18K5SO9

Kunthavai Naachiyaar Govt. Arts College (W) Autonomous, Thanjavur

B.Stat Major

Core Course – IX - Sampling Techniques

Hrs : 5
Credit : 5

Unit – I

Definitions – Parameter, Statistic, Population, Sample, Sampling distribution, Standard error. Principal steps in a sample survey, principles of sample survey, sampling and non-sampling errors. Limitations of sampling.

Unit – II

Simple Random sampling – Selecting SRSWR & SRSWOR , Merits and limitations – Derivation of sample mean & variance, unbiased estimates of mean & variance, Comparison of SRSWOR with SRSWR (Variance alone).

Unit – III

Stratified Random Sampling – Derivation of unbiased estimates of mean and variance – Optimum Allocation and Proportional Allocation – Comparison of SRS with stratified sampling (variance alone) – Gain in efficiency due to stratification, merits of stratified Random sampling.

Unit – IV

Systematic sampling – Estimation of population mean and variance, merits of systematic sampling. Comparison of SRS, Stratified and systematic sampling using variance.

Unit – V

Method of Ratio estimation – definition, notations, estimation of the mean and variance from a sample, comparison of the variance of ratio estimator with the mean per unit – Bias of the ratio estimator.

Books for Study :

1. Sampling Techniques – W.G. Cochran
2. Fundamental of Applied Statistics – V.K. Kapoor and S.P.Gupta

Core course IX - Sampling Technique Sub code: 18K5SG9

UNIT - I

Definitions - parameter, statistics, population, sample, Sampling distribution standard error, Principle steps in a Sample survey, Principles of sample survey, sampling and non-sampling errors. Limitations of sampling.

UNIT - II

Simple Random Sampling - Selecting SRSWR & SRSWOR, Merits and Limitations - Derivation of Sample mean & variance, unbiased estimates of mean & variance, comparison of SRSWOR with SRSWR (variance alone).

UNIT - III

Stratified Random sampling - Derivation of unbiased estimates of mean and variance - optimum Allocations and Proportional Allocation - comparison of SRS with Stratified sampling (variance alone) - Gain in efficiency due to Stratification, merits of stratified Random Sampling.

UNIT - IV

Systematic sampling - Estimation of population Mean and variance, merits of systematic sampling, comparison of SRS, Stratified and systematic sampling using variance.

UNIT - V

Method of Ratio estimation - definitions, Notations, estimation of the mean and variance from a sample comparison of the variance of ratio estimator with the mean per unit - Bias of the ratio estimator

UNIT - I

Introduction

The group of individuals under study is called population or universe. Thus in statistics, population is an aggregate of objects, animate or inanimate under study. The population may be finite or infinite.

Sample and Sample size:

A finite subset of statistical individuals in a population is called a sample and the number of individuals in a sample is called the sample size.

Parameters and statistics:

In order to avoid verbal confusion with the Statistical constants of the population, viz. mean variance, etc. of the population which are usually reffered to as parameters, statistical measures computed from the sample observations alone, (eg.) mean variance etc. of the sample have been termed by professor R.A. Fisher as statistics.

In practice parameter values are not known and their estimates based on the sample values are generally used. Thus, statistics which may be regarded as an estimate of the parameter, obtained from the sample is a function of the samples values only. It may be pointed out that a statistic, as it is based on sample values and as there are multiple choices of the samples

that can be drawn from a population, varies from sample to sample. The determination or the characterisation of the variation (in the value of the statistic obtained from different samples) that may be attributed to chance or fluctuations of sampling is one of the fundamental problems of the sampling theory.

Remark (unbiased estimate)

A statistic $t = t(x_1, x_2, \ldots x_n)$ a function of the sample values $x_1, x_2, \ldots x_n$ is an unbiased estimate of population parameter θ, if $E(t) =$

In other words, if,

E(Statistics) = parameter

then statistics is said to be an unbiased estimate of the parameter.

Sampling distribution.

The number of possible samples of size n that can be drawn from a finite population of size N is N^n [If N is large or infinite then we can draw a large number of such samples] For each of these samples we can compute a statistic, say: 't': eg mean, variance, etc. which will obviously vary from sample to sample. The aggregate of the various values of

the statistic under consideration so obtained (one from each sample) may be grouped into a frequency distribution which is known as the sampling distribution of the statistic. Thus, we can have the sampling distribution of the sample mean $\bar{x}$, the sample variance, etc.

Standard error:

The standard deviation of the sampling distribution of a statistic is known as its standard error. The standard errors (S.E) of some of the well known statistics are given below, where n is the sample size, σ^2 the population variance, P the population proportion and $Q = 1 - P$.

S.NO	statistic	standard error
1.	$\bar{x}$	$\sigma/\sqrt{n}$
2.	observed sample proportion 'p'	$\sqrt{PQ/n}$
3.	sample standard deviation s	$\sqrt{\sigma^2/2n}$
4	s^2	$\sigma^2\sqrt{2/n}$
5	Quartiles	$1.3626\,3\sigma/\sqrt{n}$
6	Median	$1.25331\,\sigma/\sqrt{n}$
7.	r - sample correlation coefficient	$(1-\rho^2)/\sqrt{n}$
8	μ_3	ρ being the population correlation coefficient $\sigma^3\sqrt{96/n}$

Any one example.

9.	M_4	$\sigma^4\sqrt{96/n}$
10)	Coefficient of variation (V)	$\frac{V}{\sqrt{2n}}\sqrt{1+\frac{2V^2}{104}}$ $= V/\sqrt{2n}$

The Principal steps in a sample Survey.

The main steps involved in the planning and execution of a sample survey may be grouped some what orbitary under the following heading.

* objectives of the Survey:

The first step is to define (in clear and concrete terms) the objectives of the survey. It is generally found that even the sponsoring agency is not quite clear in mind as to what it wants and now it is going to use the results. The sponsors of the survey should take to care that these objectives are commensurate with the available resources in terms of money, manpower and the time limit required for the availability of the results of the survey.

2) Defining the population to be sampled.

The population (i.e) the aggregate of objects (animate or inanimate) from which sample

chosen should be defined in clear and unambiguous terms. For eg. in sampling of farms clear but rules must be framed to define a farm regarding shape, size, etc. keeping in mind the border line cases) so as to enable the investigator to decide in the field without much hesitation whether or not include a given farm in the population.

But (practical difficulties) in handing certain segments of the population may point their elimination from scope of the survey.) consequently (for reasons of practicability or conviences the population to be Sampled por (the sampled population) is different in fact more restricted then the population for which results are wanted (the target population).)

* <u>The frame and sampling units</u>

(The population must be capable of division into what are called sampling units for purposes of sample selection)(The sampling units must be cover the entire population and they must be distinct, unambiguous and non-overlapping) in the some that every element of the population belongs to one and only sampling unit (for eg. in socio economic survey for selecting people in a four, the sampling unit might be an individual person, a family a house held or a

a block in a locality.)

(In order to cover the population decided upon, there should be same list map or other acceptable material called the frame which serves as a guide to the population to be covered the construction of the frame is often one of the major practical problems since it is the frame which determine the structure of the sample survey. The lists which have been routinely, collected for some purpose, are usually found to be incomplete or partly illegible or often contain an unknown amount of duplication (the lists collected such lists should be carefully scrutinised and examined to, ensure free from these defects and are up to date.) If they are not upto date they should be brought upto date before using them. A good frame is hard to come by and only good experience helps to construct a good frame. (only by a good experience a good frame can be constructed)

* Data to be collected.

[The data should be collected keeping in view the objectives of the survey]. the tendency should not be to collect too many data some of which are never subsequently examined and analysed. (Too many data never examined should not be collected)

A practical method is to chalk out an outline of the tables that the survey should produce, this helps in eliminating the collection of irrevent information and no essential data are omitted.

* The questionnaire or schedule

Having decided about the type of the data to be collected the important part of the sample survey is the construction of the questionnaire (to be filled in by the respondent) or schedule of enquiry (to be completed by the interviewer) which requires skill, special techniques as well as familiarity with subject matter under study. The questions should be clear, brief, corroborative, non-offending, courteous in tone, unambiguous and to the point so that not much scope of guessing is left on the part of the respondent or interviewer.

Suitable and detailed instructions for filling up the questionnaire or schedule should also be prepared.

* Method of collecting information.

The two methods commonly employed for collecting data for human population are,

(i) Interview method.

In this method, the investigator goes from house to house and interviews the individuals

personally. He asks the questions one by one and fills up the schedule on the basis of the information supplied by the individuals.

ii) Mailed questionnaire method.

Decide whether the data should be collected by interview method or mailed questionnaire method or by physical observation.

Keeping in view the costs involved and accuracy aimed at mail surveys are less costly But mail method is practicable among the educated people who are really interested in the particular survey. In interview without investigations the data collected may be worthless. In cases where data are to be collected by observations, the method of measurement, the type of measuring equipment or instrument etc are to be decided.

* Non-respondents:

Due to practical difficulties the data cannot be collected for all the sampled units.

* Selection of proper sampling design

A number of designs (plans) for the selection of a sample are available and a judicious selection will guarantee good and reliable estimates. For each sampling

plan. rough estimates of sample size n can be obtained for a desired degree of precision. The relative costs and time involved should also be considered before making a final selection of the sampling plan.

* Organisation of field work.

It is absolutely essential that the personnel should be thoroughly trained in locating the sample units, recording the measurements, the methods of collection of required data before starting the field work.

* The pretest

From practical point of view a small pretest trying has been found to be immensely useful. It always helps to decide upon effective method of asking questions and results in the improvement of the questionnaire.

* Summary and analysis of the data.

The analysis of the data may be broadly classified into following heads.

a) Scrutiny and editing of the data.

An initial quality check should be carried out by the supervisory staff while the investigators are in the field.

b) Tabulation of data

Before carrying out the tabulation of the

data, we must decide about the procedure for tabulation of the data which are incomplete due to non-response to certain items in the questionnaire and where certains questions are deleted in editing process. The method of tabulation, viz. hand tabulation or machine tabulation, will depend upon the quantity of the data.

c) Statistical Analysis:

The data has been properly scrutinished edited and tabulated, a very careful statistical analysis is to be made. Different methods of estimation may be available for the same data

d) Reporting and conclusions.

Finally a report incorporating detailed statement of the different stages of the survey should be prepared. In the presentation of the results, it is good practice to report the technical aspect of the design, viz.

* Information gained for future surveys.

Any completed surveys is helpful in providing a note of caution and taking lessons from it for designing future surveys. The information gained from any completed sample in the form of the data regarding the means, standard deviations and the nature of the variability of the principal measurements together with the

cost involved in obtaining the data serves as a potential guide for improved future sampling.

Principles of sample survey:

The theory of sampling is based on the following important principles.

* Principle of statistical regularity.

"The law of statistical regularity lays down that a moderately large no. of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group."

Large Numbers which states that." other things being equal, as the sample size increases, the result tend to be more reliable and accurate".

* Principle of validity:

By the validity of a sample design we mean that it should enable us to obtain valid tests and estimates about the parameters of the population. The samples obtained by the technique of probability sampling satisfy this principle.

* Principle of optimisation:

(i) achieving a given level of efficiency at minimum cost and

(ii) obtaining maximum possible efficiency with given level of cost.

Sampling and Non-Sampling errors

(i) Sampling errors and

(ii) Non-Sampling errors.

(i) Sampling errors:

Sampling errors have their origin in sampling and arise due to the fact that city a part of the population (ie sample) has been used to estimate population parameters and draw inferences about the population.
As such the sampling errors are absent in a complete enumeration survey.

Sampling biases are primarily due to the following reasons.

(1) Faulty selection of the sample.

Some of the bias is introduced by the use of defective sampling technique for the selection of a sample, eg purposive or judgment sampling in which the investigator deliberately selects a representative sample to obtain certain results. This bias can be overcome by strictly adhering to a simple random samples or by selecting a sample at random subject to restrictions which while improving the accuracy are of such such nature that they do not introduce bias in the results.

2) Substitution.

If difficulties arise in enumerating a particular sampling unit included in the random sample, the investigators usually substitude a convenient member of the population. this obviously leads to some bias since the characteristics possessed by the substitued unit will usually be different from those possessed by the unit originally included in the sample.

3) Faulty demarcation of sampling units.

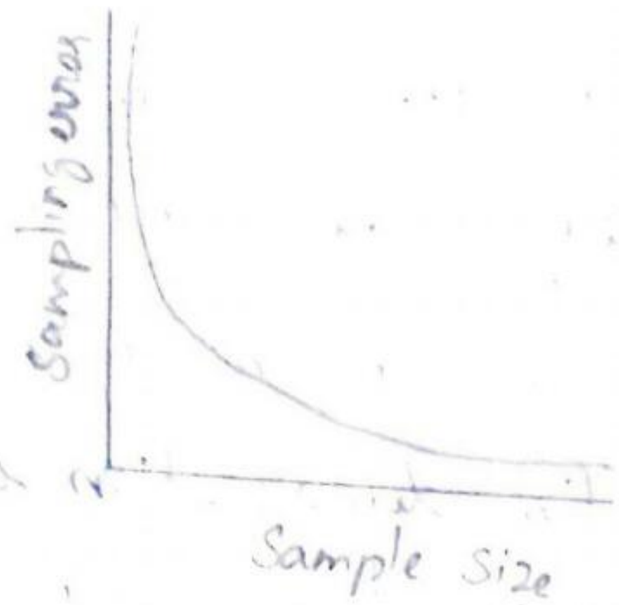
Bias due to defective demarcation of sampling units is particularly significant in area surveys such as agricultural experiments in the field or crop cutting survey etc. In such surveys, while dealing with border line cases, it depends more or less on the discretion of the investigator whether to include them in the sample or not.

4) Constant error due to improper choice of the statistics for estimating the population parameters.

For example if $x_1, x_2, \ldots x_n$ is a sample of independent observations then the sample variance $\delta^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$ as an estimate of the population variance σ^2 is biased whereas the statistic $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimate of σ^2.

Remark:

Increase in the sample size (i.e, the number of units in the sample) usually results in the decrease in sampling error. In fact, in many situations this decrease in sampling error is inversely proportional to the square root of the sample size as illustrated in the diagram on next page.



(ii) Non-sampling error:

data obtained in a complete census, although free from sampling errors, would still be subject to non-sampling errors where as data obtained in a sample survey would be subject to both sampling and non-sampling errors.

1) Faulty planning or definitions:

The planning of a survey consists in explicitly stating the objectives of the survey. These objectives are then translated into

(i) a set of definitions of the characteristics for which data are to be collected and

(ii) into a set of Specifications for collection, processing and publishing. Here the non-sampling errors can be due to

a) Data specification being inadequate and inconsistent with respect to the objectives of the survey.

b) Error due to location of the units and actual measurement of the characteristics errors in recording the measurements, errors due to ill designed questionnaire, etc.

c) Lack of trained and qualified investigators and lack of adequate supervisory staff.

Response Error:

These errors are introduced as a result of the responses furnished by the respondents and may be due to any of the following reasons.

(i) Response errors may be accidental. For example, the respondent may misunderstand a particular questions and accordingly furnish improper information unintentionally.

(ii) Prestige bias an appeal to the pride or prestige of person interviewed may introduce yet another kind of bias, called prestige bias by virtue of which he may upgrade his education, intelligence quotient, occupation, income, etc. or downgrade his age, thus resulting in wrong answers.

(iii) Self interest. Quite often, in order to safeguard ones' self interest, one may give incorrect information eg. a person may give an underestimate of his salary or production and an over statement of his expenses or requirements etc.

(iv) Bias due to interviewer. Sometimes the interviewer may affect the accuracy of the response by the way he asks questions or records them. The information obtained on suggestions from the interviewer is very likely to be influenced by interviewer's beliefs and prejudices.

v) Failure of respondents memory one source of error which is common to most of the methods of collecting informations is that of recall. Many of the questions in surveys refer to happenings or conditions in the past and there is a problem both of remembering the event and associating it with the correct time period.

Advantages of Sampling over complete census.

The main advantages or merits of sampling technique over the complete enumeration survey may be outlined as follows

1) Less Time

There is considerable saving in time and labour since only a part of the population has to be examined. The

results can be obtained more rapidly and the data can be analysed much faster since relatively fewer data have to be collected and processed

2) Reduced cost of the survey

Sampling usually results in reduction in cost in terms of money and in terms of man hours. Although the amount of labour and the expenses involved in collecting information are generally greater per unit of sample than in complete enumeration, the total cost of the sample survey is expected to be much smaller than that of a complete census. Since in most of the cases our resources are limited in terms of money and the time within which the results of the survey should be obtained, it is usually imperative to resort to sampling rather than complete enumeration.

3) Greater accuracy of results

The results of a sample survey are usually much more reliable than those obtained from a complete census due to the following reasons

i) It is always possible to determine the extent of the sampling errors and

(ii) the non-sampling errors due to number of factors such as training of field workers, measuring and recording observations, location of units, incompleteness of return returns, biases due to interviewers etc.

are likely to be of a serious nature in complete census than in a sample survey. In a sample survey non-sampling errors can be controlled more efficiently by employing more qualified and better trained personnel, better supervision and better equipment for processing and analysis of relatively limited data. Moreover, it is easier to guard against incomplete and inaccurate returns. There can be a follow up in case of non-response or incomplete response. Effective control of non-sampling errors more than compensates the errors in the estimates due to sampling. As such more sophisticated statistical techniques can be employed to obtain relatively more reliable results.

4) Greater scope

Sample survey has generally greater scope as compared with complete census. The complete enumeration is impracticable, rather inconceivable if the survey requires a highly trained personnel and more sophisticated equipment for the collection and analysis of the data. Since sample survey saves in time and money, it is possible to have a thorough and intensive enquiry because a

more detailed information can be obtained from a small group of respondents.

5) If the population is too large, as for example, trees in a jungle, we are left with no way but to resort to sampling.

6) If testing is destructive, i.e), if the quality of an article can be determined only by destroying the article in the process of testing, as for example.

(i) Jesting the quality of milk or chemical salt by analysis.

(ii) Jesting the breaking strength of chalks

(iii) Jesting of crackers and explosives.

(iv) Jesting the life of an electric tube or Bulb bulb, etc.

∴ complete enumeration is impractible and sampling technique is the only method to be used in such cases.

7) If the population is hypothetical, as for example. in coin tossing problem where the process may continue indefinitely (any number of times) sampling method is the only scientific method of estimating the parameters of the universe.

Remarks 1: Prof R.A. Fisher (1950) in a report of "The Sub-Commission on statistical sampling of united nations." Sums advantages of sampling

techniques over complete census in the following four words: Adaptability, Speed, Economy and Scientific approach.

2) From practical point of view it has been seen that the method of random sampling with suitable adaption of stratification of the universe. If it is heterogeneous or the techniques of multistage random sampling if there are clearly demarcated stages, gives fairly good results often better than those obtained by a complete census.

Limitations of sampling.

The advantages of sampling over complete census as enumerated above can be derived only if,

(i) The sampling units are drawn in a scientific manner.

(ii) appropriate sampling techniques is used and

(iii). The sample size is adequate sampling theory has its own limitations and problems which may be briefly outlined as follows.

1) Proper care should be taken in the planning and execution of the sample survey, otherwise the results obtained might be inaccurate and misleading.

2) Sampling theory requires the services of trained and qualified personnel and sophisticated equipment for its planning, execution and analysis. In the absence of these, the results of the sample survey are not trustworthy

3) However, if the information is required about each and every unit of the universe, there is no way but to ~~restard~~ resort to complete enumeration. Moreover, if time and money are not important factors or if the universe is not too large, a complete census may be better than any sampling method.

UNIT - II

SIMPLE RANDOM SAMPLING.

It is the techniques of drawing a sample in such a way that each unit of the population has an equal and independent chance of being included in the sample.

In this method an equal probability of selection is assigned to each unit of the population at the first draw. It also implies an equal probability of selecting any unit from the available units at subsequent draws.

Thus in SRS from a population of N units, the probability of drawing any unit ~~in the~~ at the ~~second draw from am~~ first draw is $1/N$, the Probability of drawing any unit in the second draw from among the available $(N-1)$ units, is $1/(N-1)$ and so on.

Let E_{ri} be the event that any specified unit is selected at the r^{th} draw. Then

$P(E_r)$ = Prob. $\{$ that the specified unit is not selecte in anyone of the Problem previous $(r-1)$ draws and then selected at the rth draw $\}$

$$\therefore P(E_r) = \prod_{i=1}^{r-1} P\{\text{It is not selected at } i\text{th draw}\} \times P\{\text{It is selected at } r\text{th draw given that it is not selected at the previous } (r-1) \text{ draws}\}.$$

(By compound probability theorem, since draws are independent)

$$\therefore \quad P(E_r) = \prod_{i=1}^{r-1}\left[1 - \frac{1}{N-(i-1)}\right] \times \frac{1}{N-(r-1)}$$

$$= \prod_{i=1}^{r-1}\left(\frac{N-i}{N-i+1}\right) \times \frac{1}{N-r+1}$$

$$= \frac{N-1}{N} \times \frac{N-2}{N-1} \times \frac{N-3}{N-2} \times \cdots \times \frac{N-r+1}{N-r+2} \times \frac{1}{N-r+1}$$

$$= 1/N$$

$$\therefore \quad P(E_r) = \frac{1}{N} = P(E_1)$$

This leads to a very interesting and important property of Simple Random Sampling without Replacement (SRSWR), viz:

"The probability of selecting a specified unit of the population at any given draw is equal to the probability of its being selected at the first draw."

Probability of selecting any specified unit in the sample

Since a specified unit can be included in the sample of size n in n mutually exclusive ways, viz. it can be selected in the sample at the rth draw ($r = 1, 2, \ldots, n$) and since

$r = 1, 2, \ldots, n$

by the addition theorem of probability, we get.

The probability that a specified unit in indeed in the sample

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

Remark:

Simple Random sampling can also be defined equivalently as follows:

Let us suppose that a sample of size n is drawn from a population of size N. There are $\binom{N}{n}$ possible samples. S.R.S. is the technique of selecting the sample in such a way that each of the $\binom{N}{n}$ samples has an equal chance or probability $p = \frac{1}{\binom{N}{n}}$ of being selected, as explained below:

In S.R.S.

Probability of selecting any unit at the first draw $= \frac{1}{N}$

Probability of selecting any unit out of the remaining (N-1) units in the second draw $= \frac{1}{N-1}$

and so on.

Probability of selecting any unit of the

remaining $N-(i-1)$ units at the ith draw

$$= \frac{1}{N-(i-1)}, (i = 3, 4, \ldots, n)$$

Since all the draws are independent, by compound probability theorem, the probability of selecting a sample of size n in a fixed specified order, is

$$\frac{1}{N(N-1)(N-2)\ldots(N-n+1)}$$

Since this probability is independent of the order of the sample and since there are $n!$ permutations of the sampled units, by addition theorem of probability, the required probability of obtaining a sample of size n (in any order) is

$$P = \frac{n!}{N(N-1)\ldots(N-n+1)} = \frac{1}{\binom{N}{n}}$$

as required.

---

* Selection of a Simple Random Sample:

Random sample refers to that method of sample selection in which every item has an equal chance of being selected. But the random sample does not depend upon the method of selection only but also on the size and nature of the population. Some procedure which

is simple and good for small population is not so for the large population. Generally, the method of selection should be independent of the properties of sampled population. Proper care has to be taken to ensure that selected sample is random. Human bias, which varies from individual to individual is inherent in any sampling scheme administered by human beings. Random sample can be obtained by any of the following methods.

a) By lottery method. System

b) 'Mechanical Randomization' or 'Random Numbers' method.

a) Lottery System.

The simplest method of selecting a random sample in the lottery system, which is illustrated below by means of an example:

Suppose we want to select 'n' candidates out of n. We assign the numbers 1 to n; one number to each candidate and write these numbers (1 to n) on n slips which are made as homogeneous as possible in shape, size, colour, etc. These slips are then put in a bag and thoroughly

shuffled and then 'n' slips are drawn one by one. The 'n' candidates corresponding do numbers on the slips drawn, will constitute a random sample.

This method of selection is quite independent of the properties of population. Generally in place of chits, cards are used. We make one card correspond to one of the unit of the Population by writing on it the number of the unit. The pack of cards is a kind of miniature of the population for sampling purposes. The cards are shuffled a number of times and then a card is drawn at random from them. This is one of the most reliable methods of selecting a random sample.

b) 'Mechanical Randomization' or 'Random Numbers' Method.

The lottery method described above is quite time consuming and cumbersome to use if the population is sufficiently large. The most practical and inexpensive method of selecting a random sample consists in the use of 'Random Number Tables', which have been so constructed that each of the digits 0, 1, 2, ... 9 appear with approximately the same frequency and independently of each other. If we have to select a sample from a

population of size N (≤ 99) then the numbers can be combined two by two to give pairs from 00 to 99. Similarly if $N \leq 999$ or $N \leq 9999$ and so on, then combining the digit three by three (or four by four and so on) we get numbers from 000 to 999 or (0000 to 9999) and so on. Since each of the digit 0, 1, 2, ... 9 occurs with approximately the same frequency and independently of each other, so does each of the pairs 00 to 99 or triplets 000 to 999 or quadruplets 0000 to 9999 and so on.

The method of drawing the random sample consists in the following steps:

(i) Identify the N units in the population with the numbers from 1 to N.

(ii) Select at random, any page of the random number tables and pick up the numbers in any row or column or diagonal at random.

(iii) The population units corresponding to the numbers selected in step (ii) constitute the random sample.

We give below different sets of random numbers commonly used in practice. The numbers in these tables have been subjected to various statistical tests for randomness of a series and their randomness has been well established for all practical purposes.

1) Tippet's (1972) Random Numbers Tables

Tippet number tables consist of 10,400 four digited numbers, giving in all 10,400 x 4 ,ie) 41,600 digits selected at random from the British Census report.

2) Fisher and Yates (1938) Tables comprise 15,000 digits arranged in ~~Kendrick~~ Fotwos. Fisher and yates obtained these tables by drawing numbers at random from the 10th to 19th digits of A.S. Thomoson's 20-figure logarithmic tables.

3) Kendall and Babington Smith's (1939) random tables consist of 1,00,000 digits grouped into 25,000 sets of 4 digited random numbers.

4) Rand Corporation (1955) random number tables consist of one million random digits consisting of 2,00,000 random numbers of 5 digits each.

Eg: 7.1 Draw a random sample (without replacement) of size 15 from a population of size 500.

Solution:

First of all we identify the 500 units in the population with the numbers from 1 to 500 then we select at random one page of random numbers from any of the random number series discussed in std starting at random with any number on that page and moving row wise, column wise or diagonally we select one by one the three digited numbers, discarding the numbers over 500, untill 15 numbers below 500 are obtained. Since here we have selected the random sample without replacement. the numbers obtained previously (in earlier selection) will also be discarded. Finally the units in the population, corresponding to these 15 numbers will constitute our random sample without replacement.

The following is an extract from the first set of 40 four digited numbers in Tippet's random number tables

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2370	7483	3408	2762	3563	1089	6913	7691
0560	5246	0112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Starting with first number and moving column-wise, the units in the population with numbers:

295, 416, 237, 056, 275, 266, 074, 052, 491, 413

241, 460, 431, 408, 112 will be desired sample of size 15 without replacement.

Notations and Terminology:

Let us consider a (finite) population of N units and let y be the character under consideration The capital letters are used to describe the characteristics of the population whereas small letters refers to sample observation. Thus for example the N population units may be denoted by $U_1, U_2, \ldots, U_N$ and the n sample units will be denoted by $u_1, u_2, \ldots, u_3$. Let $y_i (i = 1, 2, \ldots, N)$ be the value of the character for the ith unit in the population and the corresponding small letters $y_i, (i = 1, 2, \ldots n)$ denote the value of the character for ith unit selected in the sample. Then we define, population mean $= \bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} y_i \rightarrow 7.4$

$$\text{Sample mean} : \bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Sample mean $\bar{y}_n$ may also be written alternatively as

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{N} a_i Y_i$$

where

$$a_i = \begin{cases} 1, & \text{if } i\text{th unit is included in the sample} \\ 0, & \text{if } i\text{th unit is not included in the sample.} \end{cases}$$

S^2 = Mean square for the population.

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

s^2 = Mean square for the sample.

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

The population parameters will usually be denoted by either the capital letters of the English alphabet or by Greek letters and their estimates which are functions of the sample, observation are denoted by either small letters or by putting the symbol caps (^) on the corresponding parameter. Thus, $\hat{\bar{Y}}_N$ indicates the estimate of the population mean.

Theorems.

Theorem 1: In simple random sampling without replacement (SRSWOR), the sample mean is an unbiased estimate of the population mean, i.e.

$$E(\bar{y}_n) = \bar{Y}_N$$

Proof:

$$E(\bar{y}_n) = E\left[\frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right] \quad \left[\text{from sample mean } y_n = \frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right]$$

$$E(\bar{y}_n) = \frac{1}{n}\sum_{i=1}^{N} E(a_i) Y_i \rightarrow (1)$$

Since a_i takes only two values 1 and 0,

$$E(a_i) = 1 \cdot P(a_i = 1) + 0 \cdot P(a_i = 0)$$

$= 1 \cdot P$(ith unit is included in a sample of size n)

$+ 0 \cdot P$(ith unit is not included in a sample of size n)

$$= 1 \cdot \frac{n}{N} + 0\left(1 - \frac{n}{N}\right)$$

$$E(a_i) = \frac{n}{N} \rightarrow (2)$$

Hence, (2) sub in (1)

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

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

$E(\bar{y}_n) = \bar{Y}_N$ as desired.

Theorem 2:

Statement:

In SRSWOR, the sample mean square is an unbiased estimate of the Population mean square (or)

$$E(s^2) = S^2 \rightarrow (1)$$

Proof:

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

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

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

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

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

$n-1/n$

$$= \frac{1}{n-1}\left(1-1/n\right)\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j$$

$$= 1/n\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} y_i y_j \rightarrow (2)$$

Expectation on both sides.

$$E(s^2) = \frac{1}{n}E\left(\sum_{i=1}^{n} y_i^2\right) - \frac{1}{n(n-1)}E\left(\sum_{i\neq j=1}^{n} y_i y_j\right) \rightarrow (3)$$

we have,

$$E\left(\sum_{i=1}^{n} y_i^2\right) = E\left(\sum_{i=1}^{N} a_i Y_i^2\right) = \sum_{i=1}^{N} E(a_i) Y_i^2$$

Since a_i takes only two values 0 and 1

$$E(a_i) = 1 \cdot P(a_i = 1) + 0 \cdot P(a_i = 0)$$

$$= 1 \cdot P[\text{ith unit is included in a sample of size } n] +$$

$$0 \cdot P[\text{ith unit is not included in a sample of size } n]$$

$$= 1 \cdot \frac{n}{N} + 0 \cdot \left[1 - \frac{n}{N}\right]$$

$$E(a_i) = \frac{n}{N}$$

$$E\left[\sum_{i=1}^{n} y_i^2\right] = \frac{n}{N} \sum_{i=1}^{N} y_i^2 \rightarrow (4)$$

$$E\left[\sum_{i \neq j=1}^{n} y_i y_j\right] = E\left[\sum_{i \neq j=1}^{N} a_i a_j y_i y_j\right]$$

$$= \sum_{i \neq j=1}^{N} E(a_i a_j)\, Y_i Y_j$$

Now,

$$E(a_i a_j) = 1 \cdot P(a_i a_j = 1) + 0 \cdot P(a_i a_j = 0)$$

$$= P(a_i = 1) \cdot P(a_j = 1)$$

$$= P(a_j = 1) / P(a_i = 1)$$

$$E(a_i a_j) = \frac{n(n-1)}{N(N-1)}$$

Because probability of $(a_i = 1)$ = P [i^{th} unit is included in the sample of size $n = n/N$ and

$P(a_j = 1) / P(a_i = 1)$ = P (j^{th} unit is included in the sample given that i^{th} unit is included in the sample)

$$E\left(\sum_{i\neq j=1}^{n} y_i y_j\right) = \frac{n(n-1)}{N(N-1)} \sum_{i\neq j=1}^{N} Y_i Y_j \rightarrow (5)$$

Sub from (4) and (5) in (3) we get,

$$E(s^2) = \frac{1}{n}\left[\frac{n}{N}\sum_{i=1}^{N} Y_i^2\right] - \frac{1}{n(n-1)}\left[\frac{n(n-1)}{N(N-1)}\sum_{i\neq j=1}^{N} Y_i Y_j\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \frac{1}{N(N-1)}\sum_{i\neq j=1}^{N} Y_i Y_j$$

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

using (2)

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

Hence the proof.

Theorem : 3:

Statement:

In SRSWOR the variance of the sample mean is given by

$$var(\bar{y}_n) = \frac{N-n}{N}\cdot\frac{S^2}{n} \rightarrow (1)$$

Proof:

we have, $var(\bar{y}_n) = \frac{N-n}{N}\cdot\frac{S^2}{n} \rightarrow (1)$

$$var(x) = E(x^2) - [E(x)]^2$$

$$var(\bar{y}_n) = E(\bar{y}_n^2) - [E(\bar{y}_n)]^2$$

$$= E(\bar{y}_n^2) - \bar{y}_n\bar{y}_n^2 \rightarrow (2)$$

$$E(\bar{y}_n^2) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$$

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

$$= \frac{1}{n^2}\left[E\left(\sum_{i=1}^{n} y_i^2\right) + E\left(\sum_{i\neq j=1}^{n} y_i y_j\right)\right] \rightarrow (3)$$

$$(1) \Rightarrow E\left(\sum_{i=1}^{n} y_i^2\right) = \sum_{i=1}^{N} E(a_i)\, Y_i$$

$$E\left(\sum_{i=1}^{n} y_i^2\right) = n/N \sum_{i=1}^{N} Y_i^2$$

But

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2$$

$$(N-1)S^2 = \sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2$$

$$(N-1)S^2 + N\bar{Y}_N^2 = \sum_{i=1}^{N} Y_i^2$$

$$E\left(\sum_{i=1}^{n} y_i^2\right) = n/N\left[(N-1)S^2 + N\bar{Y}_N^2\right]$$

$$= n\left[\frac{(N-1)S^2}{N} + \frac{N\bar{Y}_N^2}{N}\right]$$

$$E\left(\sum_{i=1}^{n} y_i^2\right) = n\left[\frac{(N-1)S^2}{N} + \bar{Y}_N^2\right] \rightarrow (4)$$

$$(2) \Rightarrow E\left(\sum_{i\neq j=1}^{n} y_i y_j\right) = \sum_{i\neq j=1}^{n} E(a_i a_j)\, Y_i Y_j$$

$$= \frac{n(n-1)}{N(N-1)} \sum_{i\neq j=1}^{N} Y_i Y_j$$

$$= \frac{n(n-1)}{N(N-1)}\left[\left[\sum_{i=1}^{N} Y_i\right]^2 - \sum_{i=1}^{N} Y_i^2\right]$$

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

$$= \frac{n(n-1)}{N(N-1)} \left[N(N-1)\,\bar{Y}_N^2 - (N-1)S^2 \right]$$

$$E\left(\sum_{i+j=1}^{n} y_i y_j\right) = n(n-1)\left[\bar{Y}_N^2 - \frac{S^2}{N}\right] \longrightarrow (5)$$

Sub from equation number (4) and (5) in (3) we get

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

$$= \frac{1}{n^2} \cdot n\left[\frac{(N-1)S^2}{N} + \bar{Y}_N^2\right] + 1/n^2\left[n(n-1)\right]\left[\bar{Y}_N^2 - S^2/N\right]$$

$$= \frac{1}{n}\left[\frac{(N-1)S^2}{N} + \bar{Y}_N^2\right] + \frac{n-1}{n}\left[\bar{Y}_N^2 - \frac{S^2}{N}\right]$$

$$= \frac{1}{n}\left[(1 - n/N)S^2 + \bar{Y}_N^2\right] + (1 - 1/n)\left(\bar{Y}_N^2 - S^2/N\right)$$

$$= \frac{S^2}{n} - \frac{S^2}{nN} + \frac{\bar{Y}_N^2}{n} + \bar{Y}_N^2 - \frac{\bar{Y}_N^2}{n} - \frac{S^2}{N} + \frac{S^2}{nN}$$

$$= \bar{Y}_N^2 + \frac{S^2}{n} - \frac{S^2}{N}$$

$$E(\bar{y}_n^2) = \bar{Y}_N^2 + (1/n - 1/N)S^2 \longrightarrow (6)$$

Sub (6) in (2) we get,

$$\text{var}(\bar{y}_n) = E(\bar{y}_n^2) - \bar{Y}_n^2$$

$$\text{var}(\bar{y}_n) = \bar{Y}_N^2 + (1/n - 1/N)S^2 - \bar{Y}_N^2$$

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

Hence the proof.

MERITS AND LIMITATIONS OF SIMPLE RANDOM SAMPLING

MERITS 1. Since the sample units are selected at random giving each unit an equal chance of being selected, the element of subjectivity or personal bias is completely eliminated. As such a ~~sam~~ simple random sample is more representative of the population as compared to the judgement or purposive sampling.

2. The statistician can ascertain the efficiency of the estimates of the parameters by considering the sampling distribution of the statistics (estimates), e.g., $\bar{y}_n$ as an estimate of $\bar{Y}_N$ becomes more efficient as sample size n increases.

Limitations 1. The selection of a simple random sample requires an up-to-date frame, i.e, a completely catalogued population from which samples are to be drawn. Frequently, it is virtually impossible to identify the units in the population before the sample is drawn and this restricts the use of simple random sampling technique.

2. Administrative Inconvenience. A simple random sample may result in the selection of the sampling units which are widely spread geographically and in such a case the cost of collecting the data may be much in terms of time and money.

3. At times, a simple random sample might give most non-random looking results. For example, if we draw a

random sample of size 13 from pack of cards, we may get all the cards of the same suit However, the probability of such an outcome is extremely small.

4. For a given precision, simple random sampling usually requires larger sample size as compared to stratified random sampling discussed in the next section.