can be quite useful at times, the current practice of ignoring orderings and repeats will not always be followed in this book. This decision will allow sampling with replacement to be examined alongside sampling without replacement and enable the student to see exactly what their advantages and disadvantages are

Two important quantities, which characterize any sample selection procedure, are the first and second order inclusion probabilities of the population units. The first order inclusion probability of a population unit is the ordinary probability of its inclusion in the sample. The second order inclusion probability of a pair of units is the joint probability of their inclusion in the sample. The first order inclusion probabilities determine the expectations (weighted averages) of estimates based on the sample; the second order probabilities determine the reliability of these estimates.

For simple random sampling without replacement (srswor) these probabilities are clearly defined and fairly readily calculable. The *first order inclusion probability* of each population unit is the same. The probability of (selection of a unit) at the first draw is 1/N, at the second draw 1/N, and so on at the n-th draw it is 1/N. But it cannot be selected more than once, so its total inclusion probability is the sum of these, which is n/N.

For simple random sampling without replacement (srswor), the joint probability that a given population unit, say ith unit, is selected at the first draw is 1/N, and the conditional probability that any other population unit say jth unit is selected at the second draw given that the ith unit was selected at first, denoting P(j|i),is clearly 1/(N-1). So the joint probability that the ith unit is selected at the first draw and jth unit at the second is 1/N(N-1). By the same type of argument that was used to show that the probability of selection of any unit at any draw is 1/N, it can also be shown that the joint probability of selection of any two distinct population units at any two distinct draws in a given order is also 1/N(N-1). There are two possible orderings (ith first then jth, or the reverse) so the joint probability of selection of any two distinct draws in either order is 2/N(N-1). But there are n(n-1)/2 pairs of distinct draws (ignoring orderings) so the second order inclusion probability of any pair of distinct population units is



When we consider the concept of first and second order inclusion probabilities for simple random sampling with replacement, we have to make a choice whether to ignore repeats or not. If we choose to ignore repeats, a population unit’s first order inclusion probability is the probability of selecting it in sample at least once, which is one minus the probability of never selecting it at all, that is . When n is small compared to N, this expression is approximately equal to the simple random sampling without replacement value n/N, but it is in fact always somewhat smaller than n/N. The second order inclusion probability of any pair of distinct population units, ith and jth, is one minus the probability of not selecting ith, minus the probability of not selecting jth plus the probability of not selecting either ith or jth, that is



When n is small compared to N, this is roughly n2/N2, but once again it is always smaller than the simple random sampling without replacement value .

If, however, we choose not to ignore repeats it is necessary to move away from the strict notion of inclusion probabilities and consider instead the expected number of times a population unit appears in a sample and the expected number of times a pair of population units appear in a sample. When repeats are impossible, as for simple random sampling without replacement, the expected number of times a population unit appears in sample is its first order inclusion probability, and the expected number of times a pair of distinct population units appears in sample is that pair’s second order inclusion probability, so this new concept is really a generalization of the old one.

The expected number of times a given population unit appears in sample under srswr is



=

which is exactly same as the simple random sampling without replacement value.

The number of times a given pair of distinct units appears in a sample under simple random sampling with replacement is defined as the product of the times each such pair of units appears in sample in any ordering. A sample of n units (not necessarily all distinct population units) is have n(n-1) such appearances. These will be of two types. The most common type is that in which the population units appearing at two sample draws are different, but occasionally they will be the same for any given population unit. The expected number of sample unit pairs in which that population unit appears twice is:



There are N such population units, so an expected number of pairs of sample units is n(n-1)/N2 and n(n-1) units are paired in which the same population unit appears twice, leaving n(n-1)(N-1)/N in which different population units appear. There are N(N-1) pairs of different population units (regarding {i, j} as different from {j, i}, so the expected number of times any such pair appears is n(n-1)/N2.

Note that although the expected number of appearances of {i, j} is the same as the expected number of appearances of ith twice – both being n(n-1)/N2 – the number of times i and j appear in either order is 2n(n-1)/N2.

The example used earlier where two units were selected out of a population of three can be used to illustrate the procedure. Sampling is with replacement, so the nine possible samples are

AA AB AC BA BB BC

CA CB CC.

In this array, the first letter in each pair represents the population unit selected at the first draw and the second letter the one selected at the second draw.

The pairs of sample units are now considered in both possible orderings. The sample AA therefore gives rise to n(n-1) or two identical pairs, AA and AA, but the sample AB gives rise to two different pairs AB and BA. The eighteen pairs from the nine samples are

AA, AB, BA, AC, CA, BA, AB, BB,

BC, CB, CA, AC, CB, BC, CC, .

Each of the possible pairs appears twice [that is n(n-1) times] in nine (that is N2) samples. So the expected number of appearances per sample is 2/9 [that is n(n-1)/N2]. But if the population units are distinct and the order of appearance is ignored (so that the pair AB is regarded as equivalent to the pair BA) the expected number of appearances in sample is 4/9 (that is 2n(n-1)/N2.

**2.2 NOTATION**

The following notation will be used in the reminder of this chapter. Note that capital letters are used for population values and lower case letters for sample values unless otherwise specified.

 The value of the variable of ith population unit

 The value of the variable of the ith sample unit

 The population mean, 

 The sample mean,  used as an estimator   
 of the population mean, 

 The population total, 

 An estimator, of the population total, 

E An operator denoting expectation

 An operator denoting variance

 An operator denoting an estimator of variance

SE An operator denoting standard error

CL An operator denoting confidence limits

 

 

**2.3 EXPECTATION OF SAMPLE MEAN**

The expectation of a random variable was defined in Chapter 1 as the sum of the products of the probabilities of a set of exhaustive and mutually exclusive events. In survey sampling, the relevant events are the selections of population units to be in the sample and the relevant values taken are the estimand (or estimated variable) values associated with the selected units. In simple random sampling both with and without replacement, the selection probabilities are all equal and the expectation is an arithmetic mean.

**THEOREM (2.2)**

In simple random sampling with replacement and without replacement, the sample mean, , is an unbiased estimator of the population mean .**◊**

**PROOF**

We select a random sample of n units from a population of N units and observe variable yi on the ith sample unit. An obvious estimator of the population mean  is

. (2.3.1)

An estimator of population total  is

. (2.3.2)

The estimator of  is unbiased estimate of  when n is large. That is to say, if we select a very large number of samples, the average value of s over all those samples would tend to . We express this formally as:



The expectation of , by definition is,

 = 

But, from Theorem (2.1), for both srswor and srswr,

,

where  (each unit has an equal probability of inclusion in the sample). Hence .**◊**

**ALTERNATIVE PROOF**

Since there are  possible distinct samples for without replacement, then

 (2.3.3)

Now in  possible samples, each unit is appearing  times, hence

*.***◊**

Similarly estimated total may be proved as an unbiased estimator of population total i.e. .

The *Bias* of an estimator is defined as the difference between the expectation of the estimator and the true value of the quantity estimated. Hence if  is used as an estimator of and the bias is zero, then  is described as *Unbiased Estimator* of. Unbiasedness is just one criterion of an estimator’s suitability, and not the most compelling one. It is easy to construct unbiased estimators, but sometimes unsuitable for the purpose of estimation. Suppose, for instance, that an artificial variable is defined on the Ui for which the sum is zero. One such artificial variable, which is denoted by Ai, takes the value 1 for i = 1, 2,… N-1 and the value – (N-1) for i = N. Clearly the population total of the Ai, denoted by A, is zero. If  is the sample mean of the ai,formed by analogy with  then  is non – zero for all possible samples of n < N. Nevertheless  is an unbiased estimator of the population mean , which is zero. Hence all estimators of the form where c is an arbitrary constant are unbiased estimators of . If c is now allowed to become indefinitely large,  becomes inappropriate for , despite its being unbiased. To understand in what other ways  can be considered a suitable estimator for , it is necessary to consider its variance (or, since its bias is non-zero, equivalently to consider its mean squared error).

* 1. **VARIANCE OF THE SAMPLE MEAN AND ESTIMATED TOTAL**

**THEOREM (2.3)**

Suppose a simple random sample of n units is drawn from a finite population of N units. The variance of the sample mean for simple random sampling without replacement, , is

 (2.4.1)

and for simple random sampling with replacement, , is

**◊** (2.4.2)

**PROOF**

We know that

,

The variance of this estimator is



Since the variance of the sum of the random variables is equal to the sum of the variance of random variables plus the sum of the covariance, we have as:

. (2.4.3)

Now regardless of whether sampling is with or without replacement,

 

 . (2.4.4)

(Note that  or  is some times described as the *Population Variance* of yi.)

Further, for simple random sampling without replacement,



  (2.4.5)

Substituting (2.4.4) and (2.4.5) in (2.4.3) we get



 **◊**  (2.4.1)

For simple random sampling with replacement, however, all selections are independent, therefore  is zero. Hence (2.4.3) reduces to

 (2.4.6)

Substituting (2.4.4) in (2.4.6) we obtain (2.4.2) for simple random sampling with replacement.**◊**

The variance expression for sample mean and total for simple random sampling without replacement are also written as:-

 (2.4.7)

and

 . (2.4.8)

We see how (2.4.8) is built up. If there were only one unit in sample, the variance of the sample mean (that is, of the single sample observation) would be the population variance, near enough . Increasing the sample size from one unit to n units brings the variance of the sample mean down by a factor n. The estimated population total is N times the sample mean, hence the factor N2 in the formula for the variance. Finally, the fact that each unit in sample is constrained to be a different population unit means that when n = N the estimated population total must be precisely the true population total (we are ignoring all errors other than sampling error) and this has zero variance. This explains the presence of the *finite population correction* .

An analogous expression for the covariance of the unbiased estimators  and  in case of simple random sampling without replacement may be written in a straightforward way

 (2.4.9)

We see how (2.4.1) and (2.4.2) are related. Consider first the case when the sample consists of a single unit. There is then no distinction between simple random sampling with replacement and simple random sampling without replacement and the variance of  is the population variance (2.4.4). It is also the special case of (2.4.1) and (2.4.2).

When  the variance of wor is smaller than the variance of wr by the factor. Equivalently, we note that for simple random sampling without replacement the expression S2/n in the variance of  is multiplied by (1-f), which decreases as n increases, whereas for simple random sampling with replacement the same expression is always multiplied by  which does not decrease as n increases.

Another way of looking at (2.4.2) is that  in to say that the simple random sampling with replacement is same as in simple random sampling without replacement value.

People with a limited knowledge of survey sampling are sometimes under the impression that the accuracy of a sample estimator is primarily a function of the sample fraction rather than of the sample size. They would have more respect for results obtained from a sample of 5,000 out of 50,000 than they would if the results come from a sample of 20,000 out of 1,000,000, because the former is a 10 percent sample and the latter is only a 2 percent sample. Applying (2.4.1), however, we easily see that the absolute size of the sample is almost always important criterion, except when the fpc is quite large – say, greater than 0.05.

The most important point to note about the fpc is that of  is less than 2%. It reduces the variance of  by less than 20% and its standard error by less than 10% unless the fraction of population in sample is quite large, the accuracy of  entirely depend on the absolute size of the sample, n, and hardly at all on the sampling fraction. Hence, if we are comparing two populations, one containing 1,000,000 units and the other 100,000,000 units, so that any reasonably sized sample would be only a small fraction of either. We would naturally think of choosing the same sized sample from each population, and do not worry that we have selected a much thinner sample from the large population than from the smaller one.

**2.4.1 Alternative Proof Using Indicator Variable**

The following alternative proof uses *Indicator Variables*. The use of indicator variable is particularly convenient in unequal probability sampling (ups) or probability proportional to size sampling (pps). Readers who intend to master the topic of pps sampling are advised to familiarize themselves with its use in simple random sampling first.

Indicator variables are random variables, which indicate the number of times a population unit is selected in the sample. Formally the indicator variable is denoted by ai, where

ai  = ni if ith unit of the population is in sample ni times.

In simple random sampling without replacement  takes the value 1 with probability f=n/N and 0 with probability. Further, the joint probability that ai = 1 and aj = 1 is n(n-1)/N(N-1), for simple random sampling without replacement.

Now

 (2.4.10)

, (2.4.11)

and



 . (2.4.12)

In simple random sampling with replacement, however, have the multinomial distribution that arises wherever there are n independent trials having N equi-probable outcomes. Here we have

 (2.4.13)

 (2.4.14)

and

 (2.4.15)

For both simple random sampling without replacement and simple random sampling with replacement the sample means can be written using indicator variable notation as:

 (2.4.16)

Note that in (2.4.16) the ai(s) are random variables and the Yi (s) are fixed values. The expectation and variance of may therefore be written

 (2.4.17)

and

 (2.4.18)

For simple random sampling without replacement substitute the value of  from (2.4.10) in (2.4.17)

, (2.3.1)

which proves the unbiasedness of .

Substituting the value of  and  from (2.4.11) and (2.4.12) in (2.4.18) we get



 **◊**  (2.4.1)

For simple random sampling with replacement, substituting (2.4.13) in (2.4.17), we get



Again substituting (2.4.14) and (2.4.15) in (2.4.16) we have for simple random sampling with replacement



 **◊** (2.4.2)

The variance of the estimated total  for both simple random sampling with and without replacement is

 (2.4.19)

**2.5 ESTIMATOR OF THE VARIANCE OF THE SAMPLE MEAN**

The variance expressions (2.4.1) and (2.4.2) contain the expression S2, which is a function of all the population values Yi. Since only those values, which relate to the units in the sample, are known, We need to estimate variances using only those values that fall in the sample. We therefore have to have an estimator of the variance of the estimator of population mean.

**THEOREM (2.4)**

Suppose a simple random sample of n units is drawn from a finite population of N units. In unbiased estimator of the variance of the sample mean, , is

 (2.5.1)

if sampling is without replacement and

 (2.5.2)

if sampling is with replacement.**◊**

The proof of this theorem requires the following lemma:

**LEMMA (2.4.1)**

For simple random sampling without replacement, s2 is an unbiased estimator of S2 and for simple random sampling with replacement s2 is an unbiased estimator of S2 (N-1)/N.**◊**

# PROOF

For both simple random sampling without replacement and simple random sampling with replacement, we have

 







. (2.5.3)

Taking the expectation of (2.5.3) we have

.

Since from (2.4.4) we know that therefore 

Now  is the variance of the sample mean, which is  for simple random sampling without replacement and is  for simple random sampling with replacement. Hence for simple random sampling without replacement

 (2.5.4)

and for simple random sampling with replacement

 **◊** (2.5.5)

The Theorem 2.4 can be proved as under:

**2.5.1 Proof of Theorem 2.4**

For simple random sampling without replacement, the expectation of (2.5.1) is

**◊**

For simple random sampling with replacement, the expectation of (2.5.2) is

**◊**

**2.5.2 Alternate Expression of Variance for Sample mean**

s2 may be written alternatively as:

, (2.5.6)

and

. (2.5.7)

Consequently, the basic building block, which go to make up s2 is the comparisons between all pairs of sample units. The same is true of S2, which can be written as:

 (2.5.8)

Also

. (2.5.9)

This will be of considerable importance when we consider systematic sampling.

**2.6 STANDARD ERRORS AND CONFIDENCE LIMITS**

**FOR THE POPULATION MEAN**

**2.6.1 Standard Error.**

Standard error, in fact, estimates the amount of sampling error which comes in the estimation process because of sampling process. The standard error of the sample mean is the square root of the sampling distribution of sample mean i.e., consequently the standard error of the sample mean  is



 (2.6.1)

**2.6.2 Confidence Limits of Sample Mean**

It is not possible for a sample to evaluate characteristics of a population exactly, but it estimates the characteristics as accurately as possible. One way out is to find an interval, which covers the parameter with some pre-assigned probability. In case the sample estimate is the arithmetic mean, ,and observations are normally distributed with known variance, the sampling distribution of sample mean, , is normal with mean  and variance given by in the expression (2.4.1) or (2.4.2). If samples are repeated, then the interval covers  in 95% of the cases. This can be stated as



which shows that the probability of  lying between



is 0.95%. This interval is called a *confidence interval* and the probability attached to it is known as *confidence coefficient*.

In general for any probability level (1- α) the confidence limits for sample meanwith unknown variance is

 (2.6.2)

and the confidence limits for estimated population total 

, (2.6.3)

where t is the t-probability point corresponding to the desired probability. Under fairly general conditions, for non-normal populations the sampling distribution of sample mean is approximately normal provided the sample size is sufficiently large and as such the procedure for setting up the confidence limits is the same as in case of normal population. Upper and lower confidence limits are random variables. The probability that it lies above its upper (1 – α) percent confidence limit is only α percent. Similarly the probability that it lies below its lower (1 – α) percent confidence limit is also α percent. Conventionally, the most used values of α are 0.01 and 0.05. If α = 0.05 we have 95 percent confidence limits and if α = 0.01 we have 99 percent confidence limits. On a repeated sampling basis we have 19 chances in 20 that a random variable will be observed to lie between its 95 percent confidence limits, and 99 chances in 100 that it will be observed to lie between its 99 percent confidence limits.

In order to estimate confidence interval for a population mean in terms of standard error or estimated standard error, it is necessary to make some assumptions about the sample mean. The simplest and most usual assumption is that this distribution is approximately normal. Whether it in fact is approximately normal depends both on the distribution of the Yi values in the original population and on the sample size n. since the variance of the Yi is necessarily finite, the central limit theorem ensures that under simple random sampling with replacement the distribution of  is asymptotically normal. The non-normality of this distribution (as measured by the ‘excess of kurtosis’ over the normal value of 3) diminishes as 1/n. But if the original population differs a great deal from normality, this may not be necessarily fast enough for practical purposes.

**Example 2.1:**

Draw all possible samples of size 2, with and without replacement from a hypothetical population of 4 units. Verify that;  for without replacement and  for with replacement. Also find the variance of  and compare the result with that obtained by formulas (2.4.1) and (2.4.2). The values of the population are: Yi= 2, 4, 6, 8.

## SOLUTION

Population mean  and S2 = 6.67

i) Possible samples for with replacement.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 2 | 4 | 6 | 8 |
| 2 | (2, 2) | (2, 4) | (2, 6) | (2, 8) |
| 4 | (4, 2) | (4, 4) | (4, 6) | (4, 8) |
| 6 | (6, 2) | (6, 4) | (6, 6) | (6, 8) |
| 8 | (8, 2) | (8, 4) | (8, 6) | (8, 8) |

The value of and  for all possible samples are

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|  | 2 | 3 | 4 | 5 | 3 | 4 | 5 | 6 | 4 | 5 | 6 | 7 | 5 | 6 | 7 | 8 |
|  | 0 | 2.0 | 8.0 | 18.0 | 2.0 | 0 | 2.0 | 8.0 | 8.0 | 2.0 | 0 | 2.0 | 18.0 | 8.0 | 2.0 | 0 |
| P | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 |

 and 





ii) For without replacement

Since N = 4; n = 2, the possible samples under without replacement sampling scheme is  All possible samples, mean and s2 are given as::

|  |  |  |  |
| --- | --- | --- | --- |
| Possible samples |  | **s2** | P(s) |
| 2,4 | 3.0 | 2.0 | 1/6 |
| 2,6 | 4.0 | 8.0 | 1/6 |
| 2,8 | 5.0 | 18.0 | 1/6 |
| 4,6 | 5.0 | 2.0 | 1/6 |
| 4,8 | 6.0 | 8.0 | 1/6 |
| 6,8 | 7.0 | 2.0 | 1/6 |
|  | 30.0 | 40.0 |  |





If we use expressions (2.4.1) and (2.4.2) we get 

,

and  for without replacement and with replacement respectively. The standard errors of sample mean for without replacement and with replacement respectively are:

 and .

**Example 2.2:**

A simple random sample of 33 mohallahs was taken from a population of 663 mohallahs (from the first report of 1973 population census of Libya). The population (male and female) of the sampled mohallahs is given as:

3485, 3085, 2049, 2952, 1269, 754, 449, 1722, 6907, 984, 3890, 2521, 1394, 4737, 4956, 3671, 766, 1505, 14901, 1111, 1362, 2136, 2200, 14123, 1347, 458, 1923, 2793, 18585, 22422, 842, 2575, 6341.

1. Estimate average population per mohallah, [the actual average population of 663 mohallah is 3404 persons].
2. Estimate the total population of Libya; [the actual population of Libya for 663 mohallahs is 2,257,088].
3. Find the  [the s2 of the population is 25,242,133.05]
4. Find confidence limits for sample mean and total.

## SOLUTION

For this example N = 663, n = 33, Σyi = 140251, (Σyi)2/n = 59,576,530.3



1. The average population per mohallah is 140251/33 = 4250.03.
2. Estimated total of 663 mohallahs 
3. 
4. Standard error of the mean = 934.818
5. Standard error of total is 
6. 



vi)  



Further three samples (5%, 10% and 20%) were randomly selected from 663 mohallahs of Libya. Means, estimated total and their standard errors are computed and are given as:

|  |  |  |  |
| --- | --- | --- | --- |
| Sample % | **5%** | **10%** | **20%** |
| Sample units | 33 | 66 | 132 |
| Average population per mohallah | 4250 | 3969 | 3331 |
| Estimated total | 2817750 | 2631512 | 2209000 |
|  | 934.82 | 688 | 374 |
|  | 619784.33 | 456064 | 247,735 |

One can say that as the size of the sample increases standard error of sample mean decreases. Consequently one can say that as the sample size increases, sampling error decreases.

**Example 2.3**

A population of Multan district of males and females of 523 villages is given in Appendix. The total, mean and variance of Male, Females and Total population are calculated and given as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex | N0.villages |  | Mean | s2 |
| Males | 523 | 937784 | 1793.29 | 3867810.5 |
| Females | 523 | 864058 | 1793.09 | 3281968.7 |
| Total | 523 | 1800841 | 3443.29 | 14253440.5 |

The samples of 5%, 7% and 10 % have been selected.. The villages selected on the basis of 5% are given below. But the results of the samples for 7 % and 10% are given only. Estimate the total population on the basis of these results and calculate the 95% confidence limits for population mean

**Solution**

The 5% sample using the random process has been selected and population for males, females and total of males and females is given and given on next page. The mean, variance and standard error for males, females and total are calculated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex |  |  | SE of () | s2 |
| Males | 55093 | 2118.96 | 754.86 | 14815215.5 |
| Females | 51017 | 1962.19 | 688.76 | 14333971.7 |
| Total | 105786 | 4068.69 | 1444.79 | 54272633.8 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Males** | **Females** | **Total** | **Males** | **Females** | **Total** |
| 260 | 228 | 488 | 718 | 658 | 1376 |
| 180 | 152 | 332 | 870 | 843 | 1713 |
| 1542 | 1348 | 2890 | 432 | 401 | 833 |
| 2987 | 2926 | 5913 | 281 | 228 | 509 |
| 3231 | 2934 | 6165 | 662 | 587 | 1249 |
| 994 | 857 | 1851 | 700 | 699 | 1399 |
| 18913 | 17155 | 36068 | 1850 | 1758 | 3608 |
| 1303 | 1182 | 2485 | 266 | 217 | 483 |
| 177 | 171 | 348 | 416 | 365 | 781 |
| 1299 | 1224 | 2523 | 692 | 631 | 1323 |
| 8625 | 8072 | 16697 | 3061 | 2883 | 5944 |
| 3164 | 3072 | 6236 | 668 | 660 | 1328 |
| 1502 | 1466 | 2968 | 300 | 300 | 276 |

95% confidence limits for males, females and total are as:

C.L ( = males) = 639.44 to 3598.48

C.L ( = females) = 612.22 to 3312.16

C.L ( = total) = 1179.11 to 6958.27

A sample of 7 % was taken from a population of 523 villages and the statistic calculated is given as:

|  |  |  |  |
| --- | --- | --- | --- |
| Sex |  |  | Standard Error of |
| Males | 60112 | 1624.65 | 321.81 |
| Females | 55302 | 1494.65 | 297.87 |
| Total | 115414 | 3119.30 | 619.60 |

95% confidence limits for males, females and total are as:

C.L ( = males) = 944.65 to 2255.4

C.L ( = females) = 910.82 to 2077.8

C.L ( = total) = 1904.88 to 4333.72

Another sample of 10% was selected from this population relevant information s:

|  |  |  |  |
| --- | --- | --- | --- |
| Sex | Total | Mean | Standard Error of |
| Males | 94615 | 1819.52 | 239.06 |
| Females | 86643 | 1666.21 | 222.86 |
| Total | 181258 | 3485.73 | 461.48 |

C.L ( = males) = 1332.96 to 2288.08

C.L ( = females) = 1229.71 to 2103.02

C.L ( = total) = 2581.23 to 4389.73

The estimated total its standard error and confidence limits are given as:

|  |  |  |  |
| --- | --- | --- | --- |
| Sample size % |  | SE of () | C.L of () |
| 5 | 2127925 | 755625.17 | 646980 - 3608950 |
| 7 | 1639394 | 324050.8 | 1004254 - 2274534 |
| 10 | 1823037 | 241354.04 | 1349983 - 2296090 |

Some important remarks regarding these calculations are:

1. All the confidence limits contain respective population mean.
2. As the sample sizes increases standard error decreases.

**Example 2.4:**

The sample mean,  is the best linear unbiased estimator of population mean based on a simple random sampling of size n.

**SOLUTION:**

Let us consider any linear unbiased estimator



The variance of t can be written as



Substituting the value of  and from (2.4.4) and (2.4.5) we get

 





Since,  therefore



The problem is to minimize  subject to . This function may be written as



Differentiating above equation w.r.t. “a” we get

= Constant = c or .

Since, therefore n c = 1 or c = 1/n or ai = 1/n

In other words ai is constant. The unbiasedness condition of E (t) =  shows that  Hence  and 

We say that the sample mean is the best linear unbiased estimator of population mean.

**2.7. ESTIMATION FOR PROPORTIONS**

Sometimes results are required in the form of proportions and percentages. In this section the estimation of proportion (say P) of units of population belonging to a special class is considered. In medical, biological and social sciences results are needed in the form of proportion and percentage. P may be defined as a proportion or a percentage in the population, e.g percentage of unemployed people in a certain country, proportion of villages having medical facilities, proportion of educated people or percentage of people suffering from a particular disease or proportion of smoker in some class. Suppose we have N population units i.e. Y1, Y2, …. Yi, …YN. yi = 1 if ith unit possesses a certain attribute and 0 otherwise. If some numbers (say A) are falling in a particular class, then the proportion of these units may be defined as P = A/N. Similarly if we have n sample units i.e. y1, y2, …. yi, …. yn with yi = 1 ith unit possesses a certain attribute 0 otherwise, then proportion of units possessing the attribute may be defined as p = a/n.

Using above definition, the value of Yi is either 1 or 0, then

 (2.7.1)

# Since Yi takes the value 1 or 0; also takes the value 1 or 0. Thus

 (2.7.2)

The value of S2 and s2 in terms of proportions may also be derived as:

We know 

Using (2.7.1) and (2.7.2), we have

 (2.7.3)

Similarly s2 = npq / (n – 1) (2.7.4)

**THEOREM (2.5)**

The sample proportions, p, is an unbiased estimator of population proportion P both for simple random sampling without replacement and simple random sampling with replacement.**◊**

The Proof is trivial.

**2.8. VARIANCE OF SAMPLE PROPORTION AND ESTIMATED TOTAL**

**THEOREM (2.6)**

A simple random sample of size n is drawn from a finite population of size N. The variance of the proportion p for simple random sampling without replacement is

 (2.8.1)

and for sampling with replacement is

**◊** (2.8.2)

**PROOF**

We know that

 (2.4.1)

Now for proportion we use (2.7.3) in (2.4.1),

**,**

and

 () = PQ/n **◊**

The unbiased variance estimators of  may also be obtained by using (2.7.4) on the similar lines:

 (2.8.3)

 (2.8.4)

If fpc is ignored and n is large then (2.8.3) and (2.8.4) are equal to

. (2.8.5)

It matters little that for large sample size n, whether to use (n – 1) or n in the denominator of expression (2.8.5). Hence we may use

. (2.8.6)

The estimated total A′ and Var(A′) may respectively be obtained as

A′ = Np and (2.8.7)

Var(A′) = N2 Var(p). (2.8.8)

**2.8.2. Standard Error and Confidence Limits of p**

The standard error of p and A′ (estimated total = NP) may be obtained as

 (2.8.9)

and

 (2.8.10)

and confidence limits for p and A′ are

C.L(p) = p +  S.E(p) (2.8.11)

and

C.L(A′) = A′ +  S.E(A′). (2.8.12)

Also if in presenting the results, the units are classified in more than two classes. Then

, , , .

If is constant, then the probability of drawing the sample is given by the multinomial expression

, (2.8.13)

which is an extension of the binomial distribution.

## EXAMPLE 2.5

In a simple random sample of 200 colleges, 120 colleges are in favor of a proposal, 57 opposed and 23 do not express their opinion. Estimate the total number of colleges in the population who are in favor of the proposal and find 95% confidence limits for the number of colleges, which favore the proposal. The total number of colleges in the population is 2000. (Source: Cochran 1977)

### SOLUTION

We have N = 2000 and n = 200. Since 120 colleges are in favor of a proposal, hence p = 120/200 = 0.6 and q = 0.4. The estimate of colleges who are in favor of the proposal from 2000 is Np = 0.6 x 2000 = 1200.

The standard error of the estimated total is



The confidence limits are

A′ + t S.E(A′) = 1200 + 1.96 x 69.2= { 1064, 1335 }.

One can say that these limits contain true population proportion with 95% probability.

### EXAMPLE 2.6

A survey is conducted in the Faculty of Science, El-Fateh University to estimate the proportion of smokers. A simple random sample of 122 students is taken from a population of 813 students. The total number of smokers in the sample was 42. Estimate the total number of smokers in the Faculty of Science, with 95% confidence limits. Among 122 students 86 were males and 36 were females and there was no female smoker. Find the proportion of smokers ignoring the female students. Also construct 95% confidence limits of this estimate. The total number of males in the faculty was 551.

### SOLUTION

Since there are 813 students and a sample of 122 is taken from this population, so   
N = 813, n = 122.

Out of 122 only 42 are smokers, hence the proportion of smokers is

p = 42/122 = 0.3443, q = 0.6557

Since the sample is large f.p.c. may be ignored. The standard error of p is:



The estimated total number of smokers is

Np = 813 x 0.3443 = 279.916 = 280.

The standard error of the estimated total (A′) is:

S.E(A′) = 813 x 0.043 = 34.96

95% confidence limits for the total are

280 + 1.96 x 34.96 = (211 to 349).

Now there is no female smoker, the proportion of smoker is

p = 42/86 = 0.4883, q = 0.5117 S.E(p) = 0.0539. The standard error of the total is

S.E(A′) = .0539 x 551 = 26.699.

The confidence limits for proportion are

0.4883 × 551 + 1.96 × 29.699 = (211 to 327).

**2.9 ESTIMATION OF SAMPLE SIZE.**

In a sample survey, it is always a problem for an experimenter to know or to estimate the size of the sample when the results are required with least sampling error. It is always a problem whether a sample should be 2%, 5%, or 10% or any other fraction. Although the sample size has always been a matter of choice with planners, yet great care and weight is needed in its estimation. Since sample is a part of the population, it should neither be too large to involve a lot of expenditure nor too small to make the result less reliable. In fact the sample size depends on the cost involved and precision required.

The estimation of sample size is illustrated as under:

Let ‘d’ be the margin of error with some probability α by which sampling value differs from population value. The permissible error, which is a percentage difference between the estimate and parameter value, is specified as:

 (2.9.1)

Since differs from sample to sample, the probability of the margin of error being less than d is given by

 (2.9.2)

α is usually chosen such that 1 - α is 90% or 95% or some other desired level. Let us assume that  is normally distributed with mean  then

 (2.9.3)

so that

. (2.9.4)

Now consider the case of sampling without replacement. Substituting  in (2.9.4) we have

. (2.9.5)

Solving for n, we have  (2.9.6)

For large N

, (2.9.7)

Then . (2.9.8)

If N is large (2.9.6) is identical to (2.9.8). Denoting n0 = (tS/d)2, as the first approximation, the second approximation is

 (2.9.9)

the third approximation may be found as

 (2.9.10)

and so on.

**Example 2.6**:

A physician would like to know the mean fasting blood glucose of patients seen in the diabetes clinic over the past 10 years. Determine the number of records the physician should examine in order to obtain 90% (and 95%) confidence level for population if the desired width of the interval is 8 units. A Pilot sample yields a standard deviation of 60 units.

**Solution:**

Here s = 60, d = 4, as the total width is 8 which is on the both sides of the mean. Therefore, the sample size for 90% confidence is, n =  = 609, and for 95 %, n =  = 864.

**EXAMPLE 2.7**

According to population census of Multan district (Appendix II) the variation of population males and females is 3867810.5. Suppose we like to hold a census on sample basis to estimate the population of Multan Dist. What would be the sample size, when the total number of villages in this district is 523? Use 95% probability confidence coefficient.

**SOLUTION:**

S2 = 3867810.5, N = 523.

The range of distribution of population is very large and a lot of variation is in the population of the village therefore we can take d = 0.2 (20% approximately). Using (2.9.6) we can start the approximation and then by iterative method we proceed as:

First approximation: = 523

Second approximation: 

Third approximation: 

This process is repeated and







Since the difference between last three approximations is constant so 44 is the optimal sample size.

**2.9.1 Sample Size for Coefficient of Variation**

In case of coefficient of variation, the sample size may be determined as follows:

For large N the coefficient of variation of sample mean is



Or  (2.9.11)

Or  (2.9.12)

Since  is coefficient of variance per unit, therefore,

 (2.9.13)

Note: if a sample of size n gives C.V = C1for a sample estimate, then the sample estimator with C.V. = C2, the required sample size is,

n1 = n(C1/C2)2 (2.9.14)

Comparing (2.9.13) and (2.9.14) we get

, (2.9.15)

where n is the size of the first sample.

##### EXAMPLE 2.8

Suppose it is believed that about 20% of persons in a large population suffer from a certain disease. How many persons should be selected in a random sample in order that the coefficient of variation of the estimate be 10% or less.

##### SOLUTION

We know that

 C.V = 1/10, since p = 0.2 and q = 0.8, 

hence the sample size is 400.

##### EXAMPLE 2.9

A simple random sample of 50 households is taken from a population of a certain village for estimation of expenditure on meat. The estimated average expenditure on meat per household turn out to be Rs. 0.88 with standard error of Rs. 0.10. Using this information find out what sample size is needed if similar type of survey is to be carried out in some other village with permissible 95% probability level is 10% of the true value.

##### SOLUTION

The permissible margin of error at 95% probability is 10% of the true value,

Hence,

0.1 0  = t S.E

S.E = 1.96 S.E(

S.E = 0.051 = C2

We assume that the C.V. is the same in both the villages.

The C.V. of the first village is 0.1/0.88 = 0.1136 = C1

Hence

n1 = n(C1­/C2)2 = 50(0.1136/0.051)2 = 248

**2.9.2 Sample Size for Proportion**

The determination of sample size in case of proportion may be found to as.

d = t S.E(p) (2.9.16)

Using (2.7.3) and (2.7.4) in (2.9.16) we obtain

 (2.9.17)

For large N

. (2.9.18)

Solving (2.9.17) and (2.9.18) for n we get,

 (2.9.19)

and

n = t2PQ/d2. (2.9.20)

respectively.

For practical purposes population values are replaced by sample values i.e. s for S and p for P at the respective places. Denoting n0 = t2PQ/d2 the second approximation from (2.9.19) may be

 (2.9.21)

and so on.

**2.9.2.1 *Sample Size for Absolute Precision***

**Example 2.10**:

The Ministry of Health wishes to estimate the prevalence of tuberculosis among children less than 5 years of age. How many children should be taken in the sample so that the prevalence may be estimated within 5% points of the true value with 95% confidence level, if it is known that the true rate does not exceed 15%.

**Solution:**

In thés exemple we have

p = 0.15, 1–p = 0.85

Probability level or confidence level (1 – α) = 95%.

d = 5 percentage points

 and 

Using the formula, we have

 = 196 for 1– α = 95%

If population is finite then an approximation of sample size can be obtained as   
n1 = . If the population of children less than 5 years of age is 20,000, then the sample size may be estimated, by an iteration, as:

 =  = 194

This is not different from 196, so 196 or 194 may be taken as a sample size.

**Example 2.11**

Ministry of Health would like to estimate the proportion of children who are receiving medical care regularly. How large should be the sample if the estimate falls within 5% of true proportion with 95% confidence level.

**Solution:**

In this question, the assumption regarding proportion of children who are receiving regularly medical care is that 50% of the population of children is receiving medical care. Using p = 0.50, maximum sample size will be obtained.

If we take

p = 0.5; 1 - α = 0.95, 0.99; d = 0.05 ,then

 = 384 for 95%

 = 666 for 99%

Suppose N = 600, then the sample size for 95% level turned out to be:

n1 =  =  = 234 (2nd approx.)

n2 =  =  = 169 (3rd approx.)

n3 =  =  = 132 (4th approx.)

This process will continue till difference between the last two approximations becomes minimal.

**2.9.2.2 *Sample size for relative precision***

If the coefficient of variation (or for relative precision) is given, the formula for determination of sample size is

n = , **(**2.9.22)

where D denotes coefficient of variation or relative precision.

For convenience, sample sizes have been calculated for different values of p and D. [see Tables at the end of this chapter]

**Example 2.12**

Ministry of Health of Eastern Province would like to conduct a survey regarding hypertension of elderly persons (above the age of 60). It is known from the past experience that the prevalence of hypertension is 25%. How large a sample should be so that the resulting estimates falls within 10% (not 10% points) of the true proportion with 95% confidence level?

**Solution:**

In this question p = 0.25, Confidence level = 95% and relative precision is 10% of 25%. There are two ways to solve this problem.

(i) Using relative precision formula

n =  = 4610

(ii) Using absolute precision formula

Since d = 0.05 x 0.25 = 0.0125

n =  = 4610

If population size is known to be 2000, then

n1 =  =  = 1395

n2 =  =  = 822

n3 =  =  = 583

This process will continue till there is not much difference between the last two approximations. We see that after 10th approximation, we get the sample size of 212.

If p = 25% to 40% and relative precision D = 0.05 then for different values of p and with 95% confidence level, the sample sizes are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| P | 0.25 | 0.30 | 0.35 | 0.40 |
| N | 4610 | 3585 | 2854 | 2305 |

The relative precision (D) may be converted into absolute precision (d) as

d = p x D = 

The sample sizes for different values of d and p and for 95% confidence level are given as:

Sample sizes for different values of p and d

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| d →  p↓ | 0.0125 | 0.0150 | 0.0175 | 0.0200 |
| 0.25 | **4610** | 3201 | 2352 | 1801 |
| 0.30 | 5163 | **3585** | 2634 | 2017 |
| 0.35 | 5593 | 3884 | **2854** | 2184 |
| 0.40 | 5901 | 4098 | 3010 | **2305** |

If the range is given, i.e. the prevalence is 10 to 25%, then it is always advisable to use prevalence 25% for precision. If the range is 45% to 55% then for absolute precision use p = 50% but for relative precision use 55%.

In case the coefficient of variation is given, the formula for the determination of sample size will be

 (2.9.22)

##### EXAMPLE 2.13

A survey is to be made of the common disease in a large population. For any disease that affects at least 1% of the individual in the population, it is desired to estimate the total number of cases, with a coefficient of variation of not more than 20% (a) what size of sample is needed, assuming that the presence of the disease can be recognized without mistake (b) what size is needed if the total cases are wanted for male and female with the same precision (Source: Cochran, 1977).

##### SOLUTION

Since

p = 0.01 and q = 0.99, C.V = 0.20

C.V(per unit) = 

Using

 = .99/.04 = 2475

If males and females are needed separately with same C.V two samples are assumed to be taken and probability of getting disease for male and females is same; hence the sample size will be 2475 for each.

It is hard to estimate the sample size without some prior knowledge about the S or C.V. This is the reason that sampling statisticians always record S or C.V. for which they worked. If S or P are not known, their values can be estimated by four possible ways.

1. from some previous survey
2. from the experience of sampling statistician
3. from pilot studies

**2.9.1 Estimation of Sample Size When Cost is Involved**

Since cost plays major role in conduct of surveys, so in the estimation of sample size, cost aspect should be taken into account. Let the simple linear cost function be:

C = C0 + nC1, (2.9.23)

where C = total cost, C0 = overhead cost, C1 = cost per unit.

Let the loss due to  not being equal to  is

, (2.9.24)

where d is constant.

The total survey cost and the loss involved will be

 (2.9.25)



For fixed sample size (2.9.25) is

 (2.9.26)

If L(n) is plotted against n, then the graph will have minimum value at some value of n. This value of n is termed as optimum value of sample size. The optimum value of n may be obtained by the derivative of (2.9.26) with respect to n.

After ignoring finite population correction the . Putting the value of  in (2.9.26), we have

L(n) = C0 + C1n + d S2/n. (2.9.27)

Finding the differentiation of the above expression w.r.t. n and equating to zero, the value of n that minimizes cost is

 (2.9.28)

##### EXAMPLE 2.14

If the loss function due to not being equal to  and the cost function is C = C0 + C1 n show that the sample size in case of simple random sampling ignoring f.p.c. is

.

###### SOLUTION

Since the loss is given in terms of mean deviation, the loss due to  not being equal to  will be



The total survey cost and the loss involved is

L(n) = C0 + C­1n + 

Finding the derivatives with respect to n and equating to zero, we have

.

**Example 2.15**

A sample of size n is selected from a population of N units; another sample of the same size is selected from the remaining N – n units. If and are the means of the first and the second sample and  is the population mean then

1. ,
2. ,
3. ,
4. If  is the pooled estimator, then

,

which is a variance of sample of size 2n taken from a population units.

## SOLUTION

1. Let  be the mean of a sample of size n taken from a population of N units then we have already proved that 
2.  mean of the remaining units





1. 

, where









Now

 

Therefore 

 

(vi) 

 as





## Alternative Proof of Example 2.15

Let the two estimators be:



where ai is an indicator variable defined earlier.

Now 



Since  is mean of sample which is selected after the selection of first sample, therefore it can be written as



Now 





Again 



Again consider 

or 

or 

 where  will have 2n values.

Now 

   





**EXAMPLE 2.17**

To estimate the total employment in an industry comprising of 70 factories, a sample of 10 factories is selected in the following manner. Three factories with relatively larger number of employees are definitely selected for the survey and a sample of 7 factories is selected by using srswor from the remaining 67 factories. Two estimators of the total employment in that industry are proposed:

(i)  and (ii) ,

where wi denotes the number of employees in the ith sample factory. Compare the mean square error of using the following data:

|  |  |
| --- | --- |
| Total number of employees for the 3 big factories | 8050 |
| Total number of employees for the remaining 67 factories | 60810 |
| Sum of squares of the number of employees for the remaining 67 factories | 71740058 |

(Source: Murthy 1966)

## SOLUTION:

Given estimators are:

(i)  and (ii) 

Total number of employees in 3 big factories and other 67 factories are 8050 + 60810 = 68860 = Y.





 



so is a biased estimator.

Bias in   = 100823 – 68860 = 31963.



 = 8050 + 60810 = 68860 = Y.

Since, therefore  is an unbiased estimator of Y.



 

 



 = 77,015,256

 = 1098648600

Since there is no bias in 



 = 143,990,660 

.

EXERCISES

1. From a population of N = 6 units, draw all possible samples of size 2 and 3 and demonstrate that  also E(s2) = S2. The population values are Yi = 3, 4, 6, 7, 9, 12.
2. A simple random sample of 100 households is selected from the village having 850 households for the record of television sets. Survey record shows that  and  Estimate total T.V. sets in 850 houses. Find 95% confidence limits.
3. A simple random sample without replacement of 10 agriculture plots was taken from 100 plots of a village and areas under wheat was recorded as 4.5, 2.5, 3.2, 2.6, 1.9, 2.8, 3.1, 2.8, 3.0, 2.8 acres. Estimate the total cultivated area under wheat. Find the standard error of the total mean under wheat and confidence limits.
4. A simple random sample of 2070 farms is taken from a population of 207 00 farms and information is collected on the number of cattle (yi) on each farm. The following data are obtained;  Estimate total number of cattle in the population. Find the coefficient of variation. How many more farms should be selected in future survey in order that the relative error of the estimate should not be more than 5%.

(i) 258810 (ii) 336

1. Four simple random samples of different sizes are taken from 663 mohallah from the 1973 population census of Libya. The values are recorded in the following table (yi denotes the agriculture holder). Estimate total agricultural holders and find the standard error in each case. The true total is 165541.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | 10% | 5% | 10% | 20% |
| Units | 7 | 33 | 66 | 132 |
| Σyi | 1080 | 8219 | 15538 | 34801 |
| Σy21 | 240592 | 3075517 | 5446659 | 15276357 |

Comment on the finding of result.

1. In a sample of 200 colleges from a population of 2000, 140 colleges were in favour of the proposal that a new system of examination should be introduced (i) estimate the 95% confidence limits for the number of colleges in the population in the favour of the proposal (ii) do the above data furnish evidence at 95% level to reject the hypothesis that the only 50% of all the colleges are in favour of this proposal.
2. A simple random sample of 290 households was chosen from a city containing 14,828 households. Each family was asked whether it owned or rented the house and also whether it had the exclusive use of an indoor toilet. Results are as;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Owned** | | Rented | | **Total** |
|  | **Yes** | **No** | **Yes** | **No** |
|  | | 141 | 6 | 109 | 34 | 290 |

For the family who rent, estimate the percentage in the area with exclusive use of an indoor toilet and give the standard error of this estimate. Estimate the total number of renting families in the area who do not have exclusive indoor toilet facilities and give the standard error of this estimate. If the total number of renting families in the city area is 7526, make a new estimate of the number without toilet facilities and give the standard error of this estimate.

1. Material for the construction of 5000 wells was issued in a district. The list of cultivators whom the material was issued available along with the proposed location of each well. A large part of material was, however, reported to have been misused. It is proposed to estimate the proportion of wells not actually constructed, by taking a simple random sample wells with the permissible margin of error of 10% and the degree of assurance desired 95%. Determine the size of the sample required to estimate the proportion of wells not constructed for different values of the population proportion ranging from 0.5 to 0.9.
2. A simple random sample of 50 households was selected from 300 households. Out of 50 households 10 were found to possess a T.V. sets. Members of the families were 4, 4, 5, 5, 6, 5, 6, 7, 3, and 5. Estimate the followings:
3. total number of households in the population possessing T.V. sets,
4. total number of persons in the population.
5. A population contains N units, the variate value of one unit being known to be Y­1­. A without replacement sample of size n is selected from the remaining (N-1) units. Show that the estimator  has a smaller variance that  based on a random sample without replacement of size n taken from the whole population.
6. Select a simple random sample of size 10 from the given 47 villages of a certain districts of Punjab, showing area under guava crop. Estimate total area under guava crop along with standard error of your estimate. Compare your result with actual values.

|  |  |  |  |
| --- | --- | --- | --- |
| Selected  Village | Area Under  Guava Crop | Selected  Village | Area Under  Guava Crop |
| 01 | 166.15 |  | 12.57 |
| 02 | 24.73 |  | 2.00 |
| 03 | 100.77 |  | 6.72 |
| 04 | 87.14 |  | 20.75 |
| 05 | 116.28 |  | 51.65 |
| 06 | 60.22 |  | 16.42 |
| 07 | 13.59 |  | 3.90 |
| 08 | 41.70 |  | 2.44 |
| 09 | 10.52 |  | 3.90 |
| 10 | 13.85 |  | 15.31 |
| 11 | 12.92 |  | 1.44 |
| 12 | 10.73 |  | 14.88 |
| 13 | 38.64 |  | 23.01 |
| 14 | 15.92 |  | 3.44 |
| 15 | 9.09 |  | 14.32 |
| 16 | 155.51 |  | 24.39 |
| 17 | 10.34 |  | 9.88 |
| 18 | 95.16 |  | 17.66 |
| 19 | 22.40 |  | 3.26 |
| 20 | 10.97 |  | 6.89 |
| 21 | 39.07 |  | 0.84 |
| 22 | 13.70 |  | 13.02 |
| 23 | 26.64 |  | 32.85 |
| 24 | 1.40 |  |  |

1. The data given below pertain to one complete location of milk yield of 250 cows in an organized daily farm.
2. Select a simple random of size 25.
3. Estimate the mean with its standard error.
4. Construct a 95% confidence limit for the population mean.

Milk yield (in 10 kgs.) of 250 cows of one complete location (305 days) in an organized dairy farm (units are numbered column wise). {source Sing, Sing and Kumar 1981}.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 230 | 293 | 163 | 290 | 200 | 173 | 194 | 322 | 169 | 230 |
| 297 | 151 | 248 | 271 | 269 | 214 | 167 | 207 | 240 | 286 |
| 184 | 248 | 327 | 338 | 165 | 177 | 270 | 177 | 202 | 155 |
| 155 | 293 | 190 | 172 | 150 | 319 | 151 | 118 | 213 | 114 |
| 186 | 167 | 129 | 185 | 231 | 199 | 265 | 306 | 173 | 276 |
| 291 | 231 | 205 | 220 | 246 | 239 | 186 | 299 | 233 | 208 |
| 265 | 204 | 300 | 195 | 239 | 173 | 237 | 282 | 221 | 218 |
| 197 | 215 | 213 | 290 | 146 | 232 | 305 | 184 | 149 | 267 |
| 188 | 219 | 171 | 099 | 329 | 199 | 180 | 225 | 257 | 202 |
| 189 | 207 | 792 | 327 | 201 | 300 | 206 | 199 | 299 | 153 |
| 175 | 287 | 277 | 230 | 258 | 137 | 174 | 301 | 260 | 282 |
| 211 | 212 | 284 | 214 | 283 | 139 | 223 | 212 | 207 | 224 |
| 207 | 111 | 272 | 192 | 127 | 303 | 221 | 187 | 309 | 263 |
| 203 | 176 | 233 | 239 | 176 | 218 | 193 | 243 | 236 | 275 |
| 228 | 198 | 241 | 219 | 167 | 193 | 234 | 179 | 126 | 176 |
| 279 | 178 | 275 | 260 | 191 | 174 | 235 | 338 | 242 | 238 |
| 211 | 187 | 184 | 189 | 305 | 221 | 253 | 225 | 327 | 203 |
| 195 | 158 | 156 | 185 | 170 | 271 | 160 | 188 | 165 | 218 |
| 312 | 243 | 267 | 298 | 196 | 139 | 205 | 298 | 238 | 217 |
| 145 | 201 | 313 | 230 | 185 | 166 | 147 | 223 | 271 | 133 |
| 155 | 230 | 287 | 329 | 265 | 150 | 286 | 271 | 268 | 198 |
| 214 | 231 | 163 | 335 | 198 | 270 | 187 | 174 | 163 | 201 |
| 192 | 247 | 247 | 297 | 178 | 240 | 290 | 234 | 170 | 227 |
| 230 | 353 | 170 | 159 | 236 | 181 | 230 | 240 | 212 | 242 |
| 151 | 158 | 253 | 179 | 263 | 158 | 250 | 226 | 246 | 301 |

1. Following is a list of 70 villages of a Tehsil of a certain district of Punjab along with population and cultivated area of the same years select
2. a sample of 10 villages,
3. a sample of 15 villages,
4. a sample of 20 villages.

Estimate the population for three cases, estimate the standard error for each case. Compare the result and comment.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.**  **No.** | **Population** | **Cultivated**  **Area (Acres)** | **S.**  **No.** | **Population** | **Cultivated**  **Area (Acres)** |
|  | 226 | 678 |  | 904 | 760 |
|  | 670 | 663 |  | 773 | 602 |
|  | 4505 | 1290 |  | 1040 | 532 |
|  | 1732 | 1170 |  | 760 | 438 |
|  | 2874 | 1390 |  | 2084 | 633 |
|  | 2282 | 1110 |  | 828 | 277 |
|  | 793 | 760 |  | 4877 | 1640 |
|  | 895 | 730 |  | 911 | 424 |
|  | 1157 | 950 |  | 1205 | 822 |
|  | 3201 | 1700 |  | 1139 | 555 |
|  | 1117 | 909 |  | 4064 | 347 |
|  | 1236 | 1169 |  | 1114 | 744 |
|  | 5201 | 1840 |  | 547 | 372 |
|  | 848 | 660 |  | 1178 | 644 |
|  | 1238 | 1140 |  | 1159 | 732 |
|  | 1917 | 1360 |  | 441 | 622 |
|  | 1800 | 1509 |  | 555 | 342 |
|  | 2335 | 1810 |  | 827 | 387 |
|  | 4396 | 2240 |  | 2867 | 322 |
|  | 1607 | 1225 |  | 726 | 636 |
|  | 2071 | 1250 |  | 633 | 410 |
|  | 2155 | 1690 |  | 680 | 427 |
|  | 7780 | 3200 |  | 587 | 496 |
|  | 2746 | 1744 |  | 1901 | 936 |
|  | 2549 | 2400 |  | 2419 | 1226 |
|  | 1007 | 680 |  | 1258 | 836 |
|  | 1567 | 970 |  | 1225 | 634 |
|  | 5271 | 1850 |  | 1447 | 978 |
|  | 659 | 340 |  | 1314 | 724 |
|  | 3209 | 2450 |  | 1298 | 422 |
|  | 2902 | 1760 |  | 728 | 493 |
|  | 2955 | 2120 |  | 851 | 396 |
|  | 1746 | 1220 |  | 786 | 732 |
|  | 1045 | 860 |  | 663 | 422 |
|  | 666 | 620 |  | 740 | 370 |

1. A residential area has 5000 houses. It is required to estimate the proportion of houses with more than three persons living in them. The estimator is required to have standard error not exceeding 0.02. From other surveys, it would appear that the proportion for this is lying in the range 0.35 to 0.55. How large a sample is needed to meet the accuracy?

##### (HINT) We know that for large sample . The problem is to find p. We know that p is lying between the limits 0.35 and 0.55.P + t S.E(p).

1. If the loss function in terms of money is proportional to the value of the coefficient of variation and the cost function is, C = C0 + C1n. Show that the sample size ignoring f.p.c. will be  where Cy is coefficient of variation.

16. Following is a population of 47 villages showing area under fresh fruit in a certain distinct of Punjab. Select 10 percent sample of the villages and estimate the total area under fresh fruit along with standard error of the estimate. Compare the estimated total with actual one.

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. of the Selected Villages | Area UnderFresh Fruit % | S. No. of theSelected Villages | Area UnderFresh Fruit % |
|  | 127.00 |  | 8.00 |
|  | 32.00 |  | 24.00 |
|  | 74.99 |  | 3.00 |
|  | 57.00 |  | 24.00 |
|  | 68.00 |  | 38.00 |
|  | 61.01 |  | 13.00 |
|  | 6.00 |  | 4.00 |
|  | 27.00 |  | 12.00 |
|  | 68.00 |  | 1.00 |
|  | 13.17 |  | 5.00 |
|  | 8.00 |  | 1.00 |
|  | 8.00 |  | 13.00 |
|  | 22.00 |  | 19.00 |
|  | 10.00 |  | 5.00 |
|  | 30.00 |  | 12.00 |
|  | 98.99 |  | 12.00 |
|  | 7.38 |  | 7.38 |
|  | 62.99 |  | 7.00 |
|  | 13.00 |  | 2.00 |
|  | 14.00 |  | 8.00 |
|  | 30.00 |  | 1.00 |
|  | 10.00 |  | 10.00 |
|  | 13.00 |  | 22.00 |
|  | 7.00 |  | **1419.91** |

##### APPENDIX- 1 ( Random Digits)

57780 97609 52482 12783 88768 12323 64967 22970 11204 37576

68327 00067 17487 49149 25894 23639 86557 04139 10756 76285

55888 82253 67464 91628 88764 43598 45481 00331 15900 97699

84910 44827 31173 44247 56573 91759 79931 26644 27048 53704

35654 53638 00563 57230 07395 10813 99194 81592 96834 21374

46381 60071 20835 43110 31842 02855 73446 24456 24268 85291

11212 06034 77313 66896 47902 63483 09924 83635 30013 61791

49703 07226 73337 49223 73312 09534 64005 79267 76590 26066

05482 30340 24606 99042 16536 14267 84084 16198 94852 44305

92947 65090 47455 90675 89921 13036 92867 04786 76776 18675

51806 61445 32437 01129 03644 70024 07629 55805 85616 59569

16383 30577 91319 67998 72423 81307 75192 80443 09651 30068

30893 85406 42369 71836 74479 68273 78133 34506 68711 58725

59790 11682 63156 10443 99033 76460 36814 36917 37232 66218

06271 74980 46094 21881 43525 16516 26393 89082 24343 57546

93325 61834 40763 81178 17507 90432 50973 35591 36930 03184

46690 08927 32962 24882 83156 58597 88267 32479 80440 41668

82041 88942 57572 34539 43812 58483 43779 42718 46798 49079

14306 04003 91186 70093 62700 99408 72236 52722 37531 24590

63471 77583 80056 59027 37031 05819 90836 19530 07138 36431

68467 17634 84211 31776 92996 75644 82043 84157 10877 12536

94308 57895 08121 07088 65080 51928 74237 00449 86625 06626

52218 32502 82195 43867 79935 34620 37386 00243 46353 44499

46586 08309 52702 85464 06670 18796 74713 81632 34056 56461

07869 80471 69139 82408 33989 44250 79597 15182 14956 70423

46719 60281 88638 26909 32415 31864 53708 60219 44482 40004

74687 71227 59716 80619 56816 73807 94150 21991 22901 74351

42731 50249 11685 54034 12710 35159 00214 19440 61539 25717

71740 29429 86822 01187 96497 25823 18415 06087 05886 11205

96746 05938 11828 47727 02522 33147 92846 15010 96725 67903

27564 81744 51909 36192 45263 33212 71808 24753 72644 74441

21895 29683 26533 14740 94286 90342 24671 52762 22051 31743

01492 40778 05988 65760 13468 31132 37106 02723 40202 15824

55846 19271 22846 80425 00235 34292 72181 24910 25245 81239

14615 75196 40313 50783 66585 39010 76796 31385 26785 66830

77848 15755 91938 81915 65312 86956 26195 61525 97406 67988

87167 03106 52876 31670 23850 13257 77510 42393 53782 32412

73018 56511 89388 73133 12074 62538 57215 23476 92150 14737

29247 67792 10593 22772 03407 24319 19525 24672 21182 10765

17412 09161 34905 44524 20124 85151 25952 81930 43536 39705

68805 19830 87973 99691 25096 41497 57562 35553 77057 06161

40551 36740 61851 76158 35441 66188 87728 66375 98049 84604

90379 06314 21897 42800 63963 44258 14381 90884 66620 14538

09466 65311 95514 51559 29960 07521 42180 86677 94240 59783

15821 25078 19388 93798 50820 88254 20504 74158 35756 42100

10328 60890 05204 30069 79630 31572 63273 13703 52954 72793

49727 08160 81650 71690 56327 06729 22495 49756 43333 34533

71118 41798 34541 76132 40522 51521 74382 06305 11956 30611

53253 23100 03743 48999 37736 92186 19108 69017 21661 17175

12206 24205 32372 46438 67981 53226 24943 68659 91924 69555