M.Sc. (STATISTICS) (COURSE CODE: 138)

Paper – III: SAMPLING THEORY

SYLLABUS

UNIT – I

Basic concepts of sampling: Population, Sample, Sampling unit, Sampling frame, Complete enumeration survey, Sample Survey, Random number tables, Sampling errors, Non-sampling errors and sources, Important aspects at the planning stage of sample surveys, Statistical organizations-Central Statistical Organization (CSO) and National Sample Survey Organization (NSSO)

UNIT – II

Simple random sampling with and without replacements, Estimation of population mean, population total and population proportion in SRS without replacement and variances of these estimates. Determination of sample size in sampling from attribute data and variable data. Stratified random sampling, estimation of population mean and population total and population proportion and variances of these estimates, Allocation problems in stratified sampling, Gain in precision due to stratification. Determination of sample size in proportional and Neyman allocations.

UNIT – III

Systematic sampling, Variance of the estimated mean, concept of circular systematic sampling. Cluster sampling with equal cluster sizes, Variance of estimated mean, Optimum cluster size for fixed cost

UNIT – IV

PPS sampling with replacement, Procedures of selection of sample, Estimation of population total and its variance. Two-stage sampling with equal number of second stage units, Estimation of population mean, its variance and estimation of variance. Concept of multi-stage sampling.

UNIT – V

Ratio estimation, bias of the ratio-estimator, comparison of the ratio estimate with the mean per unit, conditions for optimum ratio estimate, ratio estimates in stratified sampling, regression estimation, Comparison with ratio estimate, regression estimates in stratified sampling.

Note: Two Questions are to be set from each unit

CONTENTS

CONTENTS

CHAPTER 1 BASIC CONCEPTS OF SAMPLING THEORY

NOTES

OBJECTIVES

After going through this chapter, you should be able to :

- know the basic concepts of sampling.
- know about principles of sampling theory.
- explain probability and non-probability sampling.
- to know how to use random number tables.

STRUCTURE

NOTES

1.1 INTRODUCTION

Population and Sample

In a statistical investigation, the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called population. Thus in statistics, population is an aggregate of objects, animate or inanimate under study. The population may be finite or infinite. 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.

Need for Sampling

For any statistical investigation, complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the average per capita income of the people in India, we will have to enumerate all the earning individuals in the country, which is rather a very difficult task. If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection, then 100% inspection is not desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes, namely administrative and financial implications, time factor etc. and we take the help of sampling. The situations which give rise to sampling are enumerated below :

1. When results with maximum accuracy or reliability with fixed budget, or with the minimum number of units with a specified degree of reliability are required.
2. When the units under investigation show considerable variation for the characteristic under study.
3. When the scope of the investigation is very wide and the population is not completely known.
4. When the total count of the population is not possible or is costly or destructive.

Complete Enumeration Survey

The total count of all units of the population for a certain characteristic is known as **complete enumeration,** also termed **census survey.** The money, manpower and time required for carrying out complete enumeration will

generally be large and there are many situations with limited means where complete enumeration will not be possible. There are also instances where it is not practicable to enumerate all units due to their perishable nature where recourse to selection of a few units will be helpful.

Sample Survey

When only a part, called a sample is selected from the population and examined, it is called sample enumeration or sample survey.

A sample survey will usually be less expensive than a census survey and the desired information will be obtained in less time. This does not imply that economy is the only consideration in conducting a sample survey. It is most important that a degree of accuracy of results is also maintained. Occasionally, the technique of sample survey is applied to verify the results obtained from a census survey. In many situations, a well conducted sample survey can provide much more precise results than a census survey. The advantages of sample surveys over census surveys are

1. Reduced cost of survey.
2. Greater speed of getting results.
3. Greater scope.
4. Adaptability.

Despite the above advantages, sample surveys are not always preferred to census surveys. The advantages of sampling over complete enumeration can be derived only if :

1. The units are drawn in a scientific manner.
2. An appropriate sampling technique is used.
3. The size of units selected in the sample is adequate. If information is required for each unit then census is the only answer.

1.2 PRINCIPLES OF SAMPLING THEORY

The main aim of sampling theory is to make sampling more effective so that the answer to a particular question is given in a valid, efficient and economical way. The theory of sampling is based on three important principles.

I. Principle of Validity

This principle states that the sampling design provides valid estimates about population parameters. By valid, we mean that the sample should be so selected that the estimates could be interpreted objectively and in terms of probability. Thus, the principle ensures that there is some definite and pre-assigned probability for each individual in the sampling design.

NOTES

II. Principle of Statistical Regularity

This principle can be explained in the following words :

"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. This principle stresses upon the desirability and importance of selecting sample designs where inclusion of sampling units in the sample is based upon probability theory.

III. Principle of Optimization

This principle takes into account the desirability of obtaining a sample design which gives optimum results. In other words, optimization is meant to develop methods of sample selection and of estimation which provide.

(*i*) a given level of efficiency with the minimum possible resources or

(*ii*) a given value of cost with maximum possible efficiency.

Thus the principle of optimization minimizes the risk or loss of sampling design, that is, the principle stresses upon obtaining optimum results with minimization of the total loss in terms of the cost and mean square error.

1.3 PRINCIPLE STEPS IN A SAMPLE SURVEY

I. Statement of Objectives

In a sample survey, the first step is to lay down a clear statement of objectives in the survey. The user should ensure that these objectives are commensurable with available resources in terms of money, manpower and the time limit of the survey.

II. Definition of Population

The population from which the sample is to be drawn should be defined in clear and unambiguous terms. For example, to estimate the average yield per plot for a crop, it is necessary to define the size of the plot in clear terms. The sampled population (population to be sampled) should coincide with the target population (population about which information is required). The demographic, geographical, administrative and other boundaries of the population must be specified so that there remains no ambiguity regarding the coverage of the survey, that is, the survey becomes a feasible process.

NOTES

III. Determination of Sampling Frame and Sampling Units

The main requirement of sample surveys is to fix up the sampling frame, that is, the list of all sampling units with reference to which relevant data are to be collected. It is the sampling frame which determines the sampling structure of a survey. The population should be capable of division into units which are distinct, unambiguous and non-overlapping and cover the entire population.

IV. Selection of Proper Sampling Design

If an appropriate sampling design is selected, the final estimates will be quite reliable. The size of the sample, procedure of selection and estimation of parameters along with the amount of risk involved are some of the important statistical aspects which should receive careful attention. If a number of sampling designs for taking a sample are available, then the total risk that is the cost and precision should be considered before making a final selection of the sampling design.

V. Organization of Field Work

The achievement of the aims of a sample survey depends to a large extent on reliable field work. If field work is sincerely according to the instructions laid down and if there is a careful supervision of the field staff, there remains no doubt about achieving the aims of the survey. It is therefore, necessary to make provisions for adequate supervisory staff for inspection of field work.

VI. Summary and Analysis of Data

In a sample survey, the final step of the analysis and drawing inferences from a sample to a population is a very vital and fascinating issue. Since the results of the survey are basis for policy making, it is the most essential part of the sample survey and should be handled properly. The analysis of the data collected in a survey may be broadly classified as follows :

(*a*) Scrutiny and editing of the data. (*b*) Tabulation of the data.

(*c*) Statistical analysis. (*d*) Reporting and conclusions.

Finally, a report of the findings of the survey, suggesting possible action to be taken, should be written.

NOTES

1.4 SAMPLING UNIT (DESCRIPTION)

These units may be natural units of the population such as individuals in a locality, or natural aggregates of such units such as family, or they may be artificial units such as a farm etc. Before selection the sample, the population must be divided into parts which are distinct, unambiguous and non-overlapping such that every element (smallest component part in which a population can be divided) of the population belongs to one and only one sampling unit. Since the collection of all sampling units of a specified type constitutes a population, the sampling units should be so specified that each and every element in the population occurs just in one sampling unit. Otherwise some of the elements will not be included in any sample. For example, if the sampling unit is a family, it should be so defined that an individual does not belong to two different families nor should it leave out any individuals belonging to it.

1.5 SAMPLING FRAME

A complete list of sampling units which represents the population to be covered is called sampling frame popularly known as frame. The construction of a sampling frame is sometimes a major problem. Generally, it is assumed that a frame is perfect if it is exhaustive, complete and up to date in respect of sampling units and character structures. So the frame should always be made up to date and free from errors of omission of and duplication of sampling units. A sampling frame is subject to several types of defects which may be broadly classified as follows :

I. A Frame May be Incomplete

When some sampling units of the population are either completely omitted or included more than once, then the frame is said to be incomplete.

II. A Frame May be Inaccurate

When some of the sampling units of the population are listed inaccurately or some units which do not actually exist are listed in the frame.

III. A Frame May be Inadequate

When it does not include all classes of the population which are to be taken in the survey.

NOTES

IV. A Frame May be out of Date

When it has not been updated according to the exigencies of the occasion although it was accurate, complete and adequate at the time of construction.

1.6 PROBABILITY AND NON-PROBABILITY SAMPLING

I. Probability Sampling

This is the method of selecting samples according to certain laws of probability in which each unit of the population has some definite probability of being selected in the sample.

It is to be noted here that there are number of samples of specified types S_1, S_2, ... , S_n that can be formed by grouping units of a given population and each possible sample S_i has assigned to it, a known probability of selection p_i. A clear specification of all possible samples of a given type along with their probabilities of selection is said to constitute a sampling design.

II. Non-Probability Sampling (Purposive)

This is the method of selecting samples in which the choice of selection of sampling units depends entirely on the discretion or judgment of the sampler. This is called non-probability or purposive or judgment sampling. In this technique, the sampling units are selected with some definite purpose in view. For example, if we want to give the picture that the standard of living has increased in the city of New Delhi, we may take individuals in the sample from rich and posh localities like defence colony, south extension, greater kailash etc., and ignore the localities where low income group and the middle class families live. This sampling method is used for opinion surveys, but cannot be recommended for general use as it is subject to drawbacks of prejudice and bias of the sampler. However, if the sampler is experienced and an expert, it is possible that judgment sampling may yield useful results. It however suffers from a serious drawback that it is not possible to compute the degrees of precision of the estimate from the sample values.

1.7 COMPARISON OF SAMPLE AND COMPLETE ENUMERATION SURVEY

NOTES

The information on a population may be collected in two ways. Either every unit in the population is enumerated called complete enumeration or enumeration is limited to only a part or a sample selected from the population called sample survey. A sample survey will usually be less costly than a complete census because the expense of covering all units would be greater than of covering only a sample fraction. Also it will take less time to collect and process data from a sample than from a census. But economy is not only the consideration. The most important point is whether the accuracy of the results would be adequate for the end in view. It is observed that the results obtained from carefully planned and well executed sample survey are expected to be more accurate than those from a complete census. A complete census ordinarily requires a huge and unwieldly organization and therefore many types of errors creep in which cannot be controlled adequately. In a sample survey the volume of the work is reduced considerably and it becomes possible to employ persons of higher caliber, train them suitably and supervise their work adequately. In a properly designed sample survey it is also possible to make a valid estimate of the margin of error and hence decide whether the results are sufficiently accurate. A complete census does not reveal by itself the margin of uncertainty to which it is subject. But there is not always choice of one versus other. For example, if the data are required for every small administrative area in a country, no sample survey can deliver the desired information but only the complete census.

Some of the aspects involved in planning and execution of a sample survey may be classified as follows :

I. Specification of Data Requirements

While specifying the data requirements, the sampler should always include the following points :

(*a*) Statistical statement of the desired information.

(*b*) Clear specification of the domain of study.

(*c*) Form of data which are to be collected and limitations of budget.

(*d*) Degree of precision aimed at.

II. Survey Reference and Reporting Periods

From the operational point of view, it is desirable to decide about these periods well in advance. The survey period is the time period during which the required data are collected. It is advisable to divide the survey period

into shorter sub periods to ensure an even representation of the sample. The reference period is the time period to which the data information should refer. The reporting period is the time period for which the information is collected for a unit.

III. Method of Data Collection

The planning and execution of a survey is influenced to a large extent by the method of data collection. After a very careful examination of the sampling frame, design, budget and objectives of the survey, a decision should be taken regarding the choice of the method of data collection that is to collect primary data or to use secondary data. In case the primary data is to be collected, a clear cut mode of collection should be given as to whether data are to be collected by personal interview, mail enquiry, physical measurement etc.

IV. Processing of Survey Data

Processing of the collected data in a survey may be broadly classified as :

(*a*) Scrutiny and editing of data.

(*b*) Tabulation of data.

(*c*) Statistical analysis.

It is therefore, necessary to plan the survey work in such a way that the flow of work material through various stages of data processing ensures the desired degree of precision in survey results.

V. Preparation of Reports

The guidelines as formulated by United Nations (1949) for preparation of sample survey reports should be adequate for the purpose. The report may have sections such as objectives, scope, subject coverage, method of data collection, survey reference and reporting periods, sampling design and estimation procedure, presentation of results, accuracy, and cost structure. It should be useful in giving a summary of the main results which can be used by the financial agency for policy decisions.

VI. Methods of Selecting a Random Sample

Since the theory of sampling is based on the assumption of random sampling, the technique of random sampling is of basic significance. Some of the procedures used for selecting a random sample are as follows :

VII. Lottery Method

In practice a ticket/chit may be associated with each unit of the population. Thus each sampling unit has its identification mark from one to N. The

NOTES

procedure of selecting an individual is simple. All the tickets/chits are placed in a container, in which a thorough mixing or reshuffling is possible, before each draw. Draw of tickets/chits may be continued until a sample of the required size is obtained.

This procedure of numbering units on tickets and selecting one after reshuffling becomes cumbersome when the population size is large. It may be rather difficult to achieve a thorough shuffling in practice. Human bias and prejudice may also creep in this method.

1.8 RANDOM NUMBER TABLES

A random number table is an arrangement of digits 0 to 9, in either a linear or a rectangular pattern, where each position is filled with one of these digits. A table of random numbers is so constructed that all the numbers 0, 1, ..., 9 appear independent of each other. Some random number tables in common use are :

(*i*) Tippet's random number tables.

(*ii*) Fisher and Yates tables.

(*iii*) Kendall and Smith tables.

(*iv*) A million random digits.

A practical method of selecting a random sample is to choose units one by one with the help of a table of random numbers. By considering two digit numbers, we can obtain numbers from 00 to 99 all having the same frequency. Similarly, three or more digit numbers may be obtained by combining three or more rows or columns of these tables. The simplest way of selecting a sample of the required size is by selecting a random number from 1 to N and then taking the unit bearing that number. This procedure involves a number of rejections since all the numbers greater than N appearing in the table is not considered for selection. The use of random numbers is therefore modified and some of these modified procedures are :

I. Remainder Approach

Let N be a r-digit numbers and its r-digit highest multiple be N. A random number k is chosen from 1 to N and the unit with the serial number equal to the remainder obtained on dividing k by N is selected. If the remainder is zero, the last unit is selected. As an illustration let N = 123, the highest three digit multiple of 123 is 984. For selecting a unit, one random number from 001 to 984 has to be selected. Let the random number selected be 287. Dividing 287 by 123, the remainder is 41. Hence the unit with serial number 41 is selected in the sample.

NOTES

II. Quotient Approach

Let N be an r-digit numbers and let its r-digit highest multiple be N' such that $N'/N = q$. A random number k is chosen from 0 to $(N' - 1)$. Dividing k by q the quotient r is obtained and the unit bearing the serial number $(r - 1)$ is selected in the sample. As an illustration, let $N = 16$ and hence $N' = 96$ and $q = 96/16 = 6$. Let the two-digit random number chosen be 65 which lies between 0 to 95. Dividing 65 by 6, the quotient is 10 and hence the unit bearing (10–1) = 9 is selected in the sample.

III. Independent Choice of Digits

This method of independent choice was suggested by Mathai in 1954. It consists of the selection of two random numbers which are combined to form one random number. One random number is chosen according to the first digit and other according to the remaining digits of the population size. If the number chosen is zero, the last unit is chosen. But if the number made up is greater than or equal to N, the number is rejected and the operation is repeated.

1.9 SAMPLING ERRORS AND THEIR SOURCES

Sampling error arises from the fact that samples differ from their populations in that they are usually small sub-sets of the total population. Therefore, survey sample results should be seen only as estimations. Henry (1990) note that sampling errors cannot be calculated for non-probability samples, but they can be determined for probability samples. First, to determine sample error, look at the sample size. Then, look at the sampling fraction—the percentage of the population that is being surveyed. The more people surveyed, the smaller the error. This error can also be reduced, according to Fox and Tracy (1986), by increasing the representativeness of the sample. A standard error is the extent to which we expect a calculation from a sample of individuals to differ from what the same calculation would be if we had information for the whole population. Alternatively, it can be thought of as the amount of variation around the estimated value due to the fact that only a sample of people was taken. Fortunately, sampling error can be quantified with reasonable accuracy.

Standard errors for estimates of means and proportions can be used in two ways: constructing confidence intervals and performing hypothesis tests. Strictly speaking, a confidence interval refers to what one would expect to occur under repeated sampling. So, a 95 per cent confidence interval means that, if a given value was the true population value then in 95 per cent or

NOTES

19 out of 20 repeated samples, we would expect the estimated value from a sample to lie within the range of the confidence interval. Although it is not strictly true, a 95 per cent confidence interval can also be thought of as the area bounding an estimate within which there is only a 5 per cent chance that the true population percentage is outside this range.

Instead of putting a range around the estimate, we may also want to test whether a certain estimate is different from another value. This value may be one we come up with, or another estimated value from the survey. To see whether this difference is significant in a statistical sense, a researcher can perform a hypothesis test. They would come up with a hypothesis, and then decide whether they can or cannot reject this hypothesis based on the estimated value from the sample, and the associated standard error.

The most fundamental point and interval estimation process involves the estimation of a population mean. Suppose it is of interest to estimate the population mean, μ, for a quantitative variable. Data collected from a simple random sample can be used to compute the sample mean, $\bar{x}$, where the value of $\bar{x}$ provides a point estimate of μ. When the sample mean is used as a point estimate of the population mean, some error can be expected owing to the fact that a sample, or subset of the population, is used to compute the point estimate. The **absolute value** of the difference between the sample mean, $\bar{x}$, and the population mean, μ, written $|\bar{x} - \mu|$ is called the **sampling error.** Interval estimation incorporates a probability statement about the magnitude of the sampling error. The sampling distribution of $\bar{x}$ provides the basis for such a statement.

Random sampling is not some sort of magical talisman that protects an investigator from all errors, rather it is a way of predicting the likely effects of one particular kind of sampling error. As we said above, the error whose probability is expressed in the confidence and precision figures is that error of accidently getting a sample which is not exactly representative of the population. There are a great many other sources of error in research against which random sampling gives no protection, nor does it estimate their likelihood.

One common such source of error in this class is survey questions which do not, for one reason or the other, provide measures of what you wanted to measure. For example students might lie in answering a question about academic honesty or they simply might not remember accurately how many times they had used the library in the last month. Another uncontrolled type of error is experimental error. A research assistant records the wrong time for a rat's performance in a maze or a scale is miscalibrated and hence reports the wrong weights. In addition, Random sampling is no protection against sampling a biased population. For example a written questionnaire

is biased towards those who are literate in the language of the questions. A phone survey is biased toward those who have phones and who are home to answer them.

NOTES

1.10 NON-SAMPLING ERRORS AND THEIR SOURCES

Non-sampling errors can be defined as errors arising during the course of survey activities rather than resulting from the sampling procedure. Unlike sampling errors, there is no simple and direct method of estimating the size of non-sampling errors. In most surveys, it is not practical to measure the possible effect on the statistics of the various potential sources of error arising from things other than the statistical sample. However, there has been a considerable amount of research on the kinds of errors that are likely to arise in different kinds of surveys. By examining the procedures and operations of a specific survey, experienced survey analysts may be able to assess its quality. Rarely will this produce actual error ranges, as for sampling errors. In most cases, the analyst can only state that, for example, the errors are probably relatively small and will not affect most conclusions drawn from the survey, or that the errors may be fairly large and inferences are to be made with caution. In rare instances, researchers may be able to say with some confidence in what direction the error might be.

Non-sampling errors can be classified into two groups: random errors whose effects approximately cancel out if fairly large samples are used; and biases which tend to create errors in the same direction and thus cumulate over the entire sample. With large samples, systematic errors, and resultant biases, are the principal causes for concern about the quality of a survey. For example, if there is an error in the questionnaire design, this could cause problems with the respondent's answers, which in turn, can create processing errors, etc. These types of errors often lead to a bias in the final results and analyses. In contrast to sampling variance and random non-sampling error, bias caused by systematic non-sampling errors cannot be reduced by increasing the sample size.

Non-sampling errors can occur because of problems in coverage, response, non-response, data processing, estimation and analysis. An error in coverage occurs when there is an omission, duplication or wrongful inclusion of the units in the population or sample. Omissions are referred to as under-coverage, while duplication and wrongful inclusions are called over-coverage. These are caused by defects in the survey frame: inaccuracy, incompleteness, duplication, inadequacy and obsolescence. There may be errors in sample selection, or part of the population may be omitted from the sampling frame,

NOTES

or weights to compensate for disproportionate sampling rates may be omitted. Coverage errors may also occur in field procedures.

Response errors result from data that have been requested, provided, received or recorded incorrectly. The response errors may occur because of inefficiencies with the questionnaire, the interviewer, the respondent or the survey process. Subject matter experts are often in a good position to identify flaws in such aspects of the survey. Poor questionnaire design is a common aspect of non-sampling error. It is essential that sample survey or census questions are worded carefully in order to avoid introducing bias. If questions are misleading or confusing, the responses may end up being distorted. As alluded to above, an interviewer and facilitators can influence how a respondent answers the survey questions. This may occur when the interviewer is too friendly or aloof or prompts the respondent. To prevent this, interviewers must be trained to remain neutral throughout the interview. They must also pay close attention to the way they ask each question. If an interviewer changes the way a question is worded, it may impact on the respondent's answer.

Respondents can also provide incorrect answers by their own volition. Faulty recollections (recall bias), tendencies to exaggerate or underplay events, and inclinations to give answers that are more 'socially desirable', are several reasons why a respondent may provide a false answer. Individuals may conceal the truth out of fear or suspicion of the survey process and the institutions sponsoring it (*i.e.*, governments and their agencies). Other respondent errors may arise through a failure to understand the underlying concepts or a basic lack of knowledge about the information requested. Non-sampling errors can also arise from the survey process. Using proxy responses or a lack of control over the survey procedures are just two ways of increasing the possibility of response errors. Processing errors sometimes emerge during the preparation of the final data files. For example, errors can occur while data are being coded, captured, edited or imputed. Coder bias is usually a result of poor training or incomplete instructions, variance in coder performance, data entry errors, or machine malfunction. Sometimes, errors are incorrectly identified during the editing phase. Even when errors are discovered, they can be corrected improperly because of poor imputation procedures. Non-response errors—another category of non-sampling error—can also result from having not obtained sufficient answers to survey questions. Complete non-response errors occur when the survey fails to measure some of the units in the selected sample. Reasons for this type of error may be that the respondent is unavailable or temporarily absent; the respondent is unable or refuses to participate in the survey; or the dwelling is vacant. If a significant number of people do not respond to a survey, the results may be biased, since the characteristics of the non-respondents may differ from those who have participated.

NOTES

1.11 IMPORTANT ASPECTS OF PLANNING STAGE OF SAMPLE SURVEYS

The very important and main objective of a sample survey is to obtain information about population which is a group of units defined according to the objectives of the survey. The population may consist of all the households in a village/locality, all the fields under a particular crop, a population of persons, families, fields, animals in a region, or a population of trees, birds in a forest depending upon the nature of data required. The information that we seek about the population is normally, the total number of units, aggregate values of various characteristics, averages of these characteristics per unit, proportions of units possessing specified attributes etc. Our next step is to collect a data which can be collected in two different ways:

(*i*) complete enumeration method

(*ii*) sampling method

The first approach can be considered as its special case. A sampling method is a scientific and objective procedure of selecting units from the population and provides a sample that is expected to be representative of the population. A sampling method makes it possible to estimate the population totals, averages or proportions while reducing at the same time the size of survey operations. Some of the advantages of sample surveys as compared to complete enumeration are reduction in cost, greater speed, wider scope and higher accuracy. Having agreed to go for sample surveys, the next question is how to draw the sample ?

1.12 PLANNING AND EXECUTION OF SAMPLE SURVEYS

Sample surveys are widely used as a cost effective instrument of data collection and for making valid inferences about population parameters. Three major stages of a survey are planning, data collection and tabulation of data. Some of the important aspects requiring attention at the planning stage are as follows:

(*i*) formulation of data requirements-objectives of the survey

(*ii*) ad-hoc or repetitive survey

(*iii*) method of data collection

(*iv*) questionnaire versus schedules

(*v*) survey, reference and reporting periods

(*vi*) problems of sampling frames

NOTES

(*vii*) choice of sampling design

(*viii*) planning of pilot survey

(*ix*) field work

(*x*) processing of data, and

(*xi*) preparation of report.

All the different aspects listed above are important and interdependent. We now explain the above mentioned planning aspects in detail:

1. Formulation of Data Requirements

The persons or organization requiring the statistical information are expected to formulate the objectives of the survey. The user's formulation of data requirements is not likely to be adequately precise from the statistical point of view. It is for the survey statistician to give a clear formulation of the objectives of the survey and to check up whether his formulation faithfully reflects the requirements of the users. The survey statistician's formulation of data requirements should include the following:

(*i*) clear statement of the desired information in statistical terms

(*ii*) specification of the domain of study

(*iii*) the form in which the data should be tabulated

(*iv*) the accuracy aimed at in the final results and

(*v*) cost of survey

2. Survey: Adhoc or Repetitive

An adhoc survey is one which is conducted without any intention of or provision for repeating it, whereas a repetitive survey is one, in which data are collected periodically for the same, partially replaced or freshly selected sample units. If the aim is to study only the current situation, the survey can be an adhoc one. But when changes or trends in some characteristics over time are of interest, it is necessary to carry out the survey repetitively.

3. Methods of Collecting Primary Data

There are several methods that may be used to collect information. The method to be followed has to be decided keeping in view the cost involved and the precision aimed at. The methods usually adopted for collecting primary data are described below:

(*i*) Direct Personal Interview

(*ii*) Questionnaires Sent Though Mail

(*iii*) Interviews by Enumerators

(*iv*) Telephone Interview

NOTES

4. Questionnaire Vs. Schedule

In the questionnaire approach, the informants or respondents are asked pre-specified questions and their replies to these questions are recorded by themselves or by investigators. In this case, the investigator is not supposed to influence the respondents. This approach is widely used in main enquiries. In the schedule approach, the exact form of the questions to be asked are not given and the task of questioning and soliciting information is left to the investigator, who backed by the training and instructions has to use his ingenuity in explaining the concepts and definitions to the informant for obtaining reliable information. While planning a survey, preparation of questionnaire or schedules with suitable instructions needs to be given careful consideration. Respondent's bias and Investigator's bias are likely to be different in the two methods. Simple, unambiguous suitable wordings as well as proper sequence of questions are some considerations which contribute substantially towards reducing the respondents bias. Proper training, skill of the Investigators, suitable instructions and motivation of investigators contribute towards reducing Investigator's bias.

5. Survey, Reference and Reporting Periods

(*a*) **Survey period:** The time period during which the required data is collected. Reference period: The time period to which the collective data for all the units should refer.

(*b*) **Reporting period:** The time period for which the required statistical information is collected for a unit at a time (reporting period is a part or whole of the reference period). The reporting period should be decided after conducting suitable studies to examine recall errors and other non-sampling errors. For items of information subject to seasonal fluctuations, it is desirable to have one complete year as the survey and reference period, the data being collected every month or season with suitable reporting periods for the same or different sets of sample units.

6. Sampling Frames

One of the main requirements for efficiently designing sample survey is a well constructed sampling frame. In actual practice, quite often frames are not always perfect. Various types of imperfection such as omission, duplication etc. exist in the available frame. In multi-stage sampling, the problems of securing a good sampling frame arise fro each of the stages. Usually a frame for higher stage units, such as towns, urban blocks and villages is more stable than one for lower stage units such as farms and households, which are more subject to changes. In agricultural surveys, normally the frames of first few stages of units upto village level are used from records while the

NOTES

frame of households, fields etc. within the villages are prepared afresh. This approach reduces the chances of imperfection in sampling frames.

7. Choice of Sampling Design

The choice of a suitable sampling design for a given survey situation is one of the most important step in the process of planning sample surveys. The principle generally adopted in the choice of a design is either reduction of overall cost for a pre-specified permissible error or reduction of margin of error of the estimates for given fixed cost. Generally a stratified uni-stage or multi-stage design is adopted for large scale surveys. For efficient planning, various auxiliary information which are normally available are utilized at various stages *e.g.,* the area under particular crop as available for previous years is normally used for size stratification of villages. If the information is available for each and every unit of the population and there is wide variability in the information then it may be used for selecting the sample through probability proportional to size methods. The choice of sampie units, method of selecting sample and determination of sample size are some of the important aspects in the choice of proper sample design.

8. Pilot Surveys

Where some prior information about the nature of population under study, and the operational and cost aspects of data collection and analysis is not available from part surveys. It is desirable to design and carry out a pilot survey. It will be useful for

(*i*) testing out provisional schedules and related instructions

(*ii*) evolving suitable procedure for field and tabulation work, and

(*iii*) training field and tabulation staff

9. Field Work

While planning the field work of the survey, a careful consideration is needed regarding choice of the field agency. For adhoc surveys, one may plan for adhoc staff but if survey is going to be a regular activity, the field agency should also be on a regular basis. Normally for regular surveys, the available field agency is utilized. A regular plan of work by the enumerators along with proper supervision is an important consideration for getting a good quality of data.

10. Processing of Survey Data

The analysis of data collected in a survey has broadly two facets:

(*i*) tabulation and summary of data and

(*ii*) subject analysis

The first task which is of primary importance is the reduction of collected data into meaningful tables. The tables should be presented along with the background information such as the objective(s) of the survey, the sampling design adopted, method used for data collection and tabulation, and margin of error applicable to the results. These margins of error provide the idea about the precision of estimates. Subject analysis to be taken up after preparing summary tables, should include cross tabulation of data by the meaningful, geographical, economy, demographic or other breakdowns to study their relationship and trends among various characteristics. This is a detailed technical analysis and is likely to be time consuming. Hence the part should not be tied up with the first part as otherwise the publication of the survey results might get delayed.

1.13 PREPARATION OF REPORT

Although there are no set guidelines for presentation of results and preparation of report, however some points which serve as guidelines in the preparation of sample survey reports are given below:

(*i*) Introduction and review of literature

(*ii*) Objectives

(*iii*) Scope

(*iv*) Subject coverage

(*v*) Method of data collection

(*vi*) Survey references and recording

(*vii*) Sampling design and estimation procedure

(*viii*) Tabulation procedure

(*ix*) Presentation of results

(*x*) Activity of results

(*xi*) Cost structure of the survey

(*xii*) Agency for conducting the survey

(*xiii*) References.

STATISTICAL ORGANISATIONS

1.14 CENTRAL STATISTICAL ORGANISATION (CSO)

The Central Statistical Organisation is a body which is responsible for coordination of statistical activities in the country, and evolving and

NOTES

maintaining statistical standards. Its activities include National Income Accounting; conduct of Annual Survey of Industries, Economic Censuses and its follow up surveys, compilation of Index of Industrial Production, as well as Consumer Price Indices for Urban Non-Manual Employees, Human Development Statistics, Gender Statistics, imparting training in Official Statistics, Five Year Plan work relating to Development of Statistics in the States and Union Territories; dissemination of statistical information, work relating to trade, energy, construction, and environment statistics, revision of National Industrial Classification, etc. It has a well-equipped Graphical Unit. The CSO is headed by the Director-General who is assisted by 2 Additional Director-Generals and 4 Deputy Director-Generals, Directors and Joint Directors and other supporting staff. The CSO is located in Delhi. Some portion of Industrial Statistics work pertaining to Annual Survey of industries is carried out in Calcutta.

Activities of the CSO

The following are the main activities of the organization:

1. The Enterprise Survey Unit (ESU) of the Economic Census and Surveys Division of CSO is responsible for carrying out certain Follow-up Surveys relating to non-agricultural activities. For carrying out these surveys. Economic Census frame giving count of enterprises at village/block level is used for selection of sample villages/blocks to the extent feasible.
2. Under the Follow-up Surveys, the activities of Mining and Quarrying, Manufacturing, Trade, Hotels and Restaurants, Transport, Storage and Warehoushing and Services have so far been covered.
3. Data in these surveys are collected by interview method. However, for enterprises maintaining books of accounts, information are collected on the basis of these records. A moving reference period is adopted for collection of most of the information to reduce the recall lapse. The survey period is generally of one year duration and this period is divided into 4 sub-rounds of three months each to fully capture the effect of seasonability.
4. The detailed information collected in the Follow-up Surveys relate to employment, emoluments, fixed capital, working capital, receipts, expenses, source of finance, outstanding loan, etc.
5. The ESU has so far carried out 12 Follow-up Surveys. In all 27 reports based on these surveys have been brought out. List of surveys along with the reports released is enclosed.

Objectives of the CSO

The CSO holds meetings and conferences from time to time. The main objectives of these conferences are:

- To provide a platform for discussion on the statistical issues of common interest to the Central Statistical Organisations;
- To provide an overall perspective to the development of statistical system and to make recommendations/suggestions on issues having bearing on the development of the statistical system;
- To solve the technical issues relating to statistics;
- To set up Working Groups on specific issues/tasks relating to official statistics;
- To provide guidelines in the collection of statistics and maintenance of statistical standards and quality, besides uniformity in statistical standards;
- To consider the Action Taken Report of the follow up action on the recommendations of the previous meetings of CSO; and
- To review the role of the Statistical Advisers in the Central Governments.

1.15 NATIONAL SAMPLE SURVEY ORGANISATION (NSSO)

The National Sample Survey Organisation, is an organization in the Ministry of Statistics and Programme Implementation of the Indian Government. This organisation is the largest one in India and has been conducting regular socio-economic surveys. Below, we give its origin and explain its functions:

Origin and Functions of NSSO

Initiated in 1950, the National Sample Survey (NSS), is a nation-wide, large-scale continuous survey operation conducted in the form of successive rounds. It was established on the basis of a proposal from P.C. Mahalanobis to fill up data gaps for socio-economic planning and policy making through sample surveys. In March 1970, the NSS was reorganised and all aspects of its work were brought under a single Government organisation, namely the National Sample Survey Organisation (NSSO) under the overall direction of Governing Council to impact objectivity and autonomy in the matter of collection, processing and publication of the NSS data.

The Governing Council consisted of experts from within and outside the Government and was headed by an eminent economist/statistician and the member-Secretary of the Council was Director General and Chief Executive Officer of NSSO. The Governing Council was empowered to take all technical

NOTES

decisions in respect of the survey work, from planning of survey to release of survey results. After the formation of National Statistical Commission in 2005, the Governing Council of NSSO has been dissolved from 30.08.2006 and in its place Steering Committee of National Sample Surveys has been formed on 15 December 2006. The Steering Committee consisted of 8 Non-official and 8 Official members. The Non-official members are men of eminence in either of the fields of economics, statistics and social sciences. The official members are all senior officers of the Ministry of Statistics and Programme Implementation, Planning Commission and State Directorates of Economics and Statistics. The Director General and Chief Executive Officer (DG and CEO) of NSSO is the Convener of the Steering Committee. The NSSO is headed by the Director General and Chief Executive Officer (DG and CEO) who is responsible for coordinating and supervising all activities of the organisation and is assisted by a small secretariat called Co-ordination and Publication Division (CPD). The NSSO has four divisions, namely, Survey Design and Research Division (SDRD), Field Operations Division (FOD), Data Processing Division (DPD) and Coordination and Publication Division (CPD). An Additional Director General heads each division except CPD, which is headed by a Deputy Director General.

Survey Design and Research Division (SDRD)

SDRD which is located at Kolkata, is responsible for the following activities:

- Planning of the Survey
- Formulation of sampling design
- Formulation of Concepts and Definitions
- Drawing of survey schedules
- Writing of instructions
- Preparation of validation and tabulation programmes
- Finalisation of survey results and release of reports
- Providing technical guidance on sampling techniques to various official agencies.

Field Operations Division (FOD)

FOD has its headquarters at New Delhi and Agricultural Wing at Faridabad. It has 6 Zonal offices located at Bangalore, Kolkata, Guwahati, Jaipur, Lucknow and Nagpur, 48 Regional offices and 117 sub-regional offices. This division is responsible for the following functions:

NOTES

- Collection of data through (*i*) Annual Survey of Industries (*ii*) Socio-economic surveys, (*iii*) Price collection surveys.
- Updation of Urban Frame Survey blocks.
- Sample check on area enumeration and crop cutting experiments and providing technical guidance to the states for improvement of crop statistics.
- Providing in-service training to NSSO officials.

Data Processing Division (DPD)

The Data Processing Division (DPD) of NSSO with Headquarters at Kolkata and six Data Processing Centres located at Ahmedabad, Bangalore, Kolkata, Delhi, Giridih and Nagpur provide complete IT solution from sample section, software development to processing and tabulation of data canvassed through various socio-economic surveys of National Sample Survey Organisation. This division is responsible for the following functions:

(*i*) Maintenance of Sampling Frame and updation of Urban Frame Survey database.

(*ii*) Selection of samples and preparation of sample lists for conducting socio-economic surveys undertaken by the NSSO.

(*iii*) Preparation of Data Processing Training Manual and Organising All India and D.P. Centre wise Data Processing Training Workshops.

(*iv*) Receipt of schedules from field and their checking vis-à-vis sample list.

(*v*) Scrutiny of filled-in schedules is undertaken during initial period of the survey for obtaining first hand knowledge of the quality of survey operation as well as for finalization of data processing system taking into consideration the field reality.

(*vi*) Manual checking of identification particulars and extensive scrutiny of schedules before Data-Entry, popularly known as pre-data entry scrutiny.

(*vii*) Data entry and verification of filled in schedules.

(*viii*) Validation of data through various stages covering both content check and coverage check, for which in-house validation packages are developed by the Information Technology (IT) Wing of the Division.

(*ix*) Organisation of Scrutiny Feedback Workshops in collaboration with SDRD and FOD.

(*x*) Preparation of directory and multiplier files for estimation of parameters.

NOTES

(*xi*) Tabulation of validated data as per approved Tabulation Plan for various schedules.

(*xii*) Processing and tabulation of monthly rural retail price schedules and release of quarterly bulletin.

(*xiii*) Assistance to State Statistical Bureaus in processing of state sample data through facilities available at various centres of DPD to the States.

(*xiv*) Providing training in application of computer and on data processing to the State.

(*xv*) Statistical Bureaus, ISS probationers, ISEC trainees etc. Besides in-house training programmes on computer and administrative rules and procedures are organized for the staff and officers of DPD.

Coordination and Publication Division (CPD)

CPD is located at New Delhi and is responsible for the following causes:

(*i*) Coordinating the activities of all the Divisions of NSSO.

(*ii*) Dissemination of survey results and analysis through the biannual technical journal 'Sarvekshana' and 'National Seminars' to discuss the survey results.

(*iii*) Providing technical and secretarial assistance to Steering Committee of National Sample Surveys.

(*iv*) Supplying survey data of various rounds to individuals, researchers, research institutions and other private and government bodies.

(*v*) Liaison with other Departments/Ministries on various matters concerning NSSO.

(*vi*) Providing the technical and secretarial assistance to DG and CEO of NSSO Procedures for obtaining NSS Reports

(*vii*) Copies of NSS Reports can be obtained from the Deputy Director General, SDRD, NSSO, Ministry of Statistics and Programme Implementation, Mahalanobis Bhawan, 164 G.L. Tagore Road, Calcutta-700108 on payment basis.

(*viii*) Copies of NSS Reports can also be obtained from the Deputy Director General, CPD, NSSO, Ministry of Statistics and Programme Implementation, Sardar Patel Bhawan, Parliament Street, New Delhi-110001 on payment basis.

(*ix*) To obtain the report(s) a Demand Draft in favour of Pay and Accounts Officer, Ministry of Statistics and Programme Implementation, payable

at Kolkata/New Delhi, is to be sent to the above noted offices at Kolkata or New Delhi respectively.

(*x*) The cost of survey report (hard/soft copy) includes the cost of media but are exclusive of collection costs.

(*xi*) Inland subscribers willing to obtain the report by post must send the postal charges at following rates in addition to the cost of the reports Speed Post Rs. 85 per copy and Registered Post Rs. 30 per copy.

(*xii*) Subscribers from abroad will be intimated about the exact postal charges on receipt of their request for the NSS Reports.

Procedures for obtaining NSSO Data

(*i*) Obtaining data on payment

(*a*) Validated unit level data relating to various survey rounds are available on *CD*-ROMS which can be obtained from the Deputy Director General, Computer Centre, *M/O* Statistics and *PI*, East Block No. 10 R.K. Puram, New Delhi-110066 by remitting the price along with packaging and postal charges as well as giving an undertaking duly signed in a specified format. (Click to get the format of undertaking)

(*b*) The amount is to be remitted by way of demand draft drawn in favour of Pay and Accounts Officer, Ministry of Statistics and Programme Implementation, payable at New Delhi.

(*ii*) Obtaining data free of cost by signing Memorandum of Understanding (MOU)

(*a*) The approved research Institutes and Universities can obtain NSSO data free of cost for research/studies concerning national development and planning by signing a Memorandum of Understanding (MOU)

(*b*) A specific proposal containing the following details has to be sent to the Director General and Chief Executive Officer, NSSO by the intending Research Institute/University:

1. Name and address of the University/Research Institute
2. Names, addresses of the Researchers/Project investigators
3. Academic background of the researchers
4. Title and synopsis of the research project(s)
5. Specific requirements of NSSO data
6. Expected data of completion of the research project
7. Any other details considered relevant by the researcher.

NOTES

(iii) Obtaining data free of cost for research projects duly approved by NSSO

(*a*) The approved research Institutes and Universities can obtain NSSO data free of cost for specific research projects by getting them approved by the NSSO.

(*b*) The research Institute/University is required to sign an agreement in specified format for the purpose.

(*c*) The Institute/University must send a proposal containing all the details (*i* to *vii*) above The University will sign an agreement instead of a Memorandum of Understanding (MOU).

(*d*) The data along with the necessary documentations will be made available to the researchers if the proposal is approved and the agreement is signed.

NSSO is Involved in Three Types of Surveys

1. Socio-economic surveys
2. Annual Survey of Industries
3. Agricultural surveys

In case of socio-economic surveys the entire responsibility of the survey from its design to release of the report is the responsibility of NSSO. In case of the other two activities the organisation is entrusted with collection of data from the field only. The Socio-economic surveys involve Survey Design. Field Operations. Processing of data collected and releasing of the results based on surveys. The organisation is guided by a Steering Committee established by the Government. The selection of the subject of the survey, questionnaire design, instructions to field staff etc. are finalised by the Steering Committee. The subjects on which surveys are conducted vary over the years. Important subjects like, employment-unemployment and consumer expenditure, are regularly covered. The organisation is headed by a Director General and Chief Executive Officer.

NSSO consists of four divisions, each headed by an Additional Director General:

1. Field Operation Division
2. Data Processing Division
3. Survey Design and Research Division
4. Coordination and Publication Division

NOTES

The head quarter of NSSO is located at New Delhi. It has a few regional headquarters also. The Field Operations Division has its headquarters at Delhi and Faridabad with a network of six zonal offices 48 regional Offices and 117 sub-regional offices spread throughout the country. The Data Processing Division with its headquarters in has Data Processing Centres at Delhi, Giridih, Nagpur, Bangalore, Ahmedabad and Kolkata. This division is entrusted with the responsibilities of data entry, processing and tabulation of socio-economic data. The Survey Design and Research Division having responsibility of designing the questionnaire, sample design, etc. and analyse the results has its headquarters at Kolkata. The Coordination and Publication Division is located at New Delhi.

Apart from the regular agencies engaged in data collection, a few other organisations are also actively involved in information collection on educational variables through their large-scale household sample surveys among which the National Sample Survey Organisation (NSSO) is the prominent one. The NSSO in its 52nd round (July 1995–June 1996) carried out a nation wide survey on social consumption to ascertain the extent of utilisation of facilities in the field of Education and Health. The present survey was in continuation of the two surveys on social consumption carried out by the NSSO as a part of its 35 round (July 1980–June 1981) and 42nd round (July 1986–June 1987). The survey has generated few interesting indicators that are otherwise not available from regular sources. Literacy rate (15+ population), distribution of persons (aged 15 years and above) by level of education attainment, distribution of students of age group 5–24 by primary, middle, secondary/higher secondary and higher education, gross, net (classes I–V and VI–VIII) and age-specific (age 6–10 and 11–13 years) attendance ratio, proportion of students getting free education by level of education, average amount of per capita private expenditure on education and proportions of attending, attended and never attended children by age group are some of the important indicators. The entire document contains 17 statistical statements and a few charts.

STUDENT ACTIVITY 1

1. Compare and contrast census survey and sample survey. Which survey do you find the best? Justify your comments.

STUDENT ACTIVITY 2

1. Mention three main principles of sampling theory. Discuss their advantages and disadvantages.

NOTES

1.16 SUMMARY

- 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.
- The total count of all units of the population for a certain characteristic is known as complete enumeration, also termed census survey.
- A complete list of sampling units which represents the population to be covered is called sampling frame popularly known as frame.
- A clear specification of all possible samples of a given type along with their probabilities of selection is said to constitute a sampling design.
- This is the method of selecting samples in which the choice of selection of sampling units depends entirely on the discretion or judgment of the sampler. This is called non-probability or purposive or judgment sampling.
- Since the theory of sampling is based on the assumption of random sampling, the technique of random sampling is of basic significance.
- A random number table is an arrangement of digits 0 to 9, in either a linear or a rectangular pattern, where each position is filled with one of these digits. A table of random numbers is so constructed that all the numbers 0, 1, ..., 9 appear independent of each other.

1.17 GLOSSARY

- **Population**—In statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called population.
- **Sample**—A portion of a population, serving as a basis for estimates of the attributes of the whole population.
- **Sampling Frame**—The main requirement of sample surveys is to fix up the sampling frame, that is, the list of sampling units with reference to which relevant data are to be collected. It is the sampling frame which determines the sampling structure of a survey.
- **Probability Sampling**—This is of selecting samples according to certain laws of probability in which each unit of the population has some definite probability of being selected in the sample.
- **Survey Period**—It is the time period during which the required data are collected.
- **Sample Survey**—A part or a sample selected from the population.

NOTES

1.18 REVIEW QUESTIONS

1. What is the need of sampling?
2. Define population, sampling unit and sampling frame for conducting surveys of the following subjects. Mention, other possible sampling units, if any, in each case and discuss their relative merits :
   (*i*) Election for assembly with adult franchise.
   (*ii*) Annual yield of apple fruit in H.P.
   (*iii*) Housing conditions in a rural area.
3. Explain; what do you understand by probability sampling and non-probability sampling. What are their relative advantages and disadvantages?
4. Compare sample survey and complete enumeration survey. Give some examples in support of your answer.
5. Write a short note on the following :
   (*a*) Lottery method
   (*b*) Random number tables.

1.19 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

NOTES

CHAPTER 2 SIMPLE RANDOM SAMPLING

OBJECTIVES

After going through this chapter, we should be able to :

- know simple random sampling.
- calculate values of the sample mean.
- explain expected value and sampling variance of the mean.
- know sample mean and sample mean square.

STRUCTURE

2.1 INTRODUCTION

The main purpose of statistical survey is to obtain information about the population. For this purpose, we require data which can be collected either by complete enumeration method or sampling method. The sampling method

is more scientific and objective procedure of selecting units from the population and it provides a sample which is expected to be representation of the population as a whole. The distinctive feature of survey by the use of sampling methods called sample surveys sampling error. It gives the discrepency between the sample estimates and the population values which would be obtained from enumerating all the units in the population in the same way in which the sample is enumerated. These discrepencies are unavoidable because sample estimates are based on data for only a sample of units. The sampling method enables us to estimate the average magnitude of these discrepencies to be made.

The purpose of sampling theory is to make sampling more efficient. It attempts to develop method of sample selection and of estimation that provide, at the lowest cost, estimates that are precise enough for our purpose. This principle of specified precision at minimum cost occurs repeatedly in the presentation of the theory. In order to apply this principle, we must be able to predict, for any sampling procedure which is under consideration, the precision and the cost to be expected. So far as the precision is concerned, one cannot tell exactly how large an error will be present in an estimate in any specific situation. This will require the knowledge of the true value of the population. Instead, the precision of a sampling procedure is judged by examining the frequency distribution generated for the estimate if the procedure is applied again and again to the same population. A further simplification can be introduced. With the samples of sizes which are common in practice, there is often good reason to suppose that the sample estimates are approximately normally distributed. With a normally distributed estimate, the whole shape of the frequency distribution is known if we know the mean and standard deviation (or the variance). A considerable part of the sample survey thoery is therefore concerned with the findings of these means and variances.

2.2 SIMPLE RANDOM SAMPLING WITH AND WITHOUT REPLACEMENT

In this method, an equal chance of selection is assigned to each available unit of the population at the first and each subsequent draw.

Thus, if the number of units in the population is N, the probability of selecting any unit at the first draw is $1/N$, the problem of selecting any unit from among the available units at the second draw is $\frac{1}{N-1}$ and so on. The sample so obtained is called **simple random sample.**

NOTES

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

The probability that the specified unit is selected at the r^{th} draw is clearly the product of (i) the probability of the event that it is not selected in any of the previous $(r-1)$ draws and (ii) the probability of the event that it is selected at the r^{th} draw under the assumption that it is not selected in any of the previous $(r-1)$ draws.

Now the probability that it is not selected at the first draw $= \frac{N-1}{N}$.

$\therefore$ Probability that it is not selected at the second draw given that it was not selected at the first draw $= \frac{N-2}{N-1}$ and so on.

The probability (i) is $\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{N-3}{N-2} \cdot \ldots\ldots \frac{N-(r-1)}{N-(r-2)} = \frac{N-r+1}{N}$.

The probability (ii) is $\frac{1}{N-(r-1)}$.

Thus, the probability that the specified unit of population is selected at the r^{th} draw $= \frac{N-r+1}{N} \cdot \frac{1}{N-r+1} = \frac{1}{N}$, which is the probability of its being selected at the first draw and is independent of r.

The simple way of obtaining a probability sample is to draw the units one by one with a known probability of selection assigned to each unit of the population at the first and each subsequent draw. The succesive draws may be made with or without replacements of the units selected in the preceeding draws. The former technique is called sampling with replacement and the later sampling without replacement.

2.3 NOTATIONS

We shall assume that the sampling units are drawn without replacement. We denote the following symbols.

N : The number of sampling units in the population.

y : The characteristic under consideration

y_i : The value of the characteristic for the i^{th} unit of population.

$\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$, the mean value of the characteristic per unit of population.

NOTES

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N\,\bar{y}_N{}^2\right]$, the mean square for the population.

$V(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 = \frac{N-1}{N}S^2$, the variance of single observation in the population.

n : The sample size *i.e.*, the number of units in the sample.

$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean.

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$, the sample mean square.

2.4 ESTIMATION OF POPULATION MEAN

We have $$E[\bar{y}_n] = E\left\{\frac{1}{n}\sum_{i=1}^{n} y_i\right\} \quad ...(1)$$

We may number the units in the sample serially as 1, 2, ..., r, ..., n in the order in which they were drawn.

Thus, rewriting (1), we have

$$E[\bar{y}_n] = E\left\{\frac{1}{n}\sum_{r=1}^{n} y'_r\right\} \quad ...(2)$$

where y'_r now stands for the value of the unit included in the sample at the r^{th} draw.

$$\therefore \quad E(\bar{y}_n) = \frac{1}{n}\left[E(y'_1) + E(y'_2) + ... + E(y'_r) + ... + E(y'_n)\right] \quad ...(3)$$

Now by def. $E[y'_r] = \sum_{i=1}^{N} P_{ir}\, y_i$

where P_{ir} denotes the problem of drawing a specified unit y_i at the r^{th} draw.

But $$P_{ir} = \frac{1}{N}.$$

$$\therefore \quad E[y'_r] = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_N \ (r = 1, 2, ..., n) \quad ...(4)$$

NOTES

Using, (4) in (3), we get

$$E[\bar{y}_n] = \frac{1}{n}[n\,\bar{y}_N] = \bar{y}_N \qquad ...(5)$$

Thus, $\bar{y}_n$ is an unbiased estimate of $\bar{y}_N$.

Aliter : We can write

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

where $$\alpha_i = \begin{cases} 1, & \text{if } y_i \text{ is in the sample} \\ 0, & \text{otherwise} \end{cases}$$

$$\therefore \quad E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\left[\sum_{i=1}^{N} \{E(\alpha_i)\}\, y_i\right]$$

Now $E(\alpha_i)$ = 1 × {Probability that y_i is included in the sample} + 0 × {Probability that y_i is not included in the sample}.

$$= \frac{n}{N}$$

$$\therefore \quad E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\left[\sum_{i=1}^{N} \frac{n}{N}\, y_i\right] = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_N$$

2.5 ESTIMATION OF SAMPLE MEAN SQUARE

We have $$E[y_r'^2] = \frac{1}{N}\sum_{i=1}^{N} y_i^2 \qquad ...(1)$$

$$= \bar{y}_N^2 + \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{y}_N^2$$

$$= \bar{y}_N^2 + \frac{1}{N}\left[\sum y_i^2 - N\,\bar{y}_N^2\right]$$

$$= \bar{y}_N^2 + \frac{N-1}{N}\cdot\frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N\,\bar{y}_N^2\right]$$

$$= \bar{y}_N^2 + \left(\frac{N-1}{N}\right)S^2 = \bar{y}_N^2 + \left(1 - \frac{1}{N}\right)S^2 \qquad ...(2)$$

It follows that

$$E\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2\right] = E\left[\frac{1}{n}\sum_{r=1}^{n} y_r'^2\right] = \frac{1}{n}\left[n\,E\left[y_r'^2\right]\right]$$

$$= \bar{y}_N^2 + \left(1 - \frac{1}{N}\right)S^2 \qquad ...(3)$$

Again if y_r' and y_s' are used to denote the values of the units drawn at r^{th} and r^{th} draws respectively, say y_i and y_j then we have

$$E\left(y_r'\, y_s'\right) = \sum_{i \neq j=1}^{N} P_{ir}\, P_{js/i}\, y_i y_j \qquad ...(4)$$

where $P_{js/i}$ denotes the problem of drawing y_j at the s^{th} draw given that y_i is drawn at the r^{th} draw.

Now $P_{ir} = \frac{1}{N}$ and by extension of this result

$$P_{js/i} = \frac{1}{N-1}$$

Using these values in (4), we get

$$E\left[y_r'\, y_s'\right] = \frac{1}{N(N-1)} \sum_{i \neq j=1}^{N} y_i\, y_j \qquad ...(5)$$

It follows that

$$\frac{1}{n(n-1)} E\left[\sum_{i \neq j}^{n} y_i\, y_j\right] = \frac{1}{n(n-1)} E\left[\sum_{r \neq s=1}^{n} y_r'\, y_s'\right]$$

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

$$= \frac{1}{n(n-1)}\, n(n-1)\left[\frac{1}{N(N-1)}\left\{y_i y_j\right\}\right]$$

$$\therefore \frac{1}{n(n-1)} E\left[\sum_{i \neq j}^{n} y_i\, y_j\right] = \frac{1}{N(N-1)} \sum_{i \neq j=1}^{N} y_i\, y_j \qquad ...(6)$$

NOTES

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

$$= \frac{N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2 - \frac{1}{N(N-1)}\left[\left(\sum_{i=1}^{N} y_i^2 - N\,\bar{y}_N^2\right) + N\,\bar{y}_N^2\right]$$

$$= \frac{N}{N-1}\,\bar{y}_N^2 - \frac{S^2}{N} - \frac{\bar{y}_N^2}{N-1}$$

$$= \bar{y}_N^2 - \frac{S^2}{N} \qquad ...(7)$$

Using (3) and (7), we get

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

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

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

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

Thus, $$E\left[\bar{y}_n^2\right] = \bar{y}_N^2 + \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \qquad ...(8)$$

Now we obtain expected value of the sample mean square :

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

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

using (8)

$= S^2 \Rightarrow s^2$ is unbiased estimate of S^2.

NOTES

2.6 VARIANCES OF ESTIMATES OF POPULATION TOTAL AND PROPORTION IN SRS WITHOUT REPLACEMENT

In this section, we discuss the following two estimates:

(a) Sampling Variance of the Mean

Let $V(\bar{y}_n)$ denote the sampling variance of the mean. Then

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

Now $E(\bar{y}_n) = \bar{y}_N$ and $E(\bar{y}_n^2) = \bar{y}_N^2 + \left(\frac{1}{n} - \frac{1}{N}\right)S^2$

$$\therefore \quad V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 = \frac{N-n}{N}\cdot\frac{S^2}{n}$$

The factor $\left(\frac{N-n}{N}\right)$ is a correction for the finite size of the population and is called the finite population correction factor or simply the finite multiplier. Usually the value of S^2 will not be known. Its estimate from the sample will therefore be used in estimating the sampling variance.

Thus, Est. $V(\bar{y}_n) = \frac{N-n}{N}\cdot\frac{s^2}{n}$ $(\because E(s^2) = S^2)$

The estimate of the S.E. is given by

$$\text{Est. } S.E.(\bar{y}_n) = \sqrt{\frac{N-n}{N}}\,\frac{s}{\sqrt{n}}.$$

(b) Expected Value and Sampling Variance of s

Let $s^2 = S^2 + \varepsilon$ ...(1)

where $E(\varepsilon) = 0$ and $E(\varepsilon^2) = V(s^2)$

NOTES

$$\therefore \quad S = \left(S^2 + \varepsilon\right)^{\frac{1}{2}} = S\left[1 + \frac{\varepsilon}{S^2}\right]^{\frac{1}{2}}$$

Since, $\varepsilon = s^2 - S^2$

$$\therefore \quad E[\varepsilon^2] = E[s^4 + S^4 - 2\, s^2 S^2]$$
$$= E(s^4) + S^4 - 2\, S^2 E(s^2)$$
$$= E(s^4) - S^4$$
$$= E[(s^2)^2] - [E(s^2)]^2$$
$$= V(s^2).$$

Also $E(\varepsilon) = E(s^2) - E(S^2)$
$$= S^2 - S^2 = 0$$

Since, ε will be small as compared to S^2 with a probability approaching 1 as $n \to \infty$, we may expand the RHS as a series.

Thus, $$s = S\left[1 + \frac{1}{2}\frac{\varepsilon}{S^2} - \frac{1}{8}\frac{\varepsilon^2}{S^4} + \ldots\right]$$

Neglecting powers of ε higher than second and taking expectation on both sides, we get

$$E(s) \simeq S\left[1 - \frac{1}{8}\frac{V\left(s^2\right)}{S^4}\right] \qquad \ldots(2)$$

Now $$V(s^2) = E[(s^2)^2] - E[(s^2)]^2$$

$$= \frac{1}{(n-1)^2} E\left[\sum_{i=1}^{n}\left(y_i - \bar{y}_n\right)^2\right]^2 - S^4$$

$$= \frac{1}{(n-1)^2} E\left[\sum_{i=1}^{n} y_i^{\,2} - n\,\bar{y}_n^{\,2}\right]^2 - S^4 \qquad \ldots(3)$$

Thus, the variance of (s^2) can be calculated and from (2) $E(s)$ can be calculated.

The expression for the variance of s^2 in the limiting case can also be obtained from the moments as

$$V(s^2) = \frac{\mu_4 - \mu_2^2}{n} + \frac{2}{n(n-1)}\mu_2^2 \qquad \ldots(4)$$

Using, Pearsonian notation for departure from normality, this can be written as

$$V(s^2) = S^4\left[\frac{\beta_2 - 1}{n} + \frac{2}{n(n-1)}\right] \qquad \ldots(5)$$

where $$\beta_2 = \frac{\mu_4}{S^4}$$

For normal population $\beta_2 = 3$,

$$\therefore \qquad V(s^2) = \frac{2S^4}{n-1}$$

Thus, $$S.E.\ (s^2) = \sqrt{\frac{2}{n-1}}\, S^2 \qquad ...(6)$$

To obtain variance of s, we have

$$V(s) = E[s - E(s)^2] = E\ (s)^2 - (E(s))^2$$

$$\cong S^2 - S^2 \left[1 - \frac{1}{8}\frac{V(s^2)}{S^4}\right]^2 \simeq \frac{V(s)^2}{4S^2}$$

Using (5), we get

$$V(s) \cong \frac{S^2}{4}\left[\frac{\beta_2 - 1}{n} + \frac{2}{n(n-1)}\right]$$

For normal population $\beta_2 = 3$,

$$\therefore \qquad V(s) \cong \frac{S^4}{2(n-1)}$$

$$\therefore \qquad \text{S.E.}(s) \cong S/\sqrt{2(n-1)}$$

2.7 DETERMINATION OF SAMPLE SIZE FROM ATTRIBUTE AND VARIABLE DATA

For estimation of the population mean, we know that the mean of the random sample will be approximately normally distributed if the size of the sample is not too small and if the population from which it is drawn is not very different from the normal. We may, therefore, expect that

$$|\bar{y}_n - \bar{y}_N| \leq \sqrt{\frac{N-n}{Nn}} \cdot S \text{ on an average in 68\% cases and}$$

$$|\bar{y}_n - \bar{y}_N| \leq 2\sqrt{\frac{N-n}{Nn}} \cdot S \text{ on an average in 95\% cases.}$$

NOTES

Thus, in general, we expect the following inequality :

$$\bar{y}_n - t_{(\alpha,\infty)}\sqrt{\frac{N-n}{Nn}}\,S \le \bar{y}_N \le \bar{y}_n + t_{(\alpha,\infty)}\sqrt{\frac{N-n}{Nn}}\,S \qquad ...(1)$$

Where $t_{(\alpha,\infty)}$ is the value of the normal variate corresponding to the value $1-\frac{\alpha}{2}$ of the normal probability integral, to hold on an average with a probability $1 - \alpha$. The equation (1) gives confidence limits and the interval between these limits is called confidence interval. The probability with which the inequality holds *i.e.*, $1 - \alpha$ is called the confidence coefficient.

We now calculate the size of sample required for estimating the population parameter with a specified precision. The precision is usually specified in terms of the margin of error permissible in the estimate and the coefficient of confidence with which one wants to make sure that the estimate is within the permissible margin of error. Thus, if the error permissible is, say, $\varepsilon\bar{y}_N$ and the confidence coefficient is $1 - \alpha$ then we must know the size of the sample so that

$$P\left\{\left|\bar{y}_n - \bar{y}_N\right| \ge \varepsilon\,\bar{y}_N\right\} = \alpha \qquad ...(2)$$

Thus, from equation (1), we have

$$n = \frac{\dfrac{t^2_{(\alpha,\infty)}}{\varepsilon^2}\dfrac{S^2}{\bar{y}_N^2}}{1+\dfrac{1}{N}\dfrac{t^2_{(\alpha,\infty)}}{\varepsilon^2}\dfrac{S^2}{\bar{y}_N^2}} \qquad ...(3)$$

The determination of sample size from equation (3) pressures the knowledge of $\frac{S}{\bar{y}_N}$ which is the coefficient of variation for the population. This can only be estimately roughly. Consequently, equation (3) gives only a rough idea of the size of the sample required for estimating the population mean with a specified precision.

STUDENT ACTIVITY 1

1. Show that $\overline{y}_n$ is an unbiased estimate of $\overline{y}_N$ or prove that

$$E(\overline{y}_n) = \overline{y}_N$$

STUDENT ACTIVITY 2

1. Show that s^2 is an unbiased estimate of S^2 or prove that

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

2.8 ILLUSTRATIVE EXAMPLES

NOTES

A random sample of size 2 households was drawn from a small colony of 5 households having monthly income (in Rs.) as given below:

Household :	*1*	*2*	*3*	*4*	*5*
Income (in rupees) :	*156*	*149*	*166*	*164*	*155*

(i) Calculate population mean Y, variance (σ^2) and mean square error (S^2).

(ii) Enumerate all possible samples of size 2 by the replacement method and show that :

(a) the sample mean gives an unbiased estimate of the population mean and find its sampling variance;

(b) sample variance (s^2) is an unbiased estimate of the population variance σ^2; and

(c) $v(\bar{y}) = \dfrac{(y_1 - y_2)^2}{4}$ *is an unbiased estimator of* $V(\bar{y})$, *i.e.,*

$$Ev(\bar{y}) = V(\bar{y}), = \sigma^2/2.$$

(iii) Enumerate all possible samples of size 2 by the without replacement method and show that :

(a) the sample mean gives an unbiased estimate of the population mean and find its sampling variance;

(b) the sample variance (s^2) is an unbiased estimate of the population variance S^2; and

(c) $v(\bar{y}) = 3\,(y_1 - y_2)^2/20$ *is an unbiased estimator of* $V(\bar{y})$, *i.e.,*

$$Ev(\bar{y}) = V(\bar{y}) = \left(\frac{1}{2} - \frac{1}{5}\right)S^2 = \frac{3}{10}S^2.$$

Solution. (*i*) The population mean of 5 households is given as

$$\bar{Y} = (156 + 149 + 166 + 164 + 155)/5$$

$$= \text{Rs. } 158.00$$

The population variance is given as

$$\sigma^2 = [(156^2 + 149^2 + \ldots + 155^2) - 5 \times 158^2]/5 = 38.80$$

$$\therefore \quad S^2 = N\sigma^2/(N - 1) = 5 \times 38.8/4 = 48.50$$

(*ii*) The total number of possible samples is 25 (= 5^2).

Also, each of the 25 possible samples has the same probability (1/25) of being selected. Therefore, using the data in Table I given below.

NOTES

Table I. All samples of 2 units from 5 units in simple random sampling with replacement

Sample no.	*Units in the sample*	*Probability*	*Sample observations* y_1	y_2	*Sample mean* $(\bar{y})$	*Sampling error* $(\bar{y} - \bar{Y})$	*Sampling variance* $\left(\frac{y_1 - y_2}{4}\right)^2$
(I)	*(II)*	*(III)*	*(IV)*	*(V)*	*(VI)*	*(VII)*	*(VIII)*
1	1, 1	1/25	156	156	156.0	– 2.0	0
2	1, 2	1/25	156	149	152.5	– 5.5	49/4
3	1, 3	1/25	156	166	161.0	3.0	100/4
4	1, 4	1/25	156	164	160.0	2.0	64/4
5	1, 5	1/25	156	155	155.5	– 2.5	1/4
6	2, 1	1/25	149	156	152.5	– 5.5	49/4
7	2, 2	1/25	149	149	149.0	– 9.0	0
8	2, 3	1/25	149	166	157.5	– 0.5	289/4
9	2, 4	1/25	149	164	156.5	– 1.5	225/4
10	2, 5	1/25	149	155	152.0	– 6.0	36/4
11	3, 1	1/25	166	156	161.0	3.0	100/4
12	3, 2	1/25	166	149	157.5	– 0.5	289/4
13	3, 3	1/25	166	166	166.0	8.0	0
14	3, 4	1/25	166	164	165.0	7.0	4/4
15	3, 5	1/25	166	155	160.5	2.5	121/4
16	4, 1	1/25	164	156	160.0	2.0	64/4
17	4, 2	1/25	164	149	156.5	– 1.5	225/4
18	4, 3	1/25	164	166	165.0	7.0	4/4
19	4, 4	1/25	164	164	164.0	6.0	0
20	4, 5	1/25	164	155	159.5	1.5	81/4
21	5, 1	1/25	155	156	155.5	– 2.5	1/4
22	5, 2	1/25	155	149	152.0	– 6.0	36/4
23	5, 3	1/25	155	166	160.5	2.5	121/4
24	5, 4	1/25	155	164	159.5	1.5	81/4
25	5, 5	1/25	155	155	155.0	– 3.0	0
Average					158.0		19.40

(*a*) The expected value of $\bar{y}$ is given by the average value of column (VI), which works out to be the population mean 158.0, thus verifying that the estimator is unbiased.

We find from columns (VI) and (VII) in (Table I_ that these estimates differ, in general, from $\bar{Y}$ (= 158.0), and that the error usually varies from sample to sample.

Further, it is of interest to note from column (VII) that each sample mean ($\bar{y}$) given by column (VI) is an unbiased estimate of the population mean ($\bar{Y}$) as the average of column (VII) is zero, which proves the result. The sampling variance is the mean of the squares of error [column (VII)] which works out to 19.40 (= 38.80/2).

(*b*) The sample variance (s^2) is given by $(y_1 - y_2)^2/2$ which is twice the value given in column (VIII). Also, $E(s^2) = 38.80\sigma^2$, which shows that s^2 is an unbiased estimator for the population variance in simple random sampling, *wr*.

(*c*) An estimator of $V(\bar{y})$ is given by

$$v(\bar{y}) = (y_1 - y_2)^2/4$$

the values which are given in column (8) of Table I for possible samples. The expected value of $v(\bar{y})$ is the average of the values in column (8), showing that

$$E\,[v(\bar{y})] = V(\bar{y}) = \sigma^2/2$$

Thus, the estimate $v(\bar{y})$ is unbiased.

Table II. All samples of 2 units from 5 units in simple random sampling without replacement

Sample no.	*Units in sample*	*Probability*	*Sample observations*		*Sample mean*	*Error*	*Sampling variance*
			y_1	y_2	$(\bar{y})$	$(\bar{y} - \bar{Y})$	$\frac{3}{5} \times (y_1 - y_2)^2/4$
(I)	*(II)*	*(III)*	*(IV)*	*(V)*	*(VI)*	*(VII)*	*(VIII)*
1	1, 2	1/10	156	149	152.5	– 5.5	7.35
2	1, 3	1/10	156	166	161.0	3.0	15.00
3	1, 4	1/10	156	164	160.0	2.0	9.60
4	1, 5	1/10	156	155	155.5	– 2.5	0.15
5	2, 3	1/10	149	166	157.5	– 0.5	43.35
6	2, 4	1/10	149	164	156.5	– 1.5	33.75
7	2, 5	1/10	149	155	152.0	– 6.0	5.40
8	3, 4	1/10	166	164	165.0	7.0	0.60
9	3, 5	1/10	166	155	160.5	2.5	18.15
10	4, 5	1/10	164	155	159.5	1.5	12.15
Average					158.0		14.55

(*iii*) In case of random sampling, without replacement, the number of possible samples is $10\left[= \left(\frac{5}{2}\right)\right]$. It can be seen that each of the 10

NOTES

possible samples has the same probability (1/10) of being selected. Using the data in Table II above, we have

(*a*) The expected value of $\bar{y}$, which is given by the average of column (VI) of Table II, works out to be the population mean 158.0, thus verifying that $\bar{y}$ is an unbiased estimator of $\bar{Y}$. Further, the sampling variance which is obtained by averaging the squares of errors given in column (VII) works out to be 14.55, verifying that $V(\bar{y}) = 3\sigma^2/8$ (= $3S^2/10$).

(*b*) Since, the sample variance (s^2) is given by $(y_1 - y_2)^2/2$, which can be obtained easily, the average of 10 sample mean squares gives $E(s^2)$.

Here, $E(s^2) = 485/10 = 48.5$

Also, the population mean square = 48.5

Thus, the sample mean square provided an unbiased estimator of the population mean square (S^2), verifying that

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

(*c*) An estimator of $V(\bar{y})$ is given by

$$v(\bar{y}) = 3(y_1 - y_2)^2/20$$

the values which are given in column (VIII) of Table II and may be used here. The expected value of $v(\bar{y})$, which is the average of the values in the column (VIII), is 14.55, showing that the estimator is unbiased, *i.e.*,

$$Ev(\bar{y}) = V(\bar{y}) = \frac{3S^2}{10}$$

2.9 SUMMARY

- If the number of units in the population is N, the problem of selecting any unit at the first draw is $1/N$, the problem of selecting any unit from among the available units at the second draw is $\frac{1}{N-1}$ and so on. The sample so obtained is called **simple random sample.**
- N : The number of sampling units in the population.

  Y : The characteristic under consideration

  y_i : The value of the characteristic for the i^{th} unit of population.

  $\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$, the mean value of the characteristic per unit of population.

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N\,\bar{y}_N{}^2\right]$, the mean square for the population.

$V(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 = \frac{N-1}{N}S^2$, the variance of single observation in the population.

n : The sample size *i.e.*, the number of units in the sample.

$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean.

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$, the sample mean square.

2.10 GLOSSARY

- **Finite multiplier—**The factor $\frac{N-n}{N}$ is called finite population correction factor or finite multiplier.
- **Expected value—**It is a predicted value of a variable calculated as the sum of all possible values each multiplied by the probability of its occurrence.

2.11 REVIEW QUESTIONS

1. In simple random sampling the probability of selecting any given n units in succession in a specified order is, by definition,

$$\frac{1}{N}\cdot\frac{1}{N-1}\cdots\frac{1}{N-n+1}$$

Hence, show that everyone of the ${}^{N}C_{n}$ possible clusters of size n has an equal probability of being selected.

Show further that if simple random sampling were to be defined as a method of selecting n units such that everyone of the ${}^{N}C_{n}$ possible clusters has an equal probability of being chosen, it implies that the probability of selection assigned to each available unit of the population at the first and each subsequent draw in unit-by-unit selection is equal.

NOTES

2. The following procedure has been used for selecting a sample of fields for crop-cutting experiments on rice.

"Against the name of each selected village are shown three random numbers smaller than the highest survey number in the village. Select the survey numbers corresponding to given random numbers for experiments. If the selected survey number does not grow rice, select the next bigger rice-growing survey number in its place."

Examine whether the above method will provide an equal chance of inclusion in the sample to all the paddy-growing survey numbers in the village, given the following :

1. Name of village ..... .
2. Total number of survey numbers .....300.
3. Random numbers ..... 18, 189, 239.
4. Rice-growing survey numbers ..... 49 to 88 and 189 to 300.

Show that the survey number 49 has a chance of 49/300 of being included in the sample, the survey number 189 a chance of 101/300, while the remaining survey numbers have a chance of only 1/300 each.

(I.C.A.R. 1951)

3. Consider a simple random sample of size two drawn from a finite population (y_1, y_2, y_3). Corresponding to the three possible samples, $s_1 = (y_1, y_2)$, $s_2 = (y_1, y_3)$ and $s_3 = (y_2, y_3)$, let a linear estimate $e(s)$ for estimating the population mean be defined as follows :

$$e(s_1) = \frac{2}{3} y_1 + \frac{1}{2} y_2$$

$$e(s_2) = \frac{1}{3} y_1 + \frac{1}{2} y_3$$

$$e(s_3) = \frac{1}{2} y_2 + \frac{1}{2} y_3$$

(*a*) Prove that the estimate $e(s)$ is unbiased for all values of y_1, y_2 and y_3.

(*b*) Find the variance of the estimate $e(s)$.

(*c*) Show that there exist population values y_1, y_2 and y_3 for which the estimate $e(s)$ has a smaller variance than the sample mean $\bar{y}_n$.

(*d*) In particular, show that $V(e) < V(\bar{y}_n)$ if $y_1 = 1$, $y_2 = 2$ and $y_3 = 3$.

4. A sample of size n is drawn with equal probability and without replacement from a population of size N. Let

$$\hat{\bar{y}}_N = \sum_{r=1}^{n} ar\, yr' \qquad ...(1)$$

NOTES

be any linear estimate of the population mean $\bar{y}_N$, where the a_r are some constants and yr' denotes the value of the unit included in the sample at the r^{th} draw.

(*a*) Show that $\hat{\bar{y}}_N$ is an unbiased estimate of $\bar{y}_N$ if and only if

$$\sum_{r=1}^{n} a_r = 1 \qquad ...(2)$$

(*b*) Show that under (2) the variance of $\hat{\bar{y}}_N$ is given by

$$V\left(\hat{\bar{y}}_N\right) = \frac{S^2}{N}\left[N\sum_{r=1}^{n} a_r^2 - 1\right] \qquad ...(3)$$

where S^2 is the mean square for the population.

(*c*) Show that $V\left(\hat{\bar{y}}_N\right)$ is minimized subject to (2) if $a_r = \frac{1}{n}$ for $r = 1, 2, ..., n$. Hence prove that $\bar{y}_n$ is the best unbiased estimate of $\bar{y}_N$ in the class of linear estimates given by (1).

5. To estimate the population mean $\bar{y}_N$ of a finite population of size N, a random sample of size m fixed in advance is drawn with replacement. Let u denote the number of distinct units in the sample, the i^{th} distinct unit occurring k_i times with $\sum_{i=1}^{u} k_i = m$. Consider the estimates

$$\bar{y}_m = \frac{1}{m}\sum_{i=1}^{u} k_i\, y_i$$

and

$$\bar{y}_u = \frac{1}{u}\sum_{i=1}^{u} y_i$$

where y_i is the value of the i^{th} distinct unit in the sample.

(*a*) Show that both the estimates are unbiased and that the estimate $\bar{y}_u$ is more efficient than $\bar{y}_m$ if

$$E\left(\frac{1}{u}\right) < \frac{1}{m}\left(1 + \frac{m-1}{N}\right)$$

(*b*) Show that the probability distribution of u is given by

$$P(u) = N^{-m}\left(\frac{N}{u}\right)\Delta u\, 0m$$

where the s^{th} difference of 0^t is defined by

$$\Delta s\ 0^t = \sum_{r=0}^{s}(-1)\, s - r\left(\frac{s}{r}\right) r^t$$

NOTES

$$\Delta s\, 0^{t} = \sum_{r=0}^{s} (-1)\, s - r \left(\frac{s}{r}\right) r^{t}$$

(c) Hence or otherwise, obtain the expected values of u and $\frac{1}{u}$.

6. Show that an unbiased estimate of the variance of $\overline{y}_u$ is given by

$$\text{Est. } V(\overline{y}_u) = \left(\frac{1}{u} - \frac{1}{N}\right) \cdot s^2$$

where s^2 is the sample mean square based on u distrinct units.

7. Let a_i (i = 1, 2, ..., M) denote the i^{th} name in one list and b_j (j = 1, 2, ..., N) the j^{th} name in the second list. Assume that no name appears more than once in any list. Let D be the number of names commoin to both the lists. To estimate D, draw a simple random sample of m names from the first list and n names from the second list. Let d be the number of names common to both the samples. Find an unbiased estimate of D and its variance.

2.12 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

CHAPTER 3 STRATIFIED RANDOM SAMPLING

NOTES

OBJECTIVES

After going through this chapter, we should be able to :

- define stratified sampling.
- know Neyman allocation.
- know Optimum allocation.
- explain sample sizes in different strata.
- know about estimation of the gain in precision due to stratification.
- know about different systems of allocation.

STRUCTURE

3.1 INTRODUCTION

The precision of a sample estimate of the population mean depends not only upon the size of the sample and the sampling fraction but also on the

NOTES

variability or heterogeneity of the popultion. Thus, apart from the size of the sample, the only way of increasing the precision of an estimate is to devise the sampling procedures which will effectively reduce the heterogeneity. One such procedure is stratified sampling.

The procedure consists in dividing the population into k classes and drawing a random sample composed of k random samples, one each from the different class. The classes into which the population is divided are called **strata** and the procedure is termed as **stratified random sampling.** An example of stratified random sampling is furnished by a survey for estimating the average yield of a crop per acre in which administrative areas are taken as the strata and the random samples of predetermined numbers of fields are selected from each of the strata. The geographical proximity of fields within a stratum is expected to make it more homogeneous than the entire population and thus helps to increase the precision of the estimate.

It consists in dividing the population into k classes and drawing a sample composed of k random variables samples one each from the different classes. The classes into which the population is divided are called the **strata** and the process is termed the procedure of **stratified random sampling**.

3.2 ESTIMATE OF POPULATION MEAN AND VARIANCE

Let N_i denote the size *i.e.,* The number of units in the i^{th} stratum, n_i the size of the sample to be selected therefrom and k, the number of strata with

$$\sum_{i=1}^{k} N_i = N \quad \text{and} \quad \sum_{i=1}^{k} n_i = n \qquad ...(1)$$

where N and n denote the number of units in the population and sample respectively Further let y_{ij} be the value of the j^{th} unit in the i^{th} stratum. Then the population mean $\bar{y}_N$ can be expressed as

$$\bar{y}_N = \frac{1}{N} \sum_{i=1}^{k} N_i \; \bar{y}_{Ni} = \sum_{i=1}^{k} p_i \, \bar{y}_{N_i} \qquad ...(2)$$

where $p_i = \dfrac{N_i}{N}$ and $\bar{y}_{N_i}$ is the population mean for the i^{th} stratum.

Since, the sample in each stratum is a simple random sample, $\bar{y}_{n_i}$ is an unbiased estimate of $\bar{y}_{N_i}$ and it is natural to take

$$\bar{y}_w = \sum_{i=1}^{k} P_i \, \bar{y}_{n_i} \qquad ...(3)$$

as an estimate of the population mean which is the weighted mean of the strata sample means with strata sizes as the weights. This gives an unbiased estimate of the population mean, as

$$E(\bar{y}_w) = \left[\sum_{i=1}^{k} p_i \bar{y}_{n_i}\right]$$

$$= \sum_{i=1}^{k} p_i E(\bar{y}_{n_i}) = \sum_{i} p_i \bar{y}_{N_i} = \bar{y}_N \quad ...(4)$$

Hence, the result.

Now we obtain the sampling variance of $\bar{y}_w$.

Since, the sample in the i^{th} stratum is a simple random sample,

$$V(\bar{y}_{n_i}) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 \quad ...(5)$$

where S_i^2 is the population mean square for the i^{th} stratum and is given by

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2 \quad ...(6)$$

$$\therefore \quad V(\bar{y}_w)_s = V\left[\sum_{i=1}^{k} p_i \bar{y}_{n_i}\right]$$

$$= \sum_{i=1}^{k} p_i^2 V(\bar{y}_{n_i}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 \quad ...(7)$$

Again s_i^2, the mean square of the sample from the i^{th} stratum, provides an unbiased estimate of S_i^2. It follows that an unbiased estimate of the variance of $\bar{y}_w$ is given by

$$\text{Est } V(\bar{y}_\omega)_s = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{n_i}\right) p_i^2 s_i^2 \quad ...(8)$$

Since, it is seen from (7) that the variance of the estimated population mean in stratified sampling depends upon S_i, the variabilities within the strata. It shows that the smaller the S_i, *i.e.*, the more homogeneous the strata, the greater will be the precision of the stratified sample.

3.3 DETERMINATION OF SAMPLE SIZES IN DIFFERENT ALLOCATIONS

We know that the variance of the estimated mean in stratified sampling shows that the precision of a stratified sample for given strata depends on n_i which can be fixed at our choice. The guiding principle in the determination of n_i is to choose them in such a manner so as to provide an estimate of the population mean with the desired degree of precision for a minimum cost.

NOTES

The allocation of the sample to the different strata made in accordance with thin principle is called the principle of **optimum allocation.**

If C_i is the cost per experiment in the i^{th} stratum, then the total cost is given by

$$C_i = \sum_{i=1}^{k} C_i n_i \qquad ...(1)$$

If C_i is same from stratum λ stratum, say c, then total cost is

$$C = cn \qquad ...(2)$$

To determine optimum values of n_i, when the cost function is represented by (1), we consider the function

$$\phi = V(\bar{y}_w) + \mu C \text{ where } \mu \text{ is some constant.}$$

Now

$$V(\bar{y}_w) + \mu C = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) p_i^2 S_i^2 + \mu \left\{ \sum_{i=1}^{k} C_i n_i \right\}$$

$$= \sum_{i=1}^{k} \left(\frac{p_i S_i}{\sqrt{n_i}} - \sqrt{\mu C_i n_i} \right)^2 + \text{terms independent of } n_i. \qquad ...(3)$$

Clearly, $V(\bar{y}_w)$ is minimum for fixed cost C or the cost C of a survey is minimum for a fixed value of $V(\bar{y}_w)$, when each of the square terms on the RHS of (3) is zero.

i.e., where $n_i = \frac{P_i S_i}{\sqrt{C_i \mu}}$ $(i = 1, 2, ..., k)$ ...(4)

Equation (4) obviously, shows that :

(*i*) The larger the size of the stratum, the large should be the size of the sample to be selected therefrom.

(*ii*) The larger the variability within a stratum, the larger should be the size of the sample from that stratum and

(*iii*) The cheaper the cost per sampling unit is a stratum, the larger the sample from that stratum.

Now are obtain the exact values of n_i.

Using (4) in (1), we get

$$C_0 = \sum_{i=1}^{k} \frac{p_i S_i}{\sqrt{\mu C_i}} C_i \text{, where } C_0 \text{ is the budgeted}$$

amount within which it is desired to estimate the mean with the maximum precision.

NOTES

$$\therefore \quad \sqrt{\mu} = \sum_{i=1}^{k} p_i S_i \sqrt{C_i} / C_0$$

$$\therefore \quad n_i = \frac{p_i S_i}{\sqrt{C_i}} \cdot \frac{C_0}{\sum_{i=1}^{k} p_i S_i \sqrt{C_i}} \qquad ...(5)$$

Thus, the total sample size required for estimating the population mean with maximum precision for fixed cost C_0 is given by

$$n = \frac{C_0 \sum_{i=1}^{k} \left(p_i S_i / \sqrt{C_i} \right)}{\sum_{i=1}^{k} p_i S_i \sqrt{C_i}} \qquad ...(6)$$

The allocation of the sample using equation (5) is known as **optimum allocation**. When $C_i = c$ $(i = 1, 2, \ldots, k)$ and consequently the cost of the survey is proportional to the size of the sample, the optimum values of n_i are given by

$$n_i = n \frac{p_i S_i}{\sum_{i=1}^{k} p_i S_i} \qquad ...(7)$$

$$\left| \begin{array}{l} (\because \; C = \sum_{i=1}^{k} C_i n_i) \\ \therefore \; C_0 = C_n \end{array} \right.$$

The allocation of the sample using (7) yields the estimate of the mean with the maximum precision and this allocation is known as Neyman allocation.

When the population mean is to be estimated with a given variance, say V_0, at a minimum cost, we evaluate the constant of proportionality by substituting for n_i from (4) in

$$\sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) p_i^2 S_i^2 = V_0 \qquad ...(8)$$

$$\therefore \quad \sum_{i=1}^{k} \left[\frac{\sqrt{C_i \mu}}{p_i S_i} - \frac{1}{N_i} \right] p_i^2 S_i^2 = V_0$$

$$\Rightarrow \quad \frac{1}{\sqrt{\mu}} = \frac{\sum_{i=1}^{k} p_i S_i \sqrt{C_i}}{V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i S_i^2} \qquad \because \; p_i = \frac{N_i}{N}.$$

Thus,

$$n_i = \frac{p_i S_i}{\sqrt{C_i}} \frac{\sum_{i=1}^{k} p_i S_i \sqrt{C_i}}{V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i S_i^2} \qquad ...(9)$$

NOTES

So that the minimum sample size required for estimating the mean with fixed variance V_0 under optimum allocation is given by

$$n = \frac{\sum_{i=1}^{k} \frac{p_i S_i}{\sqrt{C_i}} \sum_{i=1}^{k} p_i S_i \sqrt{C_i}}{V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i S_i^2} \quad ...(10)$$

Put $C_i = C$, we find that the minimum sample size regarding for estimating the mean with fixed variance V_0 under Neyman allocation is given by

$$n = \frac{\left(\sum_{i=1}^{k} p_i S_i \right)^2}{V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i S_i^2}$$

3.4 GAIN IN PRECISION DUE TO STRATIFICATION

In comparing the precision of the stratified with unstratified simple random sampling, we assume that the population values of the strata means and standard deviations are known. Usually, however this will not be the case. What is available is only a stratified sample and the problem is to *estimate the gain in precision due to stratification.*

Let $n_1, n_2, \ldots, n_k$ represent the stratified sample. Then the variance of the estimated mean $\bar{y}_w$ and an unbiased estimate of its variance are given by

$$V(\bar{y}_w)_s = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) p_i^2 S_i^2 \quad ...(1)$$

and

$$\text{Est. } V(\bar{y}_w)_s = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) p_i^2 s_i^2 \quad ...(2)$$

On the other hand, if the total sample is selected as a simple random sample without stratification, the varation of the estimated population mean would be :

$$V(\bar{y}_n)_R = \frac{N-n}{N} \frac{S^2}{n} \quad ...(3)$$

The problem therefore is to estimate S^2 given $\bar{y}_{n_1}, \bar{y}_{n_2}, \ldots, \bar{y}_{n_k}$ and $s_1^2, s_2^2, \ldots, s_k^2$.

NOTES

We have

$$(N-1)\,S^2 = \sum_{i=1}^{k}(N_i-1)\,S_i^{\,2} + \sum_{i=1}^{k} N_i(\bar{y}_{N_i}-\bar{y}_N)^2 \qquad ...(4)$$

$$\therefore \quad S^2 = \frac{1}{N-1}\sum_{i=1}^{k}(N_i-1)\,S_i^{\,2} + \frac{N}{N-1}\left[\sum_{i=1}^{k} p_i\bar{y}_{n_i}^2 - \bar{y}_N^{\,2}\right] \qquad ...(5)$$

$$\because \sum_{i=1}^{k} N_i\left(\bar{y}_{N_i}-\bar{y}_N\right)^2 = \sum_{i=1}^{k} N_i\,\bar{y}_{N_i}^{\,2} + \bar{y}_N^{\,2}.\,N - 2\bar{y}_N\sum_{i=1}^{k} N_i\,\bar{y}_{N_i}$$

$$= \sum_{i=1}^{k} N_i\,\bar{y}_{N_i}^{\,2} + N\,\bar{y}_N^{\,2} - 2\bar{y}_N.\,N.\,\bar{y}_N$$

$$= N[p_i(\bar{y}_{N_i}^{\,2} - \bar{y}_N^{\,2})]$$

Also $\quad V(y_{\bar{n}_i}) = E(\bar{y}_{n_i}^{\,2}) - \bar{y}_{N_i}^{\,2}$

or $\quad \bar{y}_{N_i}^{\,2} = E(\bar{y}_{N_i}^{\,2}) - V(\bar{y}_{n_i})$

Thus, $\quad \text{Est.}\,(\bar{y}_{N_i}^{\,2}) = \bar{y}_{N_i}^{\,2} - \text{Est.}V(\bar{y}_{n_i})$

$$= \bar{y}_{n_i}^{\,2} - \left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^{\,2} \qquad ...(6)$$

$$\because V(\bar{y}_{n_i}) = \left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^{\,2}$$

Similarly, an unbiased estimate of $\bar{y}_N^{\,2}$ is given by

$$\text{Est.}\,(\bar{y}_N^{\,2}) = \bar{y}_w^{\,2} - \text{Est.}\,V(\bar{y}_w)$$

$$= \bar{y}_w^{\,2} - \sum_{i=1}^{x}\left[\frac{1}{n_i}-\frac{1}{N_i}\right]p_i^2 S_i^2 \qquad ...(7)$$

Using (5), (6) and (7) and the fact that s_i^2 provides an unbiased estimate of S_i^2, an unbiased estimate of S^2 is given by

$$\text{Est. } S^2 = \frac{1}{N-1}\sum_{i=1}^{k}(N_i-1)\,s_i^2$$

$$+\frac{N}{N-1}\times\left[\sum_i p_i\left\{\bar{y}_{n_i}^{\,2} - \left(\frac{1}{n_i}-\frac{1}{N_i}\right)s_i^{\,2}\right\} - \left\{\bar{y}_\omega^{\,2} - \sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 s_i^2\right\}\right]$$

$$= \frac{1}{N-1}\sum_{i=1}^{k}(N_i-1)\,s_i^2$$

$$+\frac{N}{N-1}\times\left[\sum_{i=1}^{k} p_i\,\bar{y}_{n_i}^2 - \sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i s_i^2\,(1-p_i) - \bar{y}_w^2\right]$$

NOTES

$$= \frac{1}{N-1}\sum_{i=1}^{k}(N_i - 1)s_i^2$$

$$+ \frac{N}{N-1} \times \left[\sum_i p_i\,(\bar{y}_{n_i} - \bar{y}_w)^2 - \sum_{i=1}^{k} p_i\,(1-p_i)\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_i^2\right]$$

Putting $N_i = Np_i$, we get ...(8)

$$\text{Est. } S^2 = \sum_{i=1}^{k} p_i s_i^2 + \frac{N}{N-1}\left[\sum_i p_i\left(\bar{y}_{n_i} - \bar{y}_w\right)^2 - \sum_{i=1}^{w} p_i\,(1-p_i)\,\frac{s_i^2}{n_i}\right] \quad ...(9)$$

$$\because \text{ Consider } \frac{1}{N-1}\sum_{i=1}^{k}(Np_i - 1)s_i^{\,2} + \frac{N}{N-1}\left[\sum_{i=1}^{k} p_i(1-p_i)\,\frac{s_i^{\,2}}{Np_i}\right]$$

$$= \frac{1}{N-1}\sum_{i=1}^{k} Np_i s_i^2 - \frac{1}{N-1}\sum_{i=1}^{k} s_i^2 + \frac{1}{N-1}\sum_{i=1}^{k} s_i^2 - \frac{1}{N-1}\sum_{i=1}^{k} p_i s_i^2 = \sum_{i=1}^{k} p_i s_i^2$$

Hence from equation (3), the unbiased estimate of $V(\bar{y}_n)_R$ based on a stratified sample is given by

$$\text{Est. } V(\bar{y}_n)_R = \frac{N-n}{Nn}\ \text{Est. } S^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)\sum_i p_i s_i^{\,2} + \frac{N-n}{(N-1)n}\left[\sum_{i=1}^{k} p_i(\bar{y}_{n_i} - \bar{y}_w)^2 - \right.$$

$$\left.\sum_{i=1}^{k} p_i(1-p_i)\frac{s_i^{\,2}}{n_i}\right] \quad ...(10)$$

Now an unbiased estimate of $V(\bar{y}_w)$ is given by

$$\text{Est. } V(\bar{y}_w)_s = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)p_i^{\,2}s_i^{\,2} \quad ...(11)$$

Hence, an unbiased estimate of the reduction in variance due to stratification is given by

$$\text{Est. } \left[V(\bar{y}_n)_R - V(\bar{y}_w)_s\right] = \sum_{i=1}^{k}\left(\frac{1}{n} - \frac{p_i}{n_i}\right)p_i s_i^{\,2}$$

$$+ \frac{N-n}{(N-1)n}\left[\sum_{i=1}^{k} p_i\left(\bar{y}_{n_i} - \bar{y}_w\right)^2 - \sum_{i=1}^{k} p_i\left(1-p_i\right)\frac{s_i^2}{n_i}\right] \quad ...(12)$$

NOTES

$$((\text{Consider } \left(\frac{1}{n}-\frac{1}{N}\right)\left[p_i s_i^2-\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i{}^2 s_i{}^2\right.$$

$$=\frac{1}{n}\sum_{i=1}^{k}p_i s_i^2-\frac{1}{N}\sum_{i=1}^{k}p_i s_i^2-\sum_{i=1}^{k}\frac{1}{n_i}p_i^2 s_i^2+\sum_{i=1}^{k}\frac{1}{Np_i}p_i^2 s_i^2$$

$$=\frac{1}{n}\sum_{i=1}^{k}p_i s_i^2-\frac{1}{N}\sum_{i=1}^{k}p_i s_i^2-\sum_{i=1}^{k}\frac{1}{n_i}p_i^2 s_i^2+\frac{1}{N}\sum p_i s_i^2$$

$$=\sum_{i=1}^{k}p_i s_i^2\left(\frac{1}{n}-\frac{p_i}{n_i}\right))$$

The ratio of (12) to (11) expressed as a percentage gives an estimate of the percentage gain in efficiency due to stratification.

These results assume a particularly simple form in the case of proportional allocation when the value of $\bar{y}_w$ is the same as that of the sample mean $\bar{y}_n$. In this case $n_i = np_i$.

$\therefore$ (12) becomes

$$\text{Est. } \left[V(\bar{y}_n)_R - V(\bar{y}_w)_P\right] = \frac{N-n}{(N-1)n}\left[\frac{1}{n}\sum_i n_i\left(\bar{y}_{n_i}-\bar{y}_w\right)^2-\frac{1}{n}\sum_{i=1}^{k}\left(1-\frac{n_i}{n}\right)s_i^2\right] \quad ...(13)$$

Assuming that the mean squares within different strata are equal

i.e., $S_i^2 = S_w^2$ (say) for $i = 1, 2, \ldots, k$ ...(14)

a better estimate of S_w^2 can be obtained by pooling the sums of squares within strata for the sample.

Thus,

$$\text{Est. } S_w^2 = s_w^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{n_i}\right)^2 \quad ...(15)$$

It can easily be verified that s_w^2 is an unbiased estimate of S_w^2. Hence a better estimate of the variance of sample mean can be obtained when a stratified sample with proportional allocation is available. This is given by

$$\text{Est } V(\bar{y}_w)_p = \frac{N-n}{Nn}s_w^2 \quad ...(16)$$

$$V(\bar{y}_w)_p = \left(\frac{1}{n}-\frac{1}{N}\right)\sum_i p_i S_i^2$$

$$\because \text{ Est. } V(\bar{y}_w)_p = \frac{N-n}{Nn}\sum_{i=1}^{k}P_i Est.S_i^2 = \frac{N-n}{Nn}s_w^2\sum_i p_i$$

NOTES

$$= \frac{N-n}{Nn} s_w^2$$

$$\text{Let } \sum_{i=1}^{k} n_i (\bar{y}_{n_i} - \bar{y}_w)^2 = (k-1)\ \bar{n}\ s_b^2 \qquad \ldots(17)$$

where $\bar{n} = \frac{n}{k}$.

Substituting s_w^2 for each s_i^2 in (13) and using (17), we get

$$\text{Est.}\left[V(\bar{y}_n)_R - V(\bar{y}_w)_P\right] = \frac{N-n}{(N-1)n}\left[\frac{1}{n}(k-1)\,\bar{n}\,s_b{}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(1-\frac{n_i}{n}\right)s_w^2\right]$$

$$= \frac{N-n}{(N-1)n^2}\left[\bar{n}(k-1)\,s_b^2 - s_w^2\left(k - \frac{\sum_{i=1}^{k} n_i}{n}\right)\right]$$

$$= \frac{N-n}{(N-1)n^2}\left[(k-1)\,\bar{n}\,s_b^2 - (k-1)s_w^2\right]$$

$$= \frac{(N-n)(k-1)}{(N-1)n^2}(\bar{n}\,s_b{}^2 - s_w^2) \qquad \ldots(18)$$

The estimate of the relative gain in efficiency due to stratification is thus obtained as the ratio of (18) and (16). *i.e.*,

$$\frac{Est.\left[V(\bar{y}_n)_R - V(\bar{y}_w)_p\right]}{Est.\,V(\bar{y}_w)_p} = \frac{N(k-1)}{(N-1)n}\left[\frac{\bar{n}s_b{}^2}{s_w^2} - 1\right] \qquad \ldots(19)$$

$$\cong \frac{k-1}{n}\left[\frac{\bar{n}s_b{}^2}{s_w^2} - 1\right] \qquad \ldots(20)$$

The quantities s_w^2 and $\bar{n}\ s_b^2$ are called the mean squares within and between strata respectively and are best calculated from the table of analysis of variance given below.

Source of variance	*d.f.*	*Sum of Squares*	*Mean squares*
Between strata	$k-1$	$\sum_{i=1}^{k} n_i(\bar{y}_{n_i} - \bar{y}_n)^2$	$\bar{n}\ s_b{}^2$
Within strata	$n-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{n_i})^2$	s_w^2
Total	$n-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_n)^2$	s^2

NOTES

The efficiency of stratification is sometimes calculated directly by comparing the overall mean square s^2 with s_w^2. The relative gain in precision is given by

$$R.G. = \frac{s^2}{s_w^2} - 1 = \frac{(n-k)s_w^2 + (k-1)\bar{n}\, s_b^{\,2}}{(n-1)\, s_w^2} - 1$$

$$= \frac{k-1}{n-1}\left(\frac{\bar{n} s_b^2}{s_w^2} - 1\right) \qquad ...(21)$$

The estimated gain in precision is $\frac{n}{n-1}$ times the estimate of gain given by equation (20).

3.5 VARIANCE OF THE WEIGHTED MEAN UNDER DIFFERENT SYSTEMS OF ALLOCATION

We know that variance of estimated population mean is given by

$$V(\bar{y}_w) = \sum_{i=1}^{k}\left[\frac{1}{n_i} - \frac{1}{n_i}\right] p_i^2 S_i^2 \qquad ...(1)$$

For optimum allocation,

$$n_i = \frac{p_i s_i}{\sqrt{C_i}} \frac{C_0}{\sum_{i=1}^{k} p_i S_i \sqrt{C_i}} \qquad ...(2)$$

Using (2) in (1), we get

$$V(\bar{y}_w)_0 = \frac{\left(\sum_{i=1}^{k} p_i S_i \sqrt{C_i}\right)^2}{C_0} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^{\,2} \qquad ...(3)$$

Now under Neyman allocation,

$$n_i = n\frac{p_i S_i}{\sum_{i=1}^{k} p_i S_i} \qquad ...(4)$$

Using (4) in (1), we get

$$V(\bar{y}_w)_N = \frac{1}{n}\left(\sum_{i=1}^{k} p_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2 \qquad ...(5)$$

For proportional allocation of sample among the strata, $n_i = np_i$,

$$\therefore \qquad V(\bar{y}_w)_P = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i S_i^2 \qquad ...(6)$$

NOTES

Occasionally, it may happen that the optimum value of n_i in a stratum may exceed N_i, the total number of units in that stratum. Under this situation, the best that can be done is to Take $n_i = N_i$ for that stratum requiring 100% sampling, while for other strata, the optimum sample size is re-calculated as

$$n_j = (C_0 - N_i C_i) \frac{p_J S_j}{\sqrt{C_j}} \frac{1}{\sum_{J=1}^{k} p_i S_j \sqrt{C_j}}; j = 1, 2, \ldots k, j \neq 1 \quad ...(7)$$

where the Σ denotes summation over all strata except the i^{th}.

The optimum variance in this case is given by

$$V(\bar{y}_w)_0 = \frac{1}{(C_0 - N_i C_i)} \left(\sum_{J=1}^{k} p_j S_j \sqrt{C_j} \right)^2 - \frac{1}{N} \sum_{J=1}^{k} p_j S_j^2 \quad ...(8)$$

The corresponding formula in the case of Neyman allocation reduces to

$$n_j = (C_0 - N_i C_i) \frac{p_j S_j}{\sqrt{C_j}} \sum_{J=1}^{k} p_j S_j \sqrt{C_j}$$

$$= (C_n - N_i C_i) \frac{p_j S_j}{\sum_{J=1}^{k} p_j S_j \sqrt{C}\sqrt{C}}$$

$$= (n - N_i) \frac{p_j S_j}{\sum_{J=1}^{k} p_j S_j}; j = 1, 2 \ldots k, j \neq i \quad ...(9)$$

and variance is

$$V(\bar{y}_w)_N = \left(\frac{1}{n - N_i} \right) \left(\sum_{J=1}^{k} p_j S_j \right)^2 - \frac{1}{N} \sum_{J=1}^{k} p_j^2 S_j^2 \quad ...(10)$$

3.6 COMPARISON OF STRATIFIED SAMPLING WITH SIMPLE RANDOM SAMPLING WITHOUT STRATIFICATION

I. Proportional Allocation

We know that the variance of the estimated population mean in stratified sampling with proportional allocation is given by

$$V(\bar{y}_w)_P = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{i=1}^{k} p_i S_i^2 \quad ...(1)$$

NOTES

The variance of mean under simple random sampling is

$$V(\bar{y}_n)_R = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \qquad ...(2)$$

where the subscript R denotes simple random sampling without stratification.

For the purpose of comparing (1) with (2), it is necessary to express S^2 in terms of S_i^2.

The total sum of squares in the population can be split up into two parts, viz, (*i*) within strata (*ii*) between strata accordingly as :

$$\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_N)^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_{N_i})^2 + \sum_{i=1}^{k} N_i\,(\bar{y}_{N_i} - \bar{y}_N)^2$$

This can be written as

$$(N-1)\,S^2 = \sum_{i=1}^{k}(N_i - 1)S_i^2 + \sum_{i=1}^{k} N_i(\bar{y}_{N_i} - \bar{y}_N)^2 \qquad ...(3)$$

For the purpose of simplification we shall assume the N_i, to be large enough to permit the approximations

$$\frac{N_i - 1}{N_i} \cong 1 \quad \text{and} \quad \frac{N-1}{N} \cong 1.$$

Dividing both sides of (3) by N and applying these approximations, we get

$$S^2 \cong \sum_{i=1}^{k} P_i S_i^{\,2} + \sum_{i=1}^{k} p_i(\bar{y}_{N_i} - \bar{y}_N)^2 \qquad ...(4) \quad \left| \frac{1}{N} = 0 \right.$$

Multipling both sides by $\frac{N-n}{Nn}$ and transferring, we get

$$V(\bar{y}_n)_R - V(\bar{y}_w)_p \cong \frac{N-n}{Nn}\sum_{i=1}^{k} p_i(\bar{y}_{N_i} - \bar{y}_N)^2 \qquad ...(5)$$

Thus, the expression shows that the more the strata differ in their population means, the larger is the gain in precision of stratified sampling with proportional allocation over simple random sampling without stratification.

II. Neyman Allocation

For the comparison of variance of the estimate under Neyman allocation and under proportional allocation, we have

NOTES

$$V(\bar{y}_w)_p - V(\bar{y}_w)_N = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i S_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} p_i S_i\right)^2 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

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

$$= \frac{1}{n}\sum_{i=1}^{k} p_i\left[S_i - \bar{S}_w\right]^2 \quad ...(6)$$

where $\bar{S}_w = \sum_{i=1}^{k} p_i S_i$

$$\because \sum_i p_i (S_i - \bar{S}_w)^2 = \sum_{i=1}^{k} p_i \left[S_i^2 + \bar{S}_w^2 - 2 s_i \bar{S}_w\right]$$

$$= \sum_{i=1}^{k} p_i S_i^2 + \left(\sum_{i=1}^{k} P_i\right)\bar{S}_w^2 - 2\left(\sum_{i=1}^{k} P_i S_i\right)\bar{S}_w$$

$$= \sum_{i=1}^{k} p_i S_i^2 + \bar{S}_w^2 - 2\bar{S}_w^2 = \sum_{i=1}^{k} p_i S_i^2 + \bar{S}_w^2$$

Thus, we see that the larger the difference between the strata *S.D.*, the larger is the gain in precision of Neyman allocation over proportional allocation.

Further by adding (5) and (6), we see that

$$V(\bar{y}_n)_R - V(\bar{y}_w)_N \cong \frac{N-n}{Nn}\sum_{i=1}^{k} p_i\left[\bar{y}_{N_i} - \bar{y}_N\right]^2 + \frac{1}{N}\sum_{i=1}^{k} p_i\left(S_i - \bar{S}_w\right)^2 \quad ...(7)$$

The equation (7) shows that the gain in precision of stratified sampling with Neyman allocation over simple random sampling arises from two factors *i.e.*, (*i*) difference between strata means and (*ii*) different between strata S.D.

III. Arbitrary Allocation

When a sample in allocated arbitrarily among the strata, the variance of the estimated mean is given by

$$V(\bar{y}_w)_S = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$

Whereas when it is selected as a simple random sample, the variance is given by

NOTES

$$V(\bar{y}_n)_R = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

Thus,

$$V(\bar{y}_n)_R - V(\bar{y}_w)_S = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 - \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$

Substituting S^2 from (4), me get

$$V(\bar{y}_n)_R - V(\bar{y}_w)_S = \left(\frac{1}{n} - \frac{1}{N}\right) \left[\sum_{i=1}^{k} p_i S_i^2 + \sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2\right] -$$

$$\sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) P_i^2 S_i^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_i p_i S_i^2 - \sum_i \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$

$$+ \left(\frac{1}{n} - \frac{1}{N}\right) \sum_i P_i (\bar{y}_{N_i} - \bar{y}_N)^2$$

$$= \sum_i \left[\frac{1}{n} - \frac{1}{N} - \frac{p_i}{n_i} + \frac{p_i}{N_i}\right] p_i S_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2$$

$$= \sum_{i=1}^{k} \left(\frac{1}{n} - \frac{p_i}{n_i}\right) p_i S_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2 \quad \ldots(8)$$

$$\because p_i = \frac{N_i}{N}.$$

The second term on RHS of (8) is always +ve. But the first term may be +ve, –ve or zero depending upon this the values of n_i

It has been shown to be non-negative when the sample is allocated according to Neyman allocation. It can be verified to be zero in case of proportional allocation and also when

$$n_i \propto p_i S_i^2$$

Thus, in the case of arbitrary allocation, the first term may not only become –ve but be larger in magnitude than the IInd term, making a stratified sample less efficient than a simple random sample.

NOTES

3.7 ILLUSTRATIVE EXAMPLES

Example 1. *Consider the strata of s.d.'s of areas for villeges in a city as shown in the following Table.*

Stratum Number	*Size of Village in Bighas*	N_i	$\bar{y}_{n_i}$	S_{w_i}	S_{a_i}
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
1	*0 – 500*	*63*	*112.1*	*56.3*	*129.6*
2	*501 – 1500*	*199*	*276.7*	*116.4*	*267.0*
3	*1501 – 2500*	*53*	*558.1*	*186.0*	*276.1*
4	*> 2500*	*25*	*960.1*	*361.3*	*982.2*

The above given table presents the summary of data for complete census of all the 340 villages. The villages were stratified by size of their agricultural area into four strata as shown in column 2 of Table. The number of villages in the different strata are given in column 3. The population values of the strata means for the area under wheat $(\bar{y}_{N_i})$ *and those of the standard deviations for the area under wheat* (S_{w_i}) *and for the agricultural area* (S_{a_i}) *are given in the subsequent columns.*

Calculate the sampling variance of the estimated area under wheat for a sample of 34 villages, if the villages are selected by

(I) simple random sampling,

(II) stratified random sampling and allocated in proportion to the sizes of the strata (N_i),

(III) the products $N_i S_{w_i}$, *and*

(IV) the products $N_i S_{a_i}$.

Solution.

Case I : For Simple Random Sampling

We know

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k} N_i S_{w_i}^2 - \sum_{i=1}^{k} S_{w_i}^2 + \sum_{i=1}^{k} N_i \bar{y}_{N_i}^2 - \frac{\left(\sum_{i=1}^{k} N_i \bar{y}_{N_i}\right)^2}{N}\right] \quad ...(1)$$

NOTES

The relevant calculations are shown in the following table :

Table : Calculation of the sampling variance

Stratum Number	N_i	S_{w_i}	$S_{w_i}^2$	$N_i S_{w_i}^2$	$\bar{y}_{N_i}$	$N_i \bar{y}_{N_i}$	$N_i \bar{y}_{N_i}^2$	$N_i S_{w_i}$	S_{a_i}	$N_i S_{a_i}$	$\frac{N_i S_{w_i}^2}{S_{a_i}}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	63	56.3	3,170	200,000	112.1	7,060	791,000	3,550	129.6	8,160	1,540
2	199	116.4	13,550	2,696,000	276.7	55,060	15,235,000	23,160	267.0	53,130	10,100
3	53	186.0	34,600	1,834,000	558.1	29,580	16,509,000	9,860	276.1	14,630	6,640
4	25	361.3	130,540	3,264,000	960.1	24,000	23,042,000	9,030	982.2	24,560	3,320
Total	340		181,860	7,994,000		115,700	55,577,000	45,600		100,480	21,600

From (i), $S^2 = \frac{1}{339}$ [7,994,000 – 182,000 + 55,577,000 – 39,372,000]

$= 70{,}850$

$\therefore \quad V(\bar{y}_n)_R = \frac{N-n}{N} \cdot \frac{1}{n} S^2 = \frac{306}{340 \times 34} \times 70{,}850 = 1{,}875$

Case II : For Stratified Sampling

(i) Proportional Allocation when n_i is proportional to N_i. Then,

$$V(\bar{y}_w)_p = \frac{N-n}{N^2 n} \sum_{i=1}^{k} N_i S_{wi}^2 = \frac{306 \times 7{,}994{,}000}{(340)^2 \times 34} = 622$$

Case III : For Neyman Allocation ($N_i \propto N_i S_{wi}$)

The allocation of the sample to the different strata will be in proportion to $N_i S_{w_i}$ shown in col. 9 of Table I. On substituting in (22), we get

$$V(\bar{y}_w)_N = \frac{1}{N^2 n} \left(\sum_{i=1}^{k} N_i S_{w_i} \right)^2 - \frac{1}{N^2} \sum_{i=1}^{k} N_i S_{wi}^2$$

$$= \frac{1}{(340)^2} \left[\frac{(45{,}600)^2}{34} - 7{,}994{,}000 \right] = 460$$

Case IV : For Allocation Proportional to $N_i S_{a_i}$

Using, $n_i = \frac{n N_i S_{a_i}}{\left(\sum_{j=1}^{k} N_j S_{a_j} \right)}$, we have

$$V(\bar{y}_w)_s = \frac{1}{N^2 n} \left(\sum_{i=1}^{k} N_i S_{a_i} \right) \left(\sum_{i=1}^{k} N_i \frac{S_{wi}^2}{S_{a_i}} \right) - \frac{1}{N^2} \sum_{i=1}^{k} N_i S_{wi}^2$$

$$= \frac{1}{(340)^2} \left\{ \frac{1}{34} \times (100{,}480)(21{,}600) - 7{,}994{,}000 \right\} = 483$$

NOTES

Example 2. *A yield survey on rice was carried out in Kegalle District (Ceylon) in Maha 1951–52. Twenty-eight villages were selected and distributed in the various strata approximately in proportion to the acreage under rice. Three plots of 1/80 acre each were harvested in each village. The values of the means and the mean squares of the village means for the different strata are given in following Table. Obtain the estimate of the district mean yield by combining the strata means in proportion to the number of villages in the strata. Calculate its variance and hence estimate the efficiency of stratification as compared to simple random sampling, treating the village means as the true means of the respective villages.*

Crop-Cutting Experiments on Rice, Kegalle District (Ceylon) Maha 1951–52

(Means and Mean Squares of Village Mean Yields per Plot)

Stratum Number	N_i	n_i	$\bar{y}_{n_i}$ *(Oz. / Plot)*	S_i^2 *(Oz. / Plot)²*
1	189	5	369	4,330.9
2	242	7	301	14,812.4
3	146	3	368	17,309.0
4	178	3	171	1,658.5
5	287	10	305	3,452.7

Sol. The relevant calculations are given in Table II. From column V we have

$$\text{Est. } \bar{y}_N = \bar{y}_w = 301.6$$

To obtain the variance of $\bar{y}_w$,

$$\text{Est. } V(\bar{y}_w)_s = \sum_{i=1}^{k} \frac{p_i^2 s_i^2}{n_i} - \sum_{i=1}^{k} \frac{p_i^2 s_i^2}{N_i}$$

$$= 298.23 - 7.57 = 290.66$$

$N/(N-1) \cong 1$, we obtain

$$\text{Est. } V(\bar{y}_n)_R \cong \frac{N-n}{Nn}\left\{\sum_{i=1}^{k} p_i s_i^2 + \sum_{i=1}^{k} p_i \bar{y}_{ni}^2 - \left(\sum_{i=1}^{k} p_i \bar{y}_{n_i}\right)^2 - \sum_{i=1}^{k}\left(\frac{p_i s_i^2}{n_i} - \frac{p_i^2 s_i^2}{n_i}\right)\right\}$$

$$= \left\{\frac{1}{28} - \frac{1}{1,042}\right\}\{7,885 + 95,300 - 90,960 - 1,646 + 298\}$$

$$= 379$$

Hence,

Efficiency of stratification $\dfrac{V(\bar{y}_n)_R}{V(\bar{y}_w)_S} = \dfrac{379}{291} = 1.30$ or 130 percent.

See Table II on Next page.

Table II : Crop-Cutting Experiments on Rice, Kegalle District (Ceylon) Maha 1951–52

(Calculation of the District Mean Yield and its Variance)

Stratum Number	N_i *(1)*	n_i *(2)*	$\bar{y}_{n_i}$ *(3)*	p_i *(4)*	$p_i\bar{y}_{n_i}$ *(5) = (3) × (4)*	$p_i\bar{y}_{ni}^{\,2}$ *(6) = (3) × (5)*	s_i^2 *(7)*	$p_i s_i^2$ *(8) = (4) × (7)*	$p_i^2 s_i^2$ *(9) = (4) × (8)*	$\frac{p_i s_i^2}{n_i}$ *(10) = (8) ÷ (2)*	$\frac{p_i^2 s_i^2}{n_i}$ *(11) = (9) ÷ (2)*	$\frac{p_i^2 s_i^2}{N_i}$ *(12) = (9) ÷ (1)*
1	189	5	369	.181382	66.9	24,700	4,330.9	785.55	142.48	157.1	28.50	0.754
2	242	7	301	.232246	69.9	21,000	14,812.4	3,440.12	798.95	491.4	114.14	3.301
3	146	3	368	.140115	51.6	19,000	17,309.0	2,425.25	339.81	808.4	113.27	2.327
4	178	3	17.1	.170825	29.2	5,000	1,658.5	283.31	48.40	94.4	16.13	0.272
5	287	10	305	.275432	84.0	25,600	3,452.7	950.98	261.93	95.1	26.19	0.913
	1,042	28			301.6	95,300		7,885.21		1,646.4	298.23	7.567

NOTES

Example 3. *A sample survey on fertilizer practices was carried out in Raipur district of Madhya Pradesh. The five tehsils in the district were taken as strata. A sample of 100 villages was selected with replacement and with probability proportional to the cultivated area in the villages. The following table gives basic data in regard to total number of villages (N_i), number of selected villages (n_i), total cultivated area (c_i), estimated area under rice per village ($\bar{z}_{n_i}$) and the estimated mean square (s_{iz}^2) for different strata. Given an unbiased estimate of the average area under rice per village in the district and its estimated variance. Also calculate the percentage gain in efficiency due to stratification as compared to the case if the strata were ignored and the villages were selected from the whole population with replacement and with probability proportional to cultivated area in the village.*

Means and Mean Squares of Area Under Rice Per Village

Stratum number	*Total No. of villages* N_i	*No. of villages in the Sample* n_i	*Total cultivated Area* c_i	$\bar{z}_{n_i} = \frac{1}{n_i}\sum_{j}^{n_i}\frac{y_{ij}}{N_i P_{ij}}$	$s_{iz^2} = \frac{\sum_{j}^{n_i}(zij - \bar{z}_{ni})^2}{n_i - 1}$
1	*524*	*21*	*477,807*	*2,757.3*	*405,839.7*
2	*553*	*16*	*317,041*	*3,575.0*	*74,545.0*
3	*657*	*12*	*252,824*	*2,641.0*	*137,865.5*
4	*892*	*26*	*568,565*	*1,923.7*	*113,578.5*
5	*1,176*	*25*	*577,985*	*1,422.8*	*75,538.7*
Total	*3,802*	*100*	*2,194,222*		

Solution. An unbiased estimate of the average area under rice is given by

$$\bar{z}_w = \sum_{i=1}^{k} p_i \bar{z}_{n_i} = 2{,}247.8$$

and its estimated variance is given by

$$\text{Est.}\left[V(\bar{z}_w)\right] = \sum_{i=1}^{k} p_i^2 \frac{s_{iz}^2}{n_i} = 1{,}338.3$$

If a sample of size n is selected from the whole population with replacement and with selection probabilities Pi (i = 1, 2, ..., N) proportional to cultivated area in the village, an unbiased estimate of the reduction in variance of the estimated average area under rice per village, due to stratification, is obtained as 8,058.6.

We have now unbiased estimates of the variance of the estimated population mean from a stratified sample as well of the reduction in variance due to

stratification. Thus an estimate of the percentage gain in efficiency due to stratification may be obtained as the ratio of the two estimates and is given by

Gain in efficiency due to stratification

$$= \frac{\text{Est.}\left[V(\bar{z}_n) - V(\bar{z}_w)\right]}{\text{Est.}\left[V(\bar{z}_w)\right]} \cdot 100 = \frac{8{,}058.6}{1{,}338.3} \cdot 100 = 601$$

Example 4. *2000 cultivators' holdings in Jharkhand (India) were stratified according to their sizes. The number of holdings (N_i), mean area under wheat per holdings $(\bar{Y}_i)$ and standard deviation (s.d.) of area under wheat per holding (S_i) are given in the following Table for each stratum :*

Stratum number	*Number of holdings* (N_i)	*Mean area under wheat per-holding* $(\bar{Y}_i)$	*s.d. of area under wheat per-holding* (S_i)
1	*394*	*5.4*	*8.3*
2	*461*	*16.3*	*13.3*
3	*381*	*24.3*	*15.1*
4	*334*	*34.5*	*19.8*
5	*169*	*42.1*	*24.5*
6	*113*	*50.1*	*26.0*
7	*148*	*63.8*	*35.2*

For a smaple of 200 farms, compute the sample size in each stratum under proportional and optimum allocations. Calculate the sampling variance of the estimated area under wheat from the sample :

(i) If the farms are selected under proportional allocation by with and without replacement methods,

(ii) If the farms are selected under Neyman's allocation by with and without replacement methods.

Also compute the gain in efficiency from these procedures as compared to simple random sampling.

NOTES

Solution. Consider the following Table :

Table : Calculations of allocations, means and sampling variances

Stratum number	N_i	$\bar{Y}_i$	S_i	W_i	nW_i	W_iS_i	$\frac{nW_iS_i}{\Sigma W_iS_i}$	$W_i\bar{Y}_i$	$W_i\bar{Y}_i^2$	$W_iS_i^2$
(I)	*(II)*	*(III)*	*(IV)*	*(V)*	*(VI)*	*(VII)*	*(VIII)*	*(IX)*	*(X)*	*(XI)*
1	394	5.4	8.3	0.1970	40	1.64	19	1.06	5.72	13.612
2	461	16.3	13.3	0.2305	46	3.07	36	3.76	61.29	40.831
3	381	24.3	15.1	0.1905	38	2.88	34	4.63	112.51	43.488
4	334	34.5	19.8	0.1670	33	3.31	39	5.76	198.72	65.538
5	169	42.1	24.5	0.0845	17	2.07	24	3.56	149.88	50.715
6	113	50.1	26.0	0.0565	11	1.47	17	2.83	141.78	38.220
7	148	63.8	35.2	0.0740	15	2.61	31	4.72	301.14	91.872
Total	2000			1.0000	200	17.05	200	26.32	971.04	344.276

For Proportional Allocation, n_i is proportional to N_i or $n_i = nW_i$

The numbers of holdings to be selected from strata are given in column (VI or III) of the table and are 40, 46, 38, 33, 17, 11 and 15, respectively.

$V\left(\bar{y}_{\text{prop}}\right)$ without replacement = $[(N - n)/nN] \sum W_iS_i^2$

= 1.5492 'Prop' stands for proportional allocations

If finite population correction is ignored, we have

$V\left(\bar{y}_{\text{prop}}\right)$ with replacement = $1/n \sum W_iS_i^2$ = 1.7214

For Optimum Allocation, we have

$$n_i \propto N_iS_i \quad \text{or} \quad n_i = n\frac{W_iS_i}{\sum W_iS_i}$$

The numbers of holdings to be selected from strata are given in column (VIII) of the table and are 19, 36, 34, 39, 24, 17 and 31, respectively.

$$V\left(\bar{y}_{\text{opt}}\right) Wor = \left(\sum W_iS_i\right)^{2/n} - \sum W_iS_i^2/N = 1.2813$$

'*Wor*' stands for without replacement

'*Wr*' stands for with replacement

If finite population correction is ignored, we get

$$V\left(\bar{y}_{\text{opt}}\right) wr = \left(\sum W_iS_i\right)^{2/n} = 1.4535$$

NOTES

For Simple Random Sampling, The variance of the estimate of mean is given by

$$V(\bar{y}_{SR})\ Wor = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

$$= V_{opt} + \frac{(N-n)}{n(N-1)}\left[\sum W_i \bar{Y}_i^2 - \left(\sum W_i \bar{Y}_i\right)^2 + \sum W_i S_i^2 - \left(\sum W_i S_i\right)^2\right] = 3.0420$$

If finite population correction is ignored, we get

$$V(\bar{y}_{SR})\, wr = V_{\text{prop}} + \sum W_i \left(\bar{Y}_i - \bar{Y}\right)^{2/n} = 3.1129$$

'*SR*' stands for stratified random sampling

(*i*) The relative precision of proportional allocation is given by

(*a*) without replacement method $V_{SR}/V_{\text{prop}} = 1.9636$

(*b*) with replacement methods $V_{SR}/V_{\text{prop}} = 1.8084$

(*ii*) The relative precision of optimum allocation is given by

(*a*) without replacement method $V_{SR}/V_{\text{opt}} = 2.3742$

(*b*) with replacement method $V_{SR}/V_{\text{opt}} = 2.1416$

Example 5. *The number of pepper standards for selected villages in each of the three strata of orissa zone is shown in the following table.*

Stratum	*Total number of villages in the stratum*	*Number of villages selected from the stratum*	*Number of pepper standards in each of the selected villages*
1	*441*	*11*	*41, 116, 19, 15, 144, 159, 212, 57, 28, 119, 76.*
2	*405*	*12*	*39, 70, 38, 37, 161, 38, 27, 119, 36, 128, 30, 208.*
3	*103*	*7*	*252, 385, 192, 296, 115, 159, 120.*

Estimate the total number of pepper standards along with its standard error in orissa zone.

Also, estimate the gain in precision due to stratification.

Solution.

An estimate of the total number of pepper standards is given by

Consider the following Table I (shown in the next page)

NOTES

Table I. Calculations of estimates of means and variances

Stratum	N_i	n_i	$\sum_j y_{ij}$	$\bar{y}_i$	$N_i \bar{y}_i$	$\sum_j y_{i,j}^2$	$\sum_j y_{i,j}^2 / n_i$	s_i^2	$\left(\frac{1}{n_i} - \frac{1}{N_i}\right)$	$N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_i^2$	$N_i s_i^2$	$\frac{N_i s_i^2}{n_i}$	$N_i \bar{y}_i^2$	$\frac{N_i^2 s_i^2}{n_i}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1	441	11	986	89.63	39,529.81	130,654	88,381.45	4227.2	0.0886	72,839,892.71	1,864,219.28	169,474.48	3,543,275.94	74,738,245.68
2	405	12	743	91.92	25,481.22	70,469	46,004.08	2224.08	0.0809	29,512,745.97	900,753.74	75,062.81	1,577,720.88	30,400,438.62
3	103	7	1520	217.71	22,365.71	389,885	330,057.14	9971.44	0.1331	14,080,301.61	1,027,062.04	146,723.15	4,856,556.93	15,112,484.24
Total	949	30								116,432,940.29	3,792,035.06	391,260.44	9,977,523.75	120,251,168.54

NOTES

$$\hat{Y}_{st} = \sum_{i}^{k} N_i \bar{y}_i = 87376.54$$ 'St' stands for stratified

An estimate of variance of $\bar{Y}_{st}$ is

$$v\left(\hat{Y}_{st}\right) = \sum_{i}^{k} N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s_i^2 = 116,432,940.29$$

or standard error of $\bar{Y}_{st} = \sqrt{116,432,940.29} = 10,790.41$

Estimate of the variance of the total number of pepper standards, when simple random sampling is assumed, is given by

$$v\left(\hat{Y}_{SR}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) N \left\{\sum_{i}^{k} N_i s_i^2 - \sum_{i}^{k} \frac{N_i s_i^2}{n_i} + \sum_{i}^{k} N_i \bar{y}_i^2 - \left(\sum_{i}^{k} N_i \bar{y}_i\right)^2\right\}$$

$$= 167,269,559.37$$

Thus, the percentage gain in precision due to stratification

$$= \frac{v\left(\hat{Y}_{SR}\right) - v\left(\hat{Y}_{st}\right)}{v\left(\hat{Y}_{st}\right)} \times 100$$ 'SR' stands for stratified random sampling.

$$= \frac{167,269,559.37 - 116,432,940.29}{116,432,940.29} \times 100 = 43.65$$

3.8 SUMMARY

- The classes into which the population is divided are called **strata** and the process is termed the procedure of **stratified random sampling.**
- The allocation of the sample to the different strata made in accordance with thin principle is called the principle of **optimum allocation.**
- The variance of estimated population mean is given by

$$V(\bar{y}_w) = \sum_{i=1}^{k} \left[\frac{1}{n_i} - \frac{1}{N_i}\right] p_i^2 S_i^2$$

- The variance of mean under simple random sampling is given by

$$V(\bar{y}_n)_R = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

- The more the strata differ in their population means, the larger is the gain in precisson of stratified sampling.

NOTES

3.9 GLOSSARY

- **Strata.** The classes into which the population is divided are called strata.

3.10 REVIEW QUESTIONS

1. A population is divided into 2 strata of sizes N_1 and N_2 units. Let ϕ denote $[n_1/n_2]/[n_1'/n_2']$, where n_1, n_2 is a general allocation of the total sample size n, and n_1', n_2' is the optimum allocation for the estimation of mean in stratified random sampling.

   (*i*) Show that the relative precision of a general allocation to the optimum allocation is given by

   $$RP = \phi (N_1U + N_2)^2/(\phi N_1U + N_2)\,(N_1U + \phi N_2)$$

   where $U = S_1/S_2$

   (*ii*) Show that the relative precision is never less than $4\phi\,(1 + \phi)^{-2}$.

2. Prove that

   $$V_{SR} = V_{\text{prop}} + \frac{(N-n)}{nN\,(N-1)}\left\{\sum_i^k N_i\left(\bar{Y}_i - \bar{Y}\right)^2 - \frac{1}{N}\sum_i^k (N - N_i)\,S_i^2\right\}$$

3. The variate x_i which is non-negative follows the exponential distribution $e^{-x_i}\,d_{x_i}$. The population is divided into 2 strata at a specified point x_0, and a stratified random sample of size n is taken with proportional allocation. Derive $V(\bar{x}_{st})$ as a function of x_0. Also find the optimum value of x_0 which will minimize the variance.

4. The variate y has rectangular distribution in the interval $(a, a + d)$. The interval is divided into k equal sub-intervals which form k strata of equal size. From each stratum, a simple random sample of size n/k units is drawn. Let V_1 and V_2 be the variances for stratified and unstratified samples of size n, respectively, prove that $V_1/V_2 = k^{-2}$.

5. For the normal population $N(0, 1)$, $f(y)$ denotes the ordinate at y and w the weight between y_1 and y_2. If the distribution is truncated at y_1 and y_2, show that the mean M and variances S^2 of the truncated distribution are given by

   $$M = \{f(y_1) - f(y_2)\}/w$$

and $$S^2 - 1 = \frac{y_1 f(y_1) - y_2 f(y_2)}{w} - \frac{[f(y_1) - f(y_2)]^2}{w^2}$$

A population is divided into 2 strata of sizes N_1 and N_2 and sd's S_1 and S_2. The cost of the survey is fixed and given by $C = c_1 n_1 + c_2 n_2$. Assuming that S_1 and S_2 are nearly equal and fpc can be ignered, show that

$$V_{prop}/V_{opt} = N(N_1 c_i + N_2 c_2)\Big/\left(N_1\sqrt{c_1} + N_2\sqrt{c_2}\right)^2.$$

If $N_1 = N_2$, compute the relative increases in precision by using proportional allocation when $c_2/c_1 = 1$ and 2.

6. An investigator desires to take a stratified random sample with the following assumptions :

Stratum	N_i	S_i	C_i *(in Rs.)*
1	400	10	4
2	600	20	9

(*i*) Estimate the values of n_1/n and n_2/n which minimize the total field cost $C = c_1 n_1 + c_2 n_2$ for a given value of $V(\bar{y}_{st})$.

(*ii*) Estimate the total sample size required, under the scheme of optimum allocation, to make $V(\bar{y}_{st}) = 1$, when fpc is ignored.

(*iii*) Also estimate the cost of the survey.

7. After the sample in Excercise 6 was taken, it was found that the field costs were actually Rs. 2 per unit in stratum 1 and Rs 10 in stratum 2.

(*i*) How much has the field cost changed from the anticipated value ?

(*ii*) If the exact field cost is known in advance, how could the value $V(\bar{y}_{st}) = 1$ be obtained for the original estimated cost given in Exercise 6.

8. For estimating the average catch of fish on a certain part of the Indian coast, the coastal strip was divided into two geographical strata of equal number of fish-landing centres. From among the large number of centres in each stratum, a simple random sample of 5 centres was selected for observation and from each such centre, out of the large number of operating units, 3 units were selected by simple random sampling for recording the weight of catch. The data obtained are given below :

NOTES

	S.No. of centres	Catches of units selected (kg)		
		1	2	3
Stratum I	1	610	754	688
	2	297	411	515
	3	1187	92	487
	4	1297	533	1130
	5	860	357	656
Stratum II	1	1085	956	980
	2	817	736	926
	3	920	616	109
	4	511	328	412
	5	906	990	736

Obtain an estimate of the average catch per unit, for the coast, with its standard error. For a given number of centres to be selected, what is the optimum break up between the two strata when the number of units selected at a centre are 3 and 4. Calculate the number of centres and the optimum break up necessary to estimate the average catch with 5 percent standard error.

9. A sample survey for estimating the number of orchards of apple was conducted in Mahasu district of Himachal Pradesh (India) during a given year. Four strata *A*, *B*, *C* and *D* of villages, according to the acreage of temperate fruit trees as obtained from the revenue records, were formed. The sizes of strata (in acres) were 0–3, 3–6, 6–15 and 15 and above, respectively. A simple random sample of villages in each stratum was selected and the number of apple orchards was noted in the selected villages. The numbers of apple orchards for various strata are given below :

Stratum	*Total number*	*Number of villages selected*	*Number of orchards in the selected villages*
A	275	15	2, 5, 1, 9, 6, 7, 0, 4, 7, 0, 5, 0, 0, 3, 0.
B	146	10	21, 11, 7, 5, 6, 19, 5, 24, 30, 24.
C	93	12	3, 10, 4, 11, 38, 11, 4, 46, 4, 18, 1, 19
D	62	11	30, 42, 20, 38, 29, 22, 31, 28, 66, 41, 15.

NOTES

Estimate the number of orchards in the district. Calculate whether there is any gain due to stratification, over simple random sampling. Using the value of s_i^2, the mean square within the i^{th} stratum as the variance, give an optimum allocation of 48 villages.

10. In a pilot sample survey conducted in the district of Wardha (India), for estimation of total cattle population in the district, a two way stratification of the district by tehsils and sizes of the villages, as judged by the number of households, was adopted. A total sample of 125 villages was distributed equally among all 9 strata. Villages were sampled with equal probability and without replacement within each stratum. All households in a village were enumerated for total cattles. The total cattle and mean square between villages within each stratum are given in the table given below.

Calculate the sampling variance of the estimated total cattle population in the district, for the sample of 60 villages, if the villages were selected by the method of

(*i*) simple random sampling without stratification, and

(*ii*) simple random sampling within each stratum and allocated in proportion to the product $N_i S_i$

Tehsils	*Size according to number of households*	*Stratum*	*Total no. of villages* N_i	*No. of villages sampled* (n_i)	*Total cattle in sampled villages*	*Mean square between villages within strata* s_i^2
Arvi	0–50	1	127	13	1486	$(78.9279)^2$
	51–125	2	109	14	5085	$(126.2198)^2$
	126 and above	3	81	14	10141	$(339.9322)^2$
Wardha	0–50	4	86	14	1224	$(73.0564)^2$
	51–125	5	121	14	4282	$(55.2853)^2$
	126 and above	6	115	14	11334	$(432.2364)^2$
Huigaughat	0–50	7	123	14	2909	$(128.2665)^2$
	51–125	8	103	14	5753	$(101.3872)^2$
	126 and above	9	63	14	10002	$(298.5456)^2$
			928	125	52228	

11. For a socio-economic survey, all the villages in a region, including the uninhabited ones, were grouped into four strata on the basis of their altitude above sea level and population density. From each stratum, 10 villages were selected with srs wr. The data on the number of households in each of the sample villages are given below :

NOTES

Stratum S. No.	Total no. of villages	Total number of households in sample villages									
		1	2	3	4	5	6	7	8	9	10
1	1411	43	84	98	0	10	44	0	124	13	0
2	4705	50	147	62	87	84	158	170	104	56	160
3	2558	228	262	110	232	139	178	334	0	63	220
4	14,997	17	34	25	34	36	0	25	7	15	31

(*i*) Obtain an estimate of the total number of households and its standard error.

(*ii*) Estimate the gain due to use of stratification as compared to unstratified srs wr.

(*iii*) Compare the efficiency of the present allocation with that of the optimum allocation, keeping the total sample size fixed.

12. Using the data given below and considering the size classes as strata, compare the efficiencies of the following alternative allocations of a sample of 3000 factories for estimating the total output. The sample is to be selected with sr wor within each stratum :

(*i*) Proportional allocation

(*ii*) Allocation proportional to total output

(*iii*) Optimum allocation

S.No.	Size class no. of workers	No. of factories	Output per factory (in '000 Rs)	Standard deviation (in '000 Rs)
1	1–49	18260	100	80
2	50–99	4315	250	200
3	100–249	2233	500	600
4	250–999	1057	1760	1900
5	1000 and above	567	2250	2500

13. A survey is to be conducted for estimating the total number of literates in a town inhabited by three communities, some particulars of which are given bleow on the basis of the results of a pilot survey :

Community	Total number of persons	Percentage of literates
1	60000	40
2	10000	80
3	30000	60

(*i*) Treating the community as strata and assuming srs wr in each stratum, allocate a total sample size of 2000 persons to the strata in an optimum manner for estimating the overall proportion of literates in the town.

(*ii*) Estimate the efficiency of stratification as compared to unstratified sampling.

3.11 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

NOTES

CHAPTER 4 SYSTEMATIC SAMPLING

OBJECTIVES

After going through this chapter, we should be able to :

- know systematic sampling.
- know stratified random sampling.
- know periodic variation.

STRUCTURE

4.1 INTRODUCTION

We now consider a method of sampling in which only the first unit is selected with the help of random numbers, the rest being selected automatically according to a predetermined pattern. The method is known as systematic sampling.

The pattern usually followed in selecting a systematic sample is simple pattern involving regular spacing of units. Thus suppose a population consists of N

NOTES

units, serially numbered form 1 to N. Suppose further that N is expressible as a product of two integers k and n, so that $N = kn$. Draw a random number less than or equal to k, say i, and select the unit with the corresponding serial number and every k^{th} unit in the population thereafter. Clearly, the sample will contain the n units i, $i + k$, $i+2k$, ..., $i + (n - 1)k$. Such a sample is known as systematic sample.

Systematic sampling resembles cluster sampling. A systematic sample being equivalent to a sample of one cluster selected out of k clusters of n units each as shown in the following dia consisting of columns and n rows each :

Cluster No.	*1*	*2*	...	*i*	...	*k*
1	1	2	...	i	...	k
2	$1+k$	$2+k$	...	$i+k$	...	$2k$
3	$1+2k$	$2+2k$	...	i+2k	...	$3k$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
J	$1+(j-1)k$	$2+(j-1)k$	...	$i+(j-1)k$	..	Jk
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	$1+(n-1)k$	$2+(n-1)k$	...	$i+(n-1)k$	...	nk

4.2 THE SAMPLE MEAN AND ITS VARIANCE

Let y_{ij} denote the observation on the unit bearing the serial number $i + (j - 1)\, k$ in the population $i = 1, 2, ..., k$; $J = 1, 2, ..., n$ and suppose that the random number drawn less than or equal to k is i. The sample selected then consists of all the units with serial numbers listed in the i^{th} column in above table. Let

$$\bar{y}_{i*} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \text{ be the sample mean} \quad ...(1)$$

$$\bar{y}_{**} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} \text{ be the population mean} \quad ...(2)$$

Since the problem of selecting the i^{th} column as the systematic sample is $\frac{1}{k}$, it follows that

$$E(\bar{y}_{i*}) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_{i*} = \bar{y}_{**} \quad ...(3)$$

$\Rightarrow$ Systematic sample mean provides an unbiased estimate of popular mean. Now, the variance of sample mean is given by

$$V[\bar{y}_{i*}]_{s_y} = \frac{1}{k}\sum_{i=1}^{k}(y_{i*} - \bar{y}_{**})^2 = \frac{(k-1)}{k}S_c^2 \quad ...(4)$$

NOTES

where S_c^2 denotes the Mean Square between the column means in the population.

Now, since we have assumed that N is an exact multiple of n *i.e.*, $N = nk$ where k is an integer. We shall briefly consider the case when $N = nk + r$ where r is less than k. It is obvious that in this case the sample size will vary, being either n or $(n + 1)$ depending upon the initial number selected.

Let y_{i*} denote the sample total and $\bar{y}_{i*}$ the sample mean.

Thus,
$$\bar{y}_{i*} = \frac{1}{n+1} y_{i*} = \frac{1}{n+1} \sum_{j=1}^{n+1} y_{ij}, \text{ if } i \le r \quad \ldots(5)$$

$$= \frac{1}{n} y_{i*} = \frac{1}{n} \sum_{j-1}^{n} y_{ij}, \text{ if } i > r \quad \ldots(6)$$

Since, $E(\bar{y}_{i*}) = \frac{1}{k} \sum_{i=1}^{k} \bar{y}_{ix} \neq \bar{y}_{**}$, it follows that this estimate $\bar{y}_{i*}$ is biased

As unbiased estimate of the population mean can be easily formed.

Consider
$$\bar{y}'_{i*} = \frac{k}{N} y_{i*} \quad \ldots(7)$$

Then,
$$E(\bar{y}'_{i*}) = \frac{1}{k} \sum_{i=1}^{k} \bar{y}'_{iy} = \frac{1}{k} \cdot \frac{k}{N} \sum_{i=1}^{k} y_{i*} = \frac{1}{N} \sum_{i=1}^{k} y_{i*}$$

$$= \frac{1}{nk} \sum_{i=1}^{k} y_{i*} = \frac{1}{k} \sum_{i=1}^{k} y_{i*} = \bar{y}_{**} \quad \ldots(8)$$

Now,
$$V(\bar{y}'_{i*}) = \frac{k^2}{N^2} V(y_{i*}) = \frac{k^2}{N^2} \frac{k-1}{k} S_c'^2 = \frac{(k-1)k}{N^2} S_c'^2 \quad \ldots(9)$$

where,
$$S_c'^2 = \frac{1}{k-1} \sum_{i=1}^{k} (y_{i*} - \bar{y}_{k*})^2$$

and
$$\bar{y}_{k*} = \frac{1}{k} \sum_{i=1}^{k} y_{i*} = \frac{N}{k} \bar{y}_{**}$$

4.3 THE CONCEPT OF CIRCULAR SYSTEMATIC SAMPLING

In systematic sampling, we observe that the sample size varies, being n or $n + 1$ depending uypon the initial number selected. The difficulty arising may be avoided by another method which is explained below :

NOTES

Select a number at random from 1 through N, start at the unit corresponding to this number and thereafter select cyclically every k^{th} unit (k is the integer nearest to the inverse of the sampling fraction) until n units have been chosen for the sample. By cyclic selection, we mean assigning the number $N + 1$ to the first unit on the list, $N + 2$ to the second unit and so on, in order to continue the selection procedure when the N-th unit has been reached. This procedure is known as **circular systematic sampling.** Clearly, this method of selection gives rise to N possible samples, each with selection probability $\frac{1}{N}$. Denote by $\bar{y}_{i*}$ the arithmetic mean of the i-th sample. Since each unit occurs in n samples, being selected once each as the first element, second element, third element etc., it can be seen that $\bar{y}_{i*}$ is an unbiased estimate of the population mean. Hence, in case of circular systematic sampling, the sample mean is an unbiased estimate of the population mean. The variance of the estimate is given by

$$V(\bar{y}_{i*})_{c.ry} = E[\bar{y}_{i*} - \bar{y}_{**}]^2 = \frac{1}{N}\sum_{i=1}^{N}[\bar{y}_{i*} - \bar{y}_{**}]^2$$

where $c.ry$ denotes circular systematic sampling.

4.4 COMPARISON OF SYSTEMATIC SAMPLING WITH RANDOM SAMPLING

We know that the variation of mean of a random sample of n units closer from a population of size N is known to be given by

$$V(\bar{y}_n)_R = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \qquad ...(1)$$

where S^2 is the Mean Square between units in the population.

Now, we first express the variance of systematic sample so that it can be compared with (1).

We have $V(\bar{y}_{i*})_{sy} = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_{i*} - \bar{y}_{**})^2$

$$= \frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{n}\sum_{j=1}^{n}y_{ij} - \bar{y}_{i*}\right]^2 = \frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{iJ} - \bar{y}_{**})\right]^2$$

$$= \frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \bar{y}_{**})^2 + \sum_{j\neq j'=1}^{n}(y_{ij} - \bar{y}_{**})(y_{ij'} + \bar{y}_{**})\right]$$

$$V(\bar{y}_{i*})_{sy} = \frac{1}{kn^2}\left[(nk-1)S^2 + \sum_{i=1}^{k}\sum_{j\neq j'=1}^{n}(y_{ij'}-\bar{y}_{**})(y_{iJ'}-\bar{y}_{**})\right] \quad ...(2)$$

NOTES

Now, intra class correlation between units of a column is given by

$$\rho = \frac{E\left[(y_{ij}-\bar{y}_{**})(y_{ij'}-\bar{y}_{**})\right]}{E(y_{ij}-\bar{y}_{**})^2}$$

$$= \frac{\sum_{i=1}^{k}\sum_{j\neq j'=1}^{n}(y_{ij}-\bar{y}_{**})(y_{ij'}-\bar{y}_{**})}{n(n-1)k} \cdot \frac{nk}{(kn-1)S^2}$$

$$\therefore \sum_{i=1}^{k}\sum_{j\neq j'=1}^{n}(y_{ij}-\bar{y}_{**})(y_{ij'}-\bar{y}_{**}) = (n-1)(kn-1)S^2\rho. \quad ...(3)$$

Using (3) in (2), we get

$$V(\bar{y}_{i*})_{sy} = \frac{kn-1}{kn^2}S^2[1+\rho(n-1)] \quad ...(4)$$

which is suitable for comparison with (1).

Thus, $$\frac{V(\bar{y}_{i*})_{sy}}{V(\bar{y}_n)_R} = \frac{(nk-1)[1+\rho(n-1)]}{n(k-1)} \quad (\because N = kn) \quad ...(5)$$

The relative precision depends on the value of ρ.

If $\rho = -\frac{1}{kn-1}$, the two methods give estimates of equal precision.

If $\rho > -\frac{1}{kn-1}$, systematic sampling is less accurate than random sampling.

If $\rho < -\frac{1}{kn-1}$, systematic sampling is superior to random sampling.

The minimum value which ρ can take is $-\frac{1}{n-1}$, when the variance of a systematic sample will be zero and the reduction in variance over random

sampling will therefore be 100%. The maximum value of $\rho = 1$, when the efficiency of systematic relative to random sampling will be $\frac{(k-1)}{(nk-1)}$.

NOTES

In general, it is difficult to know what values ρ will take population distributed in space or time and so no general conclusions can be drawn about the efficiency of systematic relative to random sampling.

Now, we express the variation of the systematic sample in terms of a further break up to the extra class correlation coefficient. From equation (3) ρ is expressed in terms of the sum of $kn(n-1)$ products of y observations;

$2k(n-1)$ of these products relate to y deviations separated by one row;

$2k(n-2)$ of these products relate to y deviations separated by two rows etc.;

we may therefore rewrite equation (3) as

$$(n-1)(kn-1)\rho S^2 = 2\sum_{i=1}^{k}\sum_{\alpha=1}^{n-1}\sum_{j=1}^{n-\alpha}\left(y_{ij}-\bar{y}_{**}\right)\left(y_{i\,j+\alpha}-\bar{y}_{**}\right).$$

We now introduce the non-circular serial correlation coefficient. ρ_α for log $k\alpha$ defined by

$$\rho_\alpha = \frac{E\left[\left(y_{ij}-\bar{y}_{**}\right)\left(y_{i\,j+\alpha}-\bar{y}_{**}\right)\right]}{E\left(y_{ij}-\bar{y}_{**}\right)^2}$$

$$= \frac{1}{k(n-\alpha)}\sum_{i=1}^{k}\sum_{j=1}^{n-\alpha}\left(y_{ij}-\bar{y}_{**}\right)\left(y_{i\,j+\alpha}-\bar{y}_{**}\right)\Big/\frac{kn-1}{kn}S^2$$

$$\therefore \quad \rho = \frac{2}{(n-1)(kn-1)S^2}\sum_{\alpha=1}^{n-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n-\alpha}\left(y_{ij}-\bar{y}_{**}\right)\left(y_{i\,j+\alpha}-\bar{y}_{**}\right)\right]$$

$$= \frac{2}{(n-1)(kn-1)S^2}\sum_{\alpha=1}^{n-1}\frac{(kn-1)S^2}{kn}\rho_\alpha\, k(n-\alpha)$$

$$= \frac{2}{n(n-1)}\sum_{\alpha=1}^{n-1}(n-\alpha)\rho_\alpha$$

Substitute the value of ρ in (4), we get

$$V(\bar{y}_{ix})_{s_y} = \frac{kn-1}{kn}\cdot\frac{S^2}{n}\left[1+\frac{2}{n}\sum_{\alpha=1}^{n-1}(n-\alpha)\rho_\alpha\right]$$

This expression is due to Madows.

4.5 COMPARISON OF SYSTEMATIC SAMPLING WITH STRATIFIED RANDOM SAMPLING

NOTES

We shall consider the population to be divided into n strata corresponding to the n rows and suppose that one unit is randomly drawn from each one of these strata, thus giving a stratified sample of size n. The variation of this sample mean will be

$$V(\bar{y}_w)_s = \frac{1}{n^2}\sum_{j=1}^{n}\left(1-\frac{1}{k}\right)S_j^2 \qquad ...(1)$$

((In statified sampling, $V(\bar{y}_w)_s = \sum_{i=1}^{k} p_i^2 V(\bar{y}_n)$

$$= \sum_{j=1}^{n} p_i^2 \cdot \left(\frac{k-1}{k}\right)S_j^2 = \frac{1}{n^2}\sum_{j=1}^{n}\left(1-\frac{1}{k}\right)S_j^2))$$

where, S_j^2 is the Mean Square of the units in the j^{th} stratum (row) defined by

$$S_j^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(y_{ij}-\bar{y}_{*j}\right)^2 \qquad ...(2)$$

Substituting (2) in (1), we get

$$V(\bar{y}_w)_s = \frac{1}{n}\left(1-\frac{1}{k}\right)\frac{1}{n(k-1)}\sum_{j=1}^{n}\sum_{i=1}^{k}\left(y_{ij}-\bar{y}_{*j}\right)^2$$

$$= \frac{1}{n}\left(1-\frac{1}{k}\right)S_{wr}^2 \qquad ...(3)$$

where, S_{wr}^2 is the pooled Mean Square between units within rows defined by

$$S_{wr}^2 = \frac{1}{n(k-1)}\sum_{j=1}^{n}\sum_{i=1}^{k}\left(y_{ij}-\bar{y}_{*J}\right)^2 \qquad ...(4)$$

To compare this variance with the variance of the mean of a systematic sample, we first write the variation of mean of systematic sampling as

$$V(\bar{y}_{i*})_{s_y} = \frac{1}{k}\left[\sum_{i=1}^{k}(\bar{y}_{i*}-\bar{y}_{**})^2\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{n}\sum_{j=1}^{n}y_{ij}-\frac{1}{n}\sum_{j=1}^{n}\bar{y}_{*j}\right]^2$$

NOTES

$$= \frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{*j}\right)\right]^2$$

$$= \frac{1}{kn^2}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{*j}\right)+\sum_{i=1}^{k}\sum_{j\neq j'=1}^{n}\left(y_{ij}-\bar{y}_{*j}\right)\left(y_{ij'}-\bar{y}_{*j'}\right)\right] \quad ...(5)$$

The second term on RHS of (5) contains $kn(n-1)$ product terms of the y deviation from the respective strata means. We may group these $kn(n-1)$ terms into $(n-1)$ groups in such a manner that the factors in each term the α^{th} group ($\alpha = 1, 2, ..., n-1$) are separated by α rows.

Thus, equation (5) can be written as

$$V\left(\bar{y}_{i*}\right)_{s_y} = \frac{1}{kn^2}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{*j}\right)^2+2\sum_{i=1}^{k}\sum_{\alpha=1}^{n-1}\sum_{j=1}^{n-\alpha}\left(y_{ij}-\bar{y}_{*j}\right)\left(y_{i\,j+\alpha}-\bar{y}_{*j+\alpha}\right)\right] \quad ...(6)$$

Now, we introduce the within stratum non-circular serial correlation coefficient $\rho_{(\alpha)\omega}$ defined by

$$\rho_{(\alpha)\omega} = \frac{E\left(y_{ij}-\bar{y}_{*j}\right)\left(y_{ij+\alpha}-\bar{y}_{*j+\alpha}\right)}{E\left(y_{ij}-\bar{y}_{j}\right)^2}$$

$$= \frac{\frac{1}{k(n-\alpha)}\sum_{j=1}^{n-\alpha}\sum_{i=1}^{k}\left(y_{ij}-\bar{y}_{*j}\right)\left(y_{i\,j+\alpha}-\bar{y}_{*j}\right)}{\frac{1}{kn}\sum_{j=1}^{n}\sum_{i=1}^{k}\left(y_{ij}-\bar{y}_{*j}\right)^2} \quad ...(7)$$

Substituting (7) in (6) and using (4), we get

$$V\left(\bar{y}_{ix}\right)_{s_y} = \frac{k-1}{k}\frac{S_{wr}^2}{n}\left[1+\frac{2}{n}\sum_{\alpha=1}^{n-1}(n-\alpha)\rho_{(\alpha)\omega}\right] \quad ...(8)$$

Using (3), we get

$$V\left(\bar{y}_{i}\right)_{s_y} = V\left(\bar{y}_{w}\right)_{s}\left[1+\frac{2}{n}\sum_{\alpha=1}^{n-1}(n-\alpha)\rho_{(\alpha)\omega}\right] \quad ...(8)$$

Thus, we see that relative efficiency of systematic and stratified sample depends upon the values of $\rho_{(\alpha)\omega}$ in the population and thus no general conclusions can be drawn.

It $\rho_{(\alpha)\omega}$ are (+ve), stratified sampling will be superior to the systematic sampling and if negative, systematic will be superior to stratified. The both will be equally efficient if $\rho_{(\alpha)\omega} = 0$.

NOTES

4.6 COMPARISON OF SYSTEMATIC SAMPLING WITH SIMPLE AND STRATIFIED RANDOM SAMPLE FOR CERTAIN SPECIFIED POPULATIONS

I. Linear Trend. Let us suppose that the values of the successive units of the population increase in accordance with a linear model, so that

$$y_h = \mu + h\theta \qquad ...(1)$$

where, μ and θ are constants and h goes from 1 to N.

Clearly,

$$\bar{y}_{**} = \frac{1}{N}\sum_{h=1}^{N}(\mu+h\theta) = \mu+\frac{1}{N}[1+2+\ ...+N]\theta = \frac{N+1}{2}\theta+\mu \qquad ...(2)$$

$$\sum_{h=1}^{N} y_h^2 = \sum_{h=1}^{N}(\mu+h\theta)^2 = N\mu^2+\frac{N(N+1)(2N+1)}{6}\theta^2+\mu N(N+1)\theta \qquad ...(3)$$

$$\therefore \quad S^2 = \frac{1}{N-1}\left[\sum_{h=1}^{N} y_h^2 - N\bar{y}_{**}^2\right] = \frac{N(N+1)}{12}\theta^2 = \frac{nk(nk+1)}{12}\theta^2 \qquad ...(4)$$

Similarly, since the observations within each row increase by the same amount, we have

$$S_{wr}^2 = \frac{k(k+1)}{12}\theta^2 \qquad ...(5)$$

Also for the same reason, since the column means corresponding to k different systematic samples also increase by the same amount, we have the Mean Square between column means given by

$$S_c^2 = \frac{k(k+1)}{12}\theta^2 \qquad ...(6)$$

Now, variances of systematic, simple random and satisfied sample are

$$V(\bar{y}_{i*})_{s_y} = \frac{k-1}{k}S_c^2$$

$$V(\bar{y}_n)_R = \left(\frac{1}{n}-\frac{1}{N}\right)S^2$$

NOTES

and $$V(\bar{y}_w)_s = \frac{1}{n}\left(1-\frac{1}{k}\right)S_{wr}^2$$

Substituting the values of S^2, S_{wr}^2 and S_c^2 turn (4), (5), and (6), we get

$$V_{s_y} = \frac{k^2-1}{12}\theta^2$$

$$V_R = \frac{(k-1)(nk+1)}{12}\theta^2$$

$$V_s = \frac{k^2-1}{12n}\theta^2$$

Hence, $$V_s : V_{s_y} : V_R = \left(\frac{k+1}{n}\right) : (k+1) : (nk+1)$$

or, $$\theta_2 \cong \frac{1}{n} : 1 : n$$

Thus, we conclude that variance of stratified sample is only $\frac{1}{n}^{\text{th}}$ the variance of a systematic sample and the systematic is also $\frac{1}{n}^{\text{th}}$ variance of random sample. Stratified sampling is thus, seen to be the most efficient of the three methods for removing the effects of a linear trend with systematic sampling following it as the next best method.

II. Periodic Variation. We consider populations in which sampling units with high and low values follow one another according to a regular pattern. Suppose such a population is represented by

$$y_h = \sin\left(\alpha+(h-1)\frac{\pi}{10}\right)$$

where, h varies from 1 to integral multiple of 20.

Clearly, the inccessive sampling units will repeat themselves after every 20th value. A systematic 5% sample from such as a population will consist of sampling units drawn from the same position of each cycle, giving an estimate which is no more accurate than a single value. A 5% random sample on the other hand will contain units from the different parts of the cycles with the result that the means of such samples will vary within a narrower range than the means of different systematic samples, this making random samples more efficient than systematic samples for removing the effect of a periodic trend.

STUDENT ACTIVITY 1

1. Derive the formulae for sample mean and variance in case of systematic sampling.

STUDENT ACTIVITY 2

1. Compare systematic sampling with random sampling.

NOTES

4.7 ILLUSTRATIVE EXAMPLES

Example 1. *Given below are the daily milk yield (in litres) records of the first lactation of a specified cow belonging to various villages of Punjab. The milk yields of the first five days were not recorded, being the colostrum period.*

Day	1	2	3	4	5	6	7	8	9	10
yield	10	11	14	10	14	9	10	8	11	10
Day	11	12	13	14	15	16	17	18	19	20
yield	6	9	8	7	9	10	11	11	13	12
Day	21	22	23	24	25	26	27	28	29	30
yield	12	10	11	11	14	15	12	17	18	16
Day	31	32	33	34	35	36	37	38	39	40
yield	13	14	14	15	16	16	16	13	16	17
Day	41	42	43	44	45	46	47	48	49	50
yield	14	16	15	14	14	15	17	15	16	17
Day	51	52	53	54	55	56	57	58	59	60
yield	25	22	23	19	18	16	22	21	21	23
Day	61	62	63	64	65	66	67	68	69	70
yield	21	19	19	19	19	19	19	19	19	19
Day	71	72	73	74	75	76	77	78	79	80
yield	18	19	21	20	17	16	18	18	18	22
Day	81	82	83	84	85	86	87	88	89	90
yield	22	22	20	20	20	18	20	21	21	20
Day	91	92	93	94	95	96	97	98	99	100
yield	18	21	22	22	20	21	21	21	21	21
Day	101	102	103	104	105	106	107	108	109	110
yield	19	20	21	20	21	20	21	20	21	20
Day	111	112	113	114	115	116	117	118	119	120
yield	19	21	18	21	20	22	21	21	21	16
Day	121	122	123	124	125	126	127	128	129	130
yield	19	15	15	16	19	12	16	14	15	17
Day	131	132	133	134	135	136	137	138	139	140
yield	16	20	15	19	16	16	20	20	18	21

NOTES

Day	*141*	*142*	*143*	*144*	*145*	*146*	*147*	*148*	*149*	*150*
yield	*22*	*22*	*21*	*22*	*21*	*21*	*21*	*18*	*20*	*17*
Day	*151*	*152*	*153*	*154*	*155*	*156*	*157*	*158*	*159*	*160*
yield	*20*	*20*	*21*	*21*	*21*	*20*	*20*	*16*	*16*	*15*
Day	*161*	*162*	*163*	*164*	*165*	*166*	*167*	*168*	*169*	*170*
yield	*18*	*19*	*18*	*20*	*19*	*18*	*16*	*14*	*14*	*13*
Day	*171*	*172*	*173*	*174*	*175*	*176*	*177*	*178*	*179*	*180*
yield	*16*	*16*	*16*	*18*	*16*	*15*	*16*	*18*	*18*	*15*
Day	*181*	*182*	*183*	*184*	*185*	*186*	*187*	*188*	*189*	*190*
yield	*18*	*16*	*17*	*18*	*17*	*16*	*13*	*14*	*13*	*12*
Day	*191*	*192*	*193*	*194*	*195*	*196*	*197*	*198*	*199*	*200*
yield	*16*	*10*	*13*	*8*	*8*	*6*	*8*	*9*	*4*	*5*
Day	*201*	*202*	*203*							
yield	*6*	*6*	*4*							

Find the efficiency of systematic sampling at 7 and 14 and day's interval of recording, with respect to corresponding simple radom sampling, in estimating the lactation yield of the cow.

Solution. We arrange the given data in the following Table :

(*i*) Here, $N = 203$, $k = 7$, $n = \frac{N}{k} = \frac{203}{7} = 29$.

Day → ↓	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	10	11	14	10	14	9	10
2	8	11	10	6	9	8	7
3	9	10	11	11	13	12	12
4	10	11	11	14	15	12	17
5	18	16	13	14	14	15	16
6	16	16	13	16	17	14	16
7	15	14	14	15	17	15	16
8	17	25	22	23	19	18	16
9	22	21	21	23	21	19	19
10	19	19	19	19	19	19	19
11	18	19	21	20	17	16	18
12	18	18	22	22	22	20	20
13	20	18	20	21	21	20	18

NOTES

14	21	22	22	20	21	21	21
15	21	21	19	20	21	20	21
16	20	21	20	21	20	19	21
17	18	21	20	22	21	21	21
18	16	19	15	15	16	19	12
19	16	14	15	17	16	20	15
20	19	16	16	20	20	18	21
21	22	22	21	22	21	21	21
22	18	20	17	20	20	21	21
23	21	20	20	16	16	15	18
24	19	18	20	19	18	16	14
25	14	13	16	16	16	18	16
26	15	16	18	18	15	18	16
27	17	18	17	16	13	14	13
28	12	16	10	13	8	8	6
29	8	9	4	5	6	6	4
Total	477	495	481	494	486	472	465
Mean $\bar{y}_{r*}$	16.45	17.07	16.59	17.03	16.76	16.28	16.03

$$\therefore \quad Y = \sum_{i=1}^{7} y_{i*} = 3370$$

Taking random start $i = 5$, we get the 5th sample.

In this case, the unbiased estimate of the total milk yield is given by

$$\hat{Y}_{s_y} = 7 \times 486 = 3402$$

The variance of systematic sampling for the population total estimate is given by

$$= \frac{N^2}{k}\left[\sum_{i=1}^{k} \bar{y}_{i.}^2 - \frac{Y^2}{k}\right]$$

$$= \frac{203^2}{7}\left[16.45^2 + 17.07^2 + 16.59^2 + 17.03^2 + 16.76^2 + 16.28^2 + 16.03^2 - \frac{3370^2}{7}\right] = 5192$$

The variance of simple random sample estimate of total milk yield is given by

$$V\left(\hat{Y}_{SR}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

where, $S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{Y^2}{N}\right]$

$\therefore \quad V\left(\hat{Y}_{SR}\right) = (203)^2\left(\frac{1}{29} - \frac{1}{203}\right)\frac{3865.7228}{202} = 23194.33$

NOTES

Hence, the relative precision of systematic sampling over simple random sampling will be

$$\frac{23194.33}{5192} \times 100 = 446.73\%$$

(*ii*) We arrange the given data for 14 day's in the table given below. For N = 203, k = 14, we have n = 15 or 14

Taking a random start i = 10, we get the 10th sample. In this case, the unbiased estimate of total milk yield is given by

$$\hat{Y}_{s*} = 14 \times 246 = 3444$$

Its variance is given by,

$$V\left(\hat{Y}_{s*}\right) = k\sum_{i} y_i^2 - Y^2 = 14 \times 812028 - (2270)^2 = 11492$$

Day→ ↓	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
1	10	11	14	10	14	9	10	8	11	10	6	9	8	7
2	9	10	11	11	13	12	12	10	11	11	14	15	12	17
3	18	16	13	14	14	15	16	16	16	13	16	17	14	16
4	15	14	14	15	17	15	16	17	25	22	23	19	18	16
5	22	21	21	23	21	19	19	19	19	19	19	19	19	19
6	18	19	21	20	17	16	18	18	18	22	22	22	20	20
7	20	18	20	21	21	20	18	21	22	22	20	21	21	21
8	21	21	19	20	21	20	21	20	21	20	21	20	19	21
9	18	21	20	22	21	21	21	16	19	15	15	16	19	12
10	16	14	15	17	16	20	17	19	16	16	20	20	18	21
11	22	22	21	22	21	21	21	18	20	17	20	20	21	21
12	21	20	20	16	16	15	18	19	18	20	19	18	16	14
13	14	13	16	16	16	18	16	15	16	18	18	15	18	16
14	17	18	17	16	13	14	13	12	16	10	13	8	8	6
15	8	9	4	5	6	6	4	–	–	–	–	–	–	–
Total (y_{i*})	249	247	246	248	247	241	238	228	248	235	246	239	231	227
Mean ($\bar{y}_{i*}$)	16.60	16.47	16.40	16.53	16.47	16.06	15.87	16.28	17.71	16.78	17.57	17.07	16.50	16.21

NOTES

The variance of a simple random sample estimate of total milk yield will be

$$V\left(\hat{Y}_{SR}\right) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S^2$$

$$= (203)^2\ (1/14 - 1/203) \times 3865.7228/202$$

$$= 52187.2578$$

Therefore, the relative precision of systematic estimate over simple random sample estimate is

$$52187.2578 \times 100/11492 = 454.12\%$$

4.8 SUMMARY

- The method of selecting only the first unit with the help of random numbers, the rest being selected automatically according to a predetermined pattern is known as systematic sampling.
- Variance of a stratified sample is $\frac{1}{n}^{th}$ time the variance of a systematic sample.
- Variance of a systematic sample is $\frac{1}{n}^{th}$ time the variance of a random sample.
- Stratified sampling is the most efficient technique as compared to systematic sampling and random sampling.

4.9 GLOSSARY

- **Systematic sample :** A sample which contains the n units i, $i + k$, $i + 2k$, ..., $i + (n - 1)\ k$, is known as a systematic sample.

4.10 REVIEW QUESTIONS

1. What is systematic sampling? Give the circumstances under which it is to be preferred to simple random sampling. Explain how you will estimate the variance of a systematic sample with a random start.
2. Discuss the situations under which systematic samples are preferred to other types of samples in censuses and surveys. Show that a systematic sample mean is a more efficient estimator of the population mean than a simple random mean, but less efficient than a stratified random sample mean in a population with linear trend.

NOTES

3. Describe precisely the procedure of drawing a systematic sample of n units from a population of N units, where N is not necessarily a multiple of n. (*i*) How would you estimate the mean of a characteristic in the population and examine its unbiasedness? (*ii*) How would you combine the estimates from three systematic samples of size n into a single-pooled estimate which is unbiased and estimate its standard error?

4. In a two dimensional population with $N^2 = n^2k^2$ units, the linear trend is given by $y_{ij} = i + j$ ($i, j = 1, 2, ..., nk$), where y_{ij} is the item value in the i^{th} row and j^{th} column.

   A systematic square grid sample of n^2 units is taken by taking two independent random starts (i_1, j_1) each between 0 and k and $(i_1 + xk, j_1 + yk)$ for $x, y = 0, 1, ..., (n - 1)$. Show that the mean of this sample has the same precision as the sample mean of size n^2.

5. In a population with quadratic trend $y_i = i^2$ ($i = 1, ..., 25$), compare the values of $E\left(\bar{y}_{sy} - \bar{Y}\right)^2$ given by every k^{th} systematic sample of size 5 by (*i*) Yates method (*ii*) Singh *et al* method.

6. Data on the number of seedings in every individual foot of a sown bed which is 80 feet in length, are given below :

1–10	*11–20*	*21–30*	*31–40*	*41–50*	*51–60*	*61–70*	*71–80*
26	16	27	37	4	36	20	21
28	9	20	14	5	20	21	26
11	22	25	14	11	43	15	16
16	26	39	24	9	27	14	18
7	17	24	18	25	20	13	11
22	39	25	17	16	21	9	19
44	21	18	14	13	18	25	27
26	14	44	38	22	19	17	29
31	40	55	36	18	27	7	31
26	30	39	29	9	30	30	29

   (*i*) Find the standard error of the estimate of the total number of seedings based on a systematic sample consisting of every 10^{th} foot of the sown bed.

   (*ii*) Also find the relative efficiency of systematic sampling when compared to :

   (*a*) simple random sampling, wor, with the sample size as 8 one-foot bed lengths.

   (*b*) a stratified sample of size 8 with 2 units per stratum.

NOTES

7. For investigating the possibility of estimating the catch of marine fish, a pilot survey was conducted in a sample of fishing centres on the Malabar Coast of India. At each landing centre in the sample, a count was made of the number of boats landing every hour from 6 am to 6 pm. Of the boats landing each hour, the first one was selected for observation on the weight of fish, product of this with the number of boats giving an estimate of the fish during the hour. Data on the number of boats landing and the catch of fish at a particular centre on a particular day are given below :

Hours	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
No. of boats (x)	42	52	19	5	23	56	36	59	14	14	2	6
Catch of fish (y) (quintals)	563	887	223	88	352	1295	934	1265	466	433	98	0

Calculate the relative efficiencies of linear systematic sampling as compared to srs wor for estimating the population totals of x and y when the sample sizes are 2, 3, and 6, taking each hour as the sampling unit.

8. Table given below furnishes complete enumeration data on the length of strip (x) and volume of timber (y) for each strip in three blocks of the Black Mountain Forest, California.

Block I			*Block II*			*Block III*		
Strip no.	x	y	*Strip no.*	x	y	*Strip no.*	x	y
1	12	762	1	9	471	1	6	165
2	12	651	2	9	426	2	6	224
3	12	461	3	9	448	3	6	192
4	12	521	4	9	402	4	6	161
5	12	653	5	9	372	5	6	104
6	12	544	6	9	372	6	5	94
7	12	542	7	9	411	7	5	102
8	12	590	8	9	323	8	5	115
9	11	533	9	9	381	9	4	110
10	11	517	10	9	430	10	4	109
11	11	520	11	9	434	11	4	83
12	11	539	12	9	324	12	4	36
13	10	509	13	9	543	13	4	61
14	10	449	14	9	607	14	4	92
15	10	492	15	8	416	15	4	75
16	10	498	16	8	326	16	4	64

NOTES

(*i*) Examine the behaviour of the sampling variance of estimates of volume of timber based on systematic samples of sizes 2, 3, 4, 6, 8 and 12.

(*ii*) Compare the efficiency of systematic sampling with those of simple random sampling with and without replacement for the sample sizes considered in (*i*).

(*iii*) Study the efficiency of sampling the strips with probability proportional to the length of the strips with replacement.

9. Given below are data for 10 systematic samples of size 4 from a population of 40 units.

Systematic Sample Numbers

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
0	1	2	1	4	5	6	7	7	9
7	8	9	10	12	13	15	6	16	17
18	18	19	20	21	20	24	13	28	29
29	30	31	31	33	32	35	37	38	63

Work out the relative efficiency of systematic sampling over random sampling.

10. Table given below shows a list of 70 villages in a tehsil of India along with their population in 1981 and cultivated area in the same year. Making use of the population in 1981 as preliminary information, rearrange the villages in linear order for estimating the total cultivated area from a systematic sample.

S. no.	*Population*	*Cultivated area (acres)*	*S. no.*	*Population*	*Cultivated area (acres)*	*S. no.*	*Population*	*Cultivated area (acres)*
1	226	678	26	1007	680	51	441	622
2	670	663	27	1567	970	52	555	342
3	4505	1290	28	5271	1850	53	827	387
4	1732	1170	29	659	340	54	2867	322
5	2874	1390	30	3209	2450	55	726	636
6	2282	1110	31	2902	1760	56	633	410
7	793	760	32	2955	2120	57	680	427
8	895	730	33	1746	1220	58	587	496
9	1157	950	34	1045	860	59	1901	936
10	3201	1700	35	666	620	60	2419	1226

NOTES

11	1117	909	36	904	760	61	1258	836
12	1236	1169	37	773	502	62	1225	634
13	5201	1840	38	1040	532	63	1447	978
14	848	660	39	760	438	64	1314	724
15	1238	1140	40	2084	633	65	1298	422
16	1917	1360	41	828	277	66	728	493
17	1800	1509	42	4877	1640	67	851	396
18	2335	1810	43	911	424	68	786	732
19	4396	2240	44	1205	822	69	663	422
20	1607	1225	45	1139	555	70	740	370
21	2071	1250	46	4064	347			
22	2155	1690	47	1114	744			
23	7780	3200	48	547	372			
24	2746	1744	49	1178	644			
25	2549	2400	50	1159	732			

(*i*) Draw five circular systematic samples of size 7 each, from rearranged frame.

(*ii*) From each of the five samples, estimate the total cultivated area in the tehsil using the figures for cultivated area for the selected villages as given in Table.

(*iii*) Obtain a single combined estimate from the five sample estimates. Also, calculate the standard error of this combined estimate.

11. Compare systematic sampling with stratified random sampling.

12. Explain the following terms :

(*a*) Linear trend

(*b*) Periodic variation.

4.11 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

CHAPTER 5 CLUSTER SAMPLING

NOTES

OBJECTIVES

After going through this chapter, we should be able to :

- know about cluster sampling.
- know about intra-class correlation.
- know about single cluster and its size.
- know about estimates of the mean and variance.

STRUCTURE

5.1 INTRODUCTION

In sampling procedure, we divide the whole population into a finite number of distinct and identifiable units called the sampling units. The smallest units into which the population can be divided are called elements of the population and the groups of elements the clusters. When the sampling unit is a cluster, the procedure of sampling is called cluster sampling.

NOTES

For many types of population a list of elements is not available and the use of an element as the sampling unit is, therefore, not feasible. The melthod of cluster sampling is available in such cases. Thus, in a city list of all the houses may be available but that of persons is rarely so. Again, list of farms are noit available, but those of villages or enumeration districts prepared for the census are available. In such cases, cluster sampling is applicable and therefore widely practised in sample surveys.

A necessary condition for the validity of the method is that every unit of the population under study must correspond to one and only one unit of the cluster segment so that the total number of sampling units in the list or frame may cover all the units of the population under study with no omission or duplication when this condition is not satisfied, then errors or biases start interupting.

In this chapter, we shall give the relevant theory which can provide guidance in the choice of a sampling unit in a sample survey.

5.2 CLUSTER SAMPLING WITH EQUAL CLUSTER SIZE

We first consider the case of equal clusters and suppose that the population is compsed of N cluster of M elements each and that a sample of n clusters is drawn from it by the method of simple random sampling.

Denote by

y_{ij} : the value of the characteristic under study for the j^{th} element $y = 1(1) M$ in the i^{th} cluster $i = 1, ..., N$.

$\bar{y}_{i*} = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$: the mean per element of the i^{th} cluster

$\bar{\bar{y}}_{N*} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i*}$: the mean of cluster means in the population

$\bar{y}_{**} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$: the mean per element in the population

$\bar{\bar{y}}_{N*} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i*}$: the mean of cluster means in a simple random sample of n clusters.

It is clear that $\bar{\bar{y}}_{N*} = \bar{y}_{**}$ ($\because$ the clusters are of equal size)

$S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{i*})$: the mean square between elements in the i^{th} cluster $i = 1$ (i) N.

NOTES

$\bar{S}w^2 = \frac{1}{N}\sum_{j=1}^{N} S_i^2$: the mean square within clusters.

$S_b^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i*} - \bar{\bar{y}}_{N*})^2$: the mean square between cluster means in the population.

and

$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{**})^2$: the mean square between elements in the population.

Clearly, $\bar{\bar{y}}_{n*}$ is unbiased estimate of $\bar{y}_{**}$.

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

$$= \frac{1}{n}\frac{1}{M} E\left(\sum_{j=1}^{M} y_{ij}\right)$$

$$= \frac{1}{nM}\sum_{i=1}^{N}\left(\sum_{j=1}^{M} y_{ij}\right)\cdot\frac{n}{N} \quad \Big| \text{ prob. of } n \text{ cluster is } \frac{n}{N}$$

$$= \frac{1}{nM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \bar{y}_{**}$$

The var of $\bar{\bar{y}}_{n*}$ is

$$V(\bar{\bar{y}}_{n*}) = \frac{N-n}{Nn} S_b^{\,2}$$

It $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, Then

$$V(\bar{y}_n) = \frac{N-n}{N}\frac{S_y^{\,2}}{n}$$

If an equivalent sample of nM elements were selected from the population of NM elements by simple random sampling, the variable of mean per element would be

$$V(\bar{y}_{nM}) = \frac{NM-nM}{NM}\frac{S^2}{nM}$$

Thus, efficiency of a cluster as the unit of sampling compared with that of an element is given by

NOTES

$$E = \frac{V(\bar{y}_{nM})}{V(\bar{\bar{y}}_{n*})} = \frac{S^2}{MS_b^{\,2}} \qquad ...(1)$$

Equation (1) shows that the efficiency of cluster sampling increases as the mean square between cluster decreases.

Also,

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i*} - \bar{\bar{y}}_{N*})^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left[(\bar{y}_{ij} - \bar{\bar{y}}_{N*}) - (y_{ij} - \bar{y}_{i*})\right]^2$$

$$= \frac{1}{N-1}\sum_{i}\left[(\bar{y}_{ij} - \bar{\bar{y}}_{N*})^2 + (y_{ij} - \bar{y}_{i*})^2 - 2(y_{ij} - \bar{\bar{y}}_{N*})(y_{ij} - \bar{y}_{i*})\right.$$

$$\therefore \quad \sum_{j=1}^{M} S_b^{\,2} = \frac{1}{N-1}\sum_{i}\sum_{j}(y_{ij} - \bar{\bar{y}}_{N*})^2 + \frac{1}{N-1}$$

$$\sum_{i}\sum_{j}(y_{ij} - \bar{y}_{i*})^2 - \frac{2}{N-1}\sum_{i}\sum_{j}(y_{ij} - \bar{\bar{y}}_{N*})(y_{ij} - \bar{y}_{i*})$$

$$i.e.,\ M\,S_b^2 = \frac{1}{N-1}(NM-1)S^2 + \frac{1}{N-1}\sum_{i}(M-1)S_i^{\,2} - \frac{2}{N-1}(M-1)N\,\bar{S}_w^{\,2}$$

$$\sum_{i}\sum_{j}(y_{ij} - \bar{\bar{y}}_{N*})(y_{ij} - \bar{y}_{i*}) = \sum_{i,j} y_{ij}^{\,2} - \sum_{i,j} y_{ij}\bar{\bar{y}}_{N*} - \sum_{i,j} y_{ij}\bar{y}_{i*} + \sum_{i,j}\bar{\bar{y}}_{N*}\bar{y}_{i*}$$

$$= \sum_{i,j} y_{ij}^{\,2} - \bar{\bar{y}}_{N*}\sum_{i,j} y_{ij} - \sum_{i}\left[\bar{y}_{i*}\sum_{i} y_{ij}\right] + \bar{\bar{y}}_{N*}\sum_{i,j}\bar{y}_{i*}$$

$$= \sum_{i,j} y_{ij}^{\,2} - \bar{\bar{y}}_{N*}\,NM\,\bar{\bar{y}}_{N*} - \sum_{i}\bar{y}_{i*}.M\,\bar{y}_{i*} + \bar{\bar{y}}_{N*} + \bar{\bar{y}}_{N*}M\cdot N\,\bar{\bar{y}}_{N}$$

$$= \sum_{i,j} y_{ij}^{\,2} - M\sum_{i}\bar{y}_{i*}^{2}$$

$$= \sum_{i,j}(y_{ij} - \bar{y}_{i*})^2$$

$$\sum_{i,j}(y_{ij} - \bar{y}_{i*})^2 = \sum_{i,j} y_{ij}^{\,2} + M\sum_{i}\bar{y}_{i*}^{2} - 2\left[\sum_{i}\bar{y}_{i*}\left(\sum_{j}\bar{y}_{ij}\right)\right]$$

$$= \sum_{i,j} y_{ij}^{2} + M\sum_{i}\bar{y}_{i*}^{2} - 2M\,\bar{y}_{i*}^{2}$$

$$= \sum_{i,j} y_{ij}^{2} - M\sum_{i}\bar{y}_{i*}^{2})$$

$$= \sum_{i}\left[\sum_{j}(y_{ij} - \bar{y}_{i*})^2\right]$$

$$= \sum_i (M-1)S_i^2 = (M-1)\,N\bar{S}_w^2$$

Thus, $$MS_b^2 = \frac{1}{N-1}\left[(NM-1)S^2 + N(M-1)\bar{S}_w^2 - 2N(M-1)\bar{S}_w^2\right]$$

$$= \frac{1}{N-1}\left[(NM-1)S^2 + N(M-1)\bar{S}_w^2\right] \quad ...(2)$$

From equation (1) and (2), it follows that the efficiency will increase with the increase in the mean square within clusters. Thus, we see that for cluster sampling to be efficient, the clusters should be so formed that the variation between cluster means is as small as possible while the variation within clusters is as large as possible.

Now, if the clusters are formed by grouping together random sample of M elements from a population of *NM* elements, one would expect the elements, within a cluster to be neither more nor less alike than the elements in different clusters. Consequently, the mean squares between and within the clusters will be of the same order and they will behave as random variables with the same expected value.

For,

E [Mean square between clusters] = $E\,(MS_b^2)$

$$= E\left[\frac{1}{N-1}M\sum_{i=1}^{N}(\bar{y}_{i*} - \bar{\bar{y}}_{N*})^2\right]$$

$$= E\left[\frac{1}{N-1}M\left\{\sum_{i=1}^{n}\bar{y}_{i*}^2 + N\,\bar{y}_{N*}^2 - 2\bar{\bar{y}}_{N*}\sum_{i=1}^{N}\bar{y}_{i*}\right\}\right]$$

$$= \frac{M}{N-1}E\left[\sum_i \bar{y}_{i*}^2 + N\,\bar{\bar{y}}_{N*}^2 - 2\bar{\bar{y}}_{N*}\,N\,\bar{\bar{y}}_{N*}\right]$$

$$= \frac{M}{N-1}\left[E\left(\sum_i \bar{y}_{i*}^2\right) - N\,\bar{y}_{N*}^2\right]$$

$$= \frac{M}{N-1}\left[\sum_{i=1}^{N}E\left(\bar{y}_{i*}^2\right) - N\,\bar{\bar{y}}_{N*}^2\right] \quad ...(3)$$

Now, $$E(\bar{y}_{i*}^2) = E\left[\bar{y}_{i*}^2 - \bar{\bar{y}}_{N*}^2 + \bar{\bar{y}}_{N*}^2\right]$$

$$= \bar{\bar{y}}_{N*}^2 + E\left[\bar{y}_{i*}^2 - \bar{y}_{N*}^2\right]$$

$$= \bar{\bar{y}}_{N*}^2 + \left\{E(\bar{y}_{i*}^2) - (E(\bar{y}_{i*}1)^2\right\} \qquad (\because E(\bar{y}_{i*}) = \bar{\bar{y}}_{N*})$$

$$= \bar{\bar{y}}_{N*}^2 + V(\bar{y}_{i*})$$

$\therefore$ $$\mathrm{E}\left[\bar{y}_{i*}^2\right] = \bar{\bar{y}}_{N*}^2 + \frac{NM-M}{NM}\frac{1}{M}S^2 \qquad \left| \; V(\bar{y}_n) = \frac{N-n}{N}\frac{1}{n}S_y^2\right.$$

NOTES

Thus, (3) gives

$$E\text{ [Mean Squares between clusters]} = \frac{M}{N-1}\left[\sum_{i-1}^{N}\left\{\bar{\bar{y}}_{N^*}^{\,2}+\frac{NM-M}{NM}\frac{S^2}{M}\right\} -N\,\bar{\bar{y}}_{N^*}^{\,2}\right] = S^2$$

Similarly,

$$E\text{ [Mean square within clusters]} = E\left[\bar{S}_w^{\,2}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}S_i^{\,2}\right]$$

$$= E\left[\frac{1}{N}\sum_i\sum_j\frac{1}{M-1}(y_{ij}-\bar{y}_{i*})^2\right]$$

$$= \frac{1}{N(M-1)}E\left[\sum_{i,j}(y_{ij}-\bar{y}_{i*})^2\right]$$

$$= \frac{1}{N(M-1)}\left[\sum_{i,j}E(y_{ij}^2)-M\sum_i E(\bar{y}_{i*}^2)\right]$$

$$= \frac{1}{N(M-1)}\left[NM\left(\bar{y}_{N*}^2+\frac{NM-1}{NM}S^1\right) -MN\left(\bar{\bar{y}}_{N*}^2+\frac{NM-M}{NM}\frac{S^2}{M}\right)\right]$$

$$\sum_{i,j}E(y_{ij}^2) = \sum_{i,j}E\left[y_{ij}^2-\bar{\bar{y}}_{N*}^2+\bar{\bar{y}}_{N*}^2\right]$$

$$= NM\,\bar{\bar{y}}_{N*}^2+E\sum_{i,j}\left(y_{ij}^2-\bar{\bar{y}}_{N*}^2\right)$$

$$= NM\,\bar{\bar{y}}_{N*}^2+E\sum_{i,j}\left(y_{ij}-\bar{\bar{y}}_{N*}\right)^2$$

$$= NM\,\bar{\bar{y}}_{N*}^2+(NM-1)E(S)^2 = \text{NM}\;\bar{\bar{y}}_{N*}^2+(NM-1)S^2$$

as S^2 is constants here

or $$E(S^2) = \sum_i S^2 pi S^2))$$

$\therefore$ E[Mean square within clusters] $= S^2$

Thus, it follows that if clusters are formed of random samples of the elements of the population, they will, on the average, be as efficient as the individual elements themselves.

5.3 VARIANCE OF ESTIMATED MEAN FOR CLUSTER SAMPLING

NOTES

A cluster cannot be regarded as comprised of a random sample of elements of the population. Usually, elements of the same cluster will resemble each other more than those belonging to different clusters. Consequently the variable of an estimate based on cluster sampling will ordinarily exceed than based on an equivalent sample of elements selected individually at random. The manner is which the variance of the estimate increases with the size of the cluster can best be studied with the help of the concept of intra-class correlation between elements of a cluster.

Let ρ denote the intra-class correlation defined by

$$\rho = \frac{E\left[(y_{ij} - \bar{\bar{y}}_{N*})(y_{ik} - \bar{\bar{y}}_{N*})\right]}{E\,(y_{ij} - \bar{\bar{y}}_{N*})^2} \qquad ...(1)$$

The numerator in (1) can be written as

$$E\left[\left\{(y_{ij} - \bar{y}_{i*}) + (\bar{y}_{i*} - \bar{\bar{y}}_{N*})\right\}\left\{(y_{ik} - \bar{y}_{i*}) + (\bar{y}_{i*} - \bar{\bar{y}}_{N*})\right\}\right]$$

$$= E\left[\left\{(y_{ij} - \bar{y}_{i*})(y_{ik} - \bar{y}_{i*})\right\} + \left\{(y_{ik} - \bar{y}_{i*})(\bar{y}_{i*} - \bar{\bar{y}}_{N*})\right\}\right.$$

$$\left. + \left\{(y_{ij} - \bar{y}_{i*})(\bar{y}_{i*} - \bar{\bar{y}}_{N*})\right\} + (\bar{y}_{i*} - \bar{\bar{y}}_{N*})^2\right]$$

$$= E\left[(y_{ij} - \bar{y}_{i*})(y_{ik} - \bar{y}_{i*})\right] + E\left[(\bar{y}_{i*} - \bar{\bar{y}}_{N*})^2\right] \qquad ...(2)$$

$$(\because\ E\left[(y_{ik} - \bar{y}_{i*})(y_{i*} - \bar{\bar{y}}_{N*})\right] = E\left[y_{ik} - \bar{y}_{i*} - \bar{y}_{i*}^2 - y_{ik}\,\bar{\bar{y}}_{N*} + \bar{y}_{i*}\,\bar{\bar{y}}_{N*}\right]$$

$$= E\,(y_{ik}\,\bar{y}_{i*}) - E(\bar{y}_{i*}^2) - \bar{y}_{N*}\,E(y_{ik}) + \bar{\bar{y}}_{N*}\,E(\bar{\bar{y}}_{i*})$$

Now, $$E\,(y_{ik}\,\bar{y}_{i*}) = \frac{1}{MN}\sum_i\sum_k y_{ik}\bar{y}_{i*} = \frac{1}{MN}\sum_i\left[\bar{y}_{i*}\sum_k y_{ik}\right]$$

$$= \frac{1}{N}\sum_i \bar{y}_{ix}\bar{y}_{i*} = \frac{1}{N}\sum_i \bar{y}_{i*}^2 = E(\bar{y}_{i*}^2)$$

Also, $$E(y_{ik}) = \frac{1}{NM}\sum_{i,j} y_{ik} = \bar{\bar{y}}_{N*}$$

and, $$E(\bar{y}_{i*}) = E\,\bar{\bar{y}}_{N*}$$

$$E\left[(y_{ik} - \bar{y}_{i*})(\bar{y}_{i*} - \bar{\bar{y}}_{N*})\right] = E\,(\bar{y}_{i*}^2) - E\,(\bar{y}_{i*}^2) - \bar{\bar{y}}_{N*}^2 + \bar{\bar{y}}_{N*}^2 = 0$$

Similarly, $$E\left[(y_{ik} - \bar{y}_{i*})(\bar{y}_{i*} - \bar{\bar{y}}_{N*})\right] = 0$$

NOTES

$$\text{Now, } E\left[(y_{ij}-\bar{y}_{i*})(y_{ik}-\bar{y}_{i*})|i\right] = \frac{1}{M(M-1)}\sum_{j\neq k=1}^{M}(y_{ij}-\bar{y}_{i*})(y_{ik}-\bar{y}_{i*})$$

$$= \frac{1}{M(M-1)}\left\{\left[\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i*})\right]^2 - \sum_{j=1}^{M}(y_{ij}-\bar{y}_{i*})^2\right\}$$

$$= \frac{1}{M(M-1)}\left\{\left(\left(\sum_{j=1}^{M}y_{ij}\right)\left(-M\bar{y}_{i*}\right)^2\right) - \sum_{j=1}^{M}(y_{ij}-\bar{y}_{i*})^2\right\}$$

$$= \frac{1}{M(M-1)}\left[\left(M\bar{y}_{i*}-M\bar{y}_{ix}\right)^2 - (M-1)S_i^2\right] = -\frac{S_i^2}{M} \quad ...(3)$$

Next, taking the expectations for varying i, we get

$$E\left[(y_{ij}-\bar{y}_{i*})(y_{ik}-\bar{y}_{i*})\right] = -\frac{1}{N}\sum_{i=1}^{N}\frac{S_i^2}{M} \qquad (\because E\ (E\ (X/4)) = E(N)$$

$$= -\frac{\bar{S}_w^2}{M} \quad ...(4)$$

$$\text{Now, } E\left[y_{i*}-\bar{\bar{y}}_{N*}\right]^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_{i*}-\bar{\bar{y}}_{N*}\right)^2 = \frac{1}{N}.(N-1)S_b^2 \quad ...(5)$$

$$\text{Also, } E\left[y_{ij}-\bar{\bar{y}}_{N*}\right]^2 = \frac{1}{NM}\sum_i\sum_j(y_{ij}-\bar{\bar{y}}_{N*})^2 = \frac{(NM-1)S^2}{NM}. \quad ...(6)$$

Using (2), (4), (5) and (6) in (1), we get

$$\rho = \frac{\frac{N-1}{N}S_b^2 - \frac{\bar{S}_w^2}{M}}{\frac{NM-1}{NM}.S^2} \quad ...(7)$$

Now, when the clusters are randomly formed, it is known that

$$E(MS_b^2) = S^2 \quad \text{and} \quad E(\bar{S}_w^2) = S^2$$

$$\therefore \quad E(\rho) = \left(\frac{1}{NM-1}\right) \quad ...(8)$$

Now, by ANOVA, we know that

Total S.S. = S.S. between clusters + S.S. within clusters

$$ie., \quad \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{\bar{y}}_{N*}\right)^2 = M\sum_{i=1}^{N}\left(\bar{y}_{i*}-\bar{\bar{y}}_{N*}\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_{i*}\right)^2$$

$$\text{or} \quad (NM-1)\,S^2 = (N-1)\,MS_b^2 + N(M-1)\,\bar{S}_w^2 \quad ...(9)$$

Eliminating $\overline{S}_w^{\,2}$ from (7) and (9), we get

$$S_b^{\,2} = \frac{NM-1}{M(N-1)}\frac{S^2}{M}[1+(M-1)\rho] \qquad ...(10)$$

Next, eleminating $S_b^{\,2}$ from same equations, we get

$$\overline{S}_w^{\,2} = \frac{NM-1}{NM}S^2[(1-\rho)] \qquad ...(11)$$

Now, the variance of the mean of n cluster means in terms of intra-class correlation ρ is obtained from

$$V(\overline{\overline{y}}_{n*}) = \frac{N-n}{Nn}S_b^{\,2} = \frac{N-n}{Nn}\frac{S^2}{M}\frac{NM-1}{M(N-1)}[1+(M-1)\rho]$$

$$\cong \frac{S^2}{nM}[1+(M-1)\rho] \qquad ...(12)$$

if N is sufficiently large.

Again, the relative efficiency of cluster sampling in terms of intra-class correlation is given by

$$R.E. = \frac{S^2}{MS_b^{\,2}} = \frac{S^2}{M}\frac{M(N-1)M}{(NM-1)S^2}\frac{1}{1+(M-1)\rho}$$

$$\cong \frac{1}{1+(M-1)\rho} \text{ for } N \text{ large} \qquad ...(13)$$

Next, the variance of mean of n cluster means can also be obtained as follows :

Denote by

$$\sigma_b^{\,2} = \frac{N-1}{N}S_b^{\,2} : \text{the variable of the mean of a single cluster.} \qquad ...(14)$$

$$S_b^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}(\overline{y}_{i*}-\overline{\overline{y}}_{N*})^2 = \frac{1}{N-1}N\left\{\frac{1}{N}\sum_i(\overline{\overline{y}}_{i*}-\overline{\overline{y}}_{N*})^2\right\} = \frac{N}{N-1}\sigma_b^{\,2}.$$

$$\overline{\sigma}_w^{\,2} = \frac{M-1}{M}\overline{S}_w^{\,2} : \text{the variable of a single element within a cluster.} \qquad ...(15)$$

$$\overline{S}_w^{\,2} = \frac{1}{N}\sum_i S_i^2 = \frac{1}{M-1}\cdot\frac{1}{N}\sum_i\sum_j(y_{ij}-\overline{y}_{i*})^2$$

$$= \frac{1}{M-1}.M\frac{1}{NM}\sum_i\sum_j(y_{ij}-\overline{y}_{i*})^2 = \frac{M}{M-1}.\overline{\sigma}_w^{\,2}$$

NOTES

NOTES

and $\sigma^2 = \frac{NM-1}{NM} S^2$: the variance of a single element drawn at random from the whole population. ...(16)

$$S^2 = \frac{1}{NM-1} \sum_{i,j} \left(y_{ij} - \bar{\bar{y}}_{N*}\right)^2 = \frac{NM}{NM-1} \cdot \frac{1}{NM}$$

$$\sum_i \sum_j \left(y_{ij} - \bar{\bar{y}}_{N*}\right)^2 = \frac{NM}{NM-1} \sigma^2$$

On substituting the values of S_b^2, $\bar{S}_w^2$ and S^2 from (14), (15) and (16) in (7), (9), (10), (11) and (12), we get

$$\rho = \frac{1}{\sigma^2}\left[\sigma_b^2 - \frac{\bar{\sigma}_w^2}{M-1}\right]$$

$$\sigma^2 = \sigma_b^2 \, \bar{\sigma}_w^2$$

$$\sigma_b^2 = \frac{\sigma^2}{M}[1 + (M-1)\rho]$$

$$\bar{\sigma}_0^2 = \frac{M-1}{M} \sigma^2 (1-\rho)$$

and

$$V(\bar{\bar{y}}_{n*}) = \left(\frac{N-n}{N-1}\right)\left(\frac{\sigma^2}{nM}\right)[1 + (M-1)\rho]$$

This formula for variance is made up of three factors. The Ist factor is finite multiplier. The IInd is the variable of mean based on nM elements selected with replacement at random. The IIIrd factor measures the contribution of cluster sampling to the variable.

Now if $M = 1$, we are left with first two factors only. If $M > 1/\rho$ will, therefore measure the relative change in the sampling variance brought about by sampling clusters instead of elements. In usual practice ρ is positive and diseases as M increases, but the rate of decrease is small relative to the rate of increase in M, so that ordinarily, increase in the size of a cluster leads to substantial increase in the sampling variance of the sample estimate.

5.4 ESTIMATION FROM THE SAMPLE OF THE EFFICIENCY OF CLUSTER SAMPLING

Data for complete population is usually unknown. Thus the information available is sample of clusters and ANOVA of the elements in the sample.

NOTES

We want to assess the relative efficiency of cluster sampling from the sample data.

Let the sample consists of n clusters. The ANOVA for the sample will take the form :

Source of variation	*D.F.*	*Mean Squares*
Between Clusters	$n-1$	$\frac{1}{n-1}\sum_{i=1}^{n} M\left(\bar{y}_{i*}-\bar{\bar{y}}_{n*}\right)^2 = MS_b^{\,2}$
within Clusters	$n\,(M-1)$	$\frac{1}{n(M-1)}\sum_{i=1}^{n}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_{i*}\right)^2 = \bar{s}_{\omega}^{\,2}$
Total Sample	$nM-1$	$\frac{1}{nM-1}\sum_{i=1}^{n}\sum_{j=1}^{M}\left(y_{ij}-\bar{\bar{y}}_{n*}\right)^2 = s^2$

In a random sample of clusters, we have

$$E(s_b^2) = S_b^2, \qquad E\ (\bar{s}_w^{\,2}) = (\bar{S}_w^{\,2})$$

However s^2 will not be an unbiased estimate of S^2, since the element on which it is based cannot be considered to be a simple random sample of elements from the population of NM units. An unbiased estimate of S^2 is however can be obtained from the relation.

$$(NM-1)S^2 = (N-1)M\ S_b^{\,2} + N(M-1)\ \bar{S}_{\omega}^{\,2}$$

$$\text{Est. } S^2 = \frac{1}{NM-1}\left[(N-1)M \text{ Est. } S_b^{\,2} + N(M-1) \text{ Est. } \bar{S}_{\omega}^{\,2}\right]$$

$$= \frac{(N-1)Ms_b^{\,2} + N(M-1)\bar{s}_w^{\,2}}{NM-1}$$

Now, Relative efficiency $= \dfrac{S^2}{MS_b^{\,2}}$

$$\therefore \qquad \text{Est. } (R.E.) = \frac{(N-1)Ms_b^{\,2} + N(M-1)\bar{s}_w^{\,2}}{(NM-1)Ms_b^{\,2}}$$

$$\cong \frac{1}{M} + \left(\frac{M-1}{M}\right)\frac{\bar{s}_w^{\,2}}{Ms_b^{\,2}} \text{ for large } N.$$

5.5 RELATIONSHIP BETWEEN THE VARIANCE OF THE MEAN OF A SINGLE CLUSTER AND ITS SIZE

NOTES

The problem now is that of estimating the variance of the estimated characteristic from a sample of clusters of any size, given the variance of an equivalent sample of clusters of a particular size. This is possible only if we can express the cluster mean square S_b^2 as a front of the cluster size M.

Fairfield Smith (1938) made an attempt to work out such a relationship. He argued that if the clusters were to consist of a random sample of elements, S_b^2 would be equal to $\frac{S^2}{M}$.

Using this fact, for most of the populations, elements of a cluster will be positively correlated, cluster means will differ from one another more than if the clusters were composed of randomly selected elements. Thus S_b^2 will ordinarily exceed S^2/M. Thus, he proposed the relationship

$$S_b^2 = \frac{S^2}{M^g}$$ where g is constant less than 1, to be calculated from the sample. ...(1)

Fairfield Smith's law, however leads to one logical difficulty. On the assumption that the total mean square between elements in the population is known and the mean square between cluster means in clusters of size M is given by the relationship proposed in (1), an expression for the within cluster mean square can be derived from

$$(NM - 1)\, S^2 = (N - 1)\, M\, S_b^2 + N\,(M - 1)\, \overline{S}_\omega^{\,2}$$

i.e.,
$$\overline{S}_\omega^{\,2} = \frac{(NM-1)S^2 - (N-1)MS^2M^{-g}}{N(M-1)} \qquad ...(2)$$

Equation (2) shows that the variability within clusters infront of N, the size of the finite population of clusters althrough strictly it should have been independent of it.

However for large N,

$$\overline{S}_\omega^{\,2} = \frac{M}{M-1} S^2\,[1 - M^{-g}] \qquad ...(3)$$

where S^2 will now represent the total mean square in the infinite population of which the finite population itself may be considered as a sample.

Equation (2) also shows that if we regard the population itself as a single cluster and N is consequently 1, the within cluster variable $\overline{S}_\omega^{\,2}$ will be ,

$$\overline{S}_\omega^{\,2} = S^2$$

NOTES

Thus, if instead of assuming the relationship (1), we assume the relation given by (3) which satisfies the condition of $\overline{S}_\omega^{\,2}$ being individual of the size of the population, the expression for the mean square between clusters can be written as

$$S_b^{\,2} = \frac{S^2}{M(N-1)}\left[\frac{NM}{M^g} - 1\right] \qquad ...(4)$$

$$((\because (NM-1)S^2 = (N-1)\,MS_b^{\,2} + N(M-1)\,\overline{S}_\omega^{\,2}$$

$$= (N-1)\,MS_b^{\,2} + N(M-1)\left[\frac{M}{M-1}S^2\left(1 - M^{-g}\right)\right]$$

$$S_b^{\,2} = \frac{S^2}{M(N-1)}\left[\frac{NM}{M^g} - 1\right]))$$

$S_b^{\,2}$ now depends on N and decreases on N increases and tends N $\frac{S^2}{M}$ as $N \to \infty$.

Mahalanobis (1940) and Jessen (1942) agreed with the Fairfield Smith's law but suggested that most economic characteristics relating to the farm data follow a slightly different law. They proposed that the mean square among the elements within a cluster is a monotonic increasing front of the size of the cluster given by

$$\overline{S}_\omega^{\,2} = aM^b \ (b > 0) \qquad ...(5)$$

where a, b are constants to be evaluated from the data.

Consequently, assuming this law to hold for the mean square within clusters, the expression for the mean square between cluster means is obtained as

$$S_b^{\,2} = \frac{(NM-1)\,S^2 - N(M-1)\,aM^b}{M(N-1)} \qquad ...(6)$$

The constants S^2, a, b are evaluated from the data.

S^2 will be estimated from

$$\text{Est. } S^2 = \frac{(N-1)\,Ms_b^{\,2} + N(M-1)\,\overline{s}_w^{\,2}}{MN-1}$$ and a, b can be obtained if we have the estimate of the mean square between elements within clusters for at least two values of M.

If we regard the total population as a single cluster containing NM elements,

$$S^2 = a(NM)^b \qquad ...(7)$$

NOTES

We may substitute the estimates of $\overline{S}_{\omega}^{\,2}$ and S^2 in (5) and (7) and solve for a and b. These values may now be substituted in

$$S_b^{\,2} = \frac{(NM-1)\,a(NM)^b - N(M-1)aM^b}{M(N-1)}$$

5.6 OPTIMUM CLUSTER SIZE FOR FIXED COST

It has been observed that cluster sampling leads to loss of precision. On the other hand, it helps to reduce the cost of a survey. In this article, we shall consider the problem of determining the optimum size of cluster which will provide the maximum information for minimum funds available.

The cost of survey based on a sample of n clusters will, apart from the overhead costs on planning and analysis, be made up of

(a) Costs due to time spent on enumerating all the elements in the sample, nM in number, including the time spent and cost of transportation within clusters and

(b) Costs due to time spent and the cost of transportation between clusters.

The cost of survey can, therefore, be expressed as a sum of two components one of which is proportional to the number of elements in the sample and the other proportional to the distance to be travelled between clusters, *i.e.*,

$$C = c_1 nM + c_2 d \qquad ...(1)$$

where C_1 represents the cost of enumerating an element including the travel cost from one element to another within the cluster, c_2 that of travelling a unit distance between clusters and d distance between clusters. It has been proved impirically that the expected value of the minimum distance between n points located at random is proportional to

$$n^{1/2} - n^{-1/2}$$

Of course, Jensen proved experimentally that the approximation $n^{1/2}$ works well in practice. Thus, taking the value of $d = n^{1/2}$, equation (i) becomes

$$C = c_1 nM + c_2 n^{1/2} \qquad ...(2)$$

where c_2 will now be proportional to the cost of travelling a unit distance.

We note that if the variance within clusters is assumed to follow the following law :

$$\overline{S}_w^{\,2} = aM^b \qquad ...(3)$$

then the variance of the estimated mean per element based on a sample of n clustrs of size M each is given by

$$V(\bar{\bar{y}}_{n*}) = \frac{N-n}{N} \cdot \frac{S_b^2}{n} \qquad ...(4)$$

where $$S_b^2 = \frac{(NM-1)S^2 - N(M-1)aM^b}{M(N-1)} \qquad ...(5)$$

NOTES

Substituting S_b^2 from equation (5), equation (4) gives

$$V(\bar{\bar{y}}_{n*}) \cong \frac{1}{n}\{S^2 - (M-1)aM^b - 1\} \qquad ...(6)$$

where the finite multiplier is ignored.

Thus, for fixed cost and for minimum value of the variance $V(\bar{\bar{y}}_{n*})$, the optimum value of n, *i.e.*, cluster size can be calculated.

5.7 UNEQUAL CLUSTERS

Estimates of the Mean and their Variance

Let the i^{th} cluster consists of M_i elements $i = 1, 2, \ldots N$ and $M_o = \sum_{i=1}^{N} M_i$ denote the total number of elements. Let $\overline{M} = \frac{M_o}{N}$ be the average number of elements per cluster in the population. Then the mean of the characterstic per element in the i^{th} cluster will be given by

$$\bar{y}_{i*} = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$$

The mean per element in the population is given by

$$\bar{y}_{**} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{M_i} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{y}_{i*}}{M_o}$$

Several estimates of the population value of the mean per element can be formed from a random sample of n clusters.

I. Estimate $\bar{\bar{y}}_{n*}$. Consider A.M. of the cluster means given by

$$\bar{\bar{y}}_{n*} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i*} \qquad ...(1)$$

Now, $$E(\bar{\bar{y}}_{n*}) = E\left[\frac{1}{n}\sum^{n}\bar{y}_{i*}\right] = \frac{1}{N}.\sum_{i=1}^{N}\bar{y}_{i*} = \bar{\bar{y}}_{N*}$$

NOTES

$\Rightarrow \bar{\bar{y}}_{n*}$ is not an unbiased estimate of the population mean.

In general $\bar{\bar{y}}_{N*} \neq \bar{\bar{y}}_{**}$, since the clusters are of unequal size.

The bias of the estimate is given by

$$\text{Bias}\,(\bar{\bar{y}}_{n*}) = \bar{\bar{y}}_{n*} - \bar{y}_{**} = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i*} - \frac{\sum_{i=1}^{N} M_i\bar{y}_{i*}}{M_0}$$

$$= \frac{\bar{M}}{M_0}\sum^{N}\bar{y}_{i*} - \frac{\sum^{N} M_i\bar{y}_{i*}}{M_0} = -\frac{1}{M_0}\sum_{i=1}^{N}\bar{y}_{i*}(M_i - \bar{M})$$

$$= -\frac{1}{M_0}\sum_{i=1}^{N}(M_i - \bar{M})\left[\bar{y}_{i*} - \bar{\bar{y}}_{N*}\right]$$

$$\because \sum^{N}\left(M_i - \bar{M}\right)\left[\bar{y}_{i*} - \bar{\bar{y}}_{N*}\right] = \sum_{i}(M_i - \bar{M})\,\bar{y}_{i*} - \bar{\bar{y}}_{N*}\sum_{i=1}^{N}(M_i - \bar{M})$$

$$= \sum_{i}(M_i - \bar{M})\,\bar{y}_{i*} - 0 \text{ an algebraic sum}$$

$$\therefore \quad \text{Bias}\,(\bar{\bar{y}}_{nx}) = -\frac{N-1}{M_0}S_{M\bar{y}} \qquad ...(2)$$

$$\left(\because \frac{1}{N-1}\sum_{i}(M_i - \bar{M})\,(\bar{y}_{i*} - \bar{\bar{y}}_{N*}) = C_N \quad (M, 5.)\right.$$

Hence, the estimate $\bar{\bar{y}}_{n*}$ is not unbiased unless M_i and $\bar{y}_{i*}$ are uncorrelated.

To, obtain M.S.E. of the estimate $\bar{\bar{y}}_{n*}$, we have

$$\text{M.S.E}\,(\bar{\bar{y}}_{n*}) = V(\bar{\bar{y}}_{n*}) + [\text{Bias}\,(\bar{\bar{y}}_{n*})]^2$$

$$= \frac{N-n}{Nn}S_b^{\,2} + \frac{(N-1)^2}{M_o^{\,2}}S_{M\bar{y}}^{\,2} \qquad ...(3)$$

NOTES

where, $$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{y}_{ix} - \bar{\bar{y}}_{N*}\right)^2$$

and $$S_{M\bar{y}} = \frac{1}{N-1}\sum_{i=1}^{N}\left(M_i - \bar{M}\right)\left(\bar{y}_{i*} - \bar{\bar{y}}_{N*}\right)$$

The unbiased of estimate of $V(\bar{\bar{y}}_{n*})$ is given by

$$\text{Est. } V(\bar{\bar{y}}_{n*}) = \left(\frac{N-n}{Nn}\right)s_b^2 \quad \text{where } s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_{i*} - \bar{\bar{y}}_{n*}\right)^2$$

Thus, from (3) M.S.E. $(\bar{y}_{n*})$ can be calculated.

II. Estimate $(\bar{y}'_{n*})$. A simple unbiased estimate of the population mean $\bar{y}_{**}$ can also be formed. Consider A.M. based on cluster totals given by

$$\bar{y}'_{n*} = \frac{1}{n\bar{M}}\sum^{n} M_i\bar{y}_{i*} = \frac{1}{n}\sum^{n}\frac{M_i}{\bar{M}}\bar{y}_{i*} \qquad ...(4)$$

Now, $$E\left(\bar{y}'_{n*}\right) = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_{i*}}{\bar{M}}\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{M_i\bar{y}_{i*}}{\bar{M}} = \frac{1}{M_0}\sum_{i=1}^{N}M_i\bar{y}_{i*} = \bar{y}_{**}$$

Thus, $\bar{y}'_{n*}$ is an unbiased estimate of $\bar{y}_{**}$.

The sampling variable of this estimate is given by

$$V(\bar{y}'_{n*}) = \frac{M-n}{N}\frac{1}{n}S_b'^2$$

where, $$S_{b'}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{M_i\bar{y}_{i*}}{\bar{M}} - \bar{y}_{**}\right]^2$$

It will be noticed that the variable of $\bar{y}'_{n*}$ depends upon the variation of the product $M_i\bar{y}_{i*}$ and is therefore likely to be larger than of $\bar{\bar{y}}_{n*}$, unless M_i and $\bar{y}_{i*}$ very in such a way that there product is almost constant.

NOTES

To obtain unbiased estimate of the variable, consider

$$S_{b'}^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i\bar{y}_{i*}}{\bar{M}} - \bar{y}_{n*}\right)^2$$

It can be seen that $S_{b'}^{\,2}$ is an unbiased estimate of $S_{b'}^{\,2}$. Thus it follows that an unbiased estimate of the variable is given by

$$\text{Est. } V(\bar{y}_{n*}) = \frac{N-n}{N}\cdot\frac{1}{n}s_{b'}^{\,2}$$

III. Estimate $\bar{y}_{n*}''$. A third estimate which is biased but consistent is given by

$$\bar{y}_{n*}'' = \frac{\sum_{i=1}^{n} M_i\bar{y}_{i*}}{\sum_{i=1}^{n} M_i}$$

It is weighted mean of the cluster means and is the ratio of two random variables. The estimate is biased but the bias decreases with the increase in n. A first approx. to the variable of this estimate is given by replacing y_i by $M_i\bar{y}_{i*}$ and x_i by M_i in the following equation.

$$V_1(R_n) = \frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\frac{1}{\bar{x}_n^{\,2}}\frac{1}{N}\sum_{i=1}^{N}(y_i - R_N x_i)^2$$

$$\therefore \quad V_1(\bar{y}_{n*}'') = \frac{N-n}{N-1}\frac{1}{n}\frac{1}{\bar{M}^2}\frac{1}{N}\sum_{i=1}^{N}(M_i\bar{y}_{i*} - M_i\bar{y}_{**})^2$$

$$= \frac{N-n}{nN}S_{b''}^{\,2}$$

where, $$S_{b''}^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}\frac{M_i^2}{\bar{M}^2}(\bar{y}_{i*} - \bar{y}_{**})^2$$

Further, a consistent estimate of the variable is obtained from

$$\text{Est. } V_1(\bar{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - R_n x_i)^2$$

$$\therefore \quad \text{Est. } V_1(\bar{y}_{n*}'') = \frac{N-n}{N}\frac{s_b''^2}{n}$$

where, $$S_{b''}^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{M_i^2}{\bar{M}_n^2}\left(\bar{y}_{iy} - \bar{y}_{n*}''\right)^2$$

IV. Estimate $\bar{y}'''_{n*}$. We know that $\bar{\bar{y}}_{n*}$ is a biased estimate of the population mean $\bar{y}_{**}$ and that

$$\text{Bias } (\bar{\bar{y}}_*) = -\frac{N-1}{N\bar{\bar{M}}} S_{M\bar{y}}$$

Since, $$E[s_{M\bar{y}}] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(M_i - \bar{M}_n\right)\left(\bar{y}_{i*} - \bar{y}_{n*}\right)\right] = S_{M\bar{y}},$$

it follows that we can form a fourth estimate $\bar{y}'''_{n*}$ given by

$$\bar{y}'''_{n*} = \bar{y}_{n*} + \frac{N-1}{N\bar{\bar{M}}} s_{M\bar{y}} \qquad ...(5)$$

Thus, $\bar{y}'''_{n*}$ is clearly an unbiased estimate of $\bar{y}_{**}$. It is to be noted that this estimate is unbiased ratio type estimate

$$y''_R = \bar{r}_n\, \bar{x}_N + \frac{n(N-1)}{N(N-1)}(\bar{y}_n - \bar{r}_n\, \bar{x}_n) \qquad ...(6)$$

Replacing y_i by $\frac{M_i}{\bar{\bar{M}}}\bar{y}_{i*}$ and x_i by $\frac{M_i}{\bar{\bar{M}}}$, (6) reduces to (5).

The exact variable of (5) cannot be calculated but to Ist degree of appox,

$$V_1\left(\bar{y}'''_{nx}\right) = \frac{S_b'''^2}{n}$$

where, $$S_b'''^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(\frac{M_i}{\bar{\bar{M}}}\bar{y}_{i*} - \bar{y}_{y**}\right) - \frac{\bar{\bar{y}}_{N*}}{\bar{\bar{M}}}\left(M_i - \bar{M}\right)\right]^2$$

The variance of $\bar{y}'''_{n*}$ will be smaller than the variance of the ratio estimate $\bar{y}''_{n*}$, provided the regression Coeff. of $\frac{M_i}{\bar{\bar{M}}}\bar{y}_{i*}$ or $\frac{M_i}{\bar{\bar{M}}}$ is nearer to $\bar{\bar{y}}_{N*}$ then to $\bar{y}_{**}$.

It can be seen that a consistent estimate of the variable is given by

$$\text{Est. } V_1(\bar{y}'''_{n*}) = \frac{s_b'''^2}{n}$$

where, $$s_b'''^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\frac{M_i\bar{y}_{i*}}{\bar{\bar{M}}} - \bar{y}_{n*}\right) - \frac{\bar{\bar{y}}_{n*}}{\bar{\bar{M}}}\left(M_i - \bar{M}_*\right)\right]^2$$

NOTES

STUDENT ACTIVITY 1

1. Show that the efficiencly of cluster sampling increases as the mean square between cluster decreases.

STUDENT ACTIVITY 2

1. Show that the efficiency of cluster sampling increases as the mean square within the cluster increases.

NOTES

5.8 ILLUSTRATIVE EXAMPLES

Example 1. *A pilot sample survey for study of cultivation practices and yield of guava was conducted by IASRI in Uttar Pradesh (India). Out of a total of 412 bearing trees, 15 clusters of size 4 trees each were selected and yields (in kg) were recorded as given in the following Table.*

Cluster	*1st tree*	*2nd tree*	*3rd tree*	*4th tree*
1	5.53	4.84	0.69	15.79
2	26.11	10.93	19.08	11.18
3	11.08	0.65	4.21	7.56
4	12.66	32.52	16.92	37.02
5	0.87	3.56	4.81	57.54
6	6.40	11.68	40.05	5.15
7	54.21	34.63	52.55	37.96
8	1.94	35.97	29.54	25.98
9	37.94	47.07	16.94	28.11
10	56.92	17.69	26.24	6.77
11	27.59	38.10	24.76	6.53
12	45.98	5.17	1.17	6.53
13	7.13	34.35	12.18	9.86
14	14.23	16.89	28.93	21.70
15	3.53	40.76	5.15	1.25

(i) Estimate the average yield (in kg) per tree of guava in the Umerpur–Neerna village of Allahabad along with its standard error.

(ii) Estimate the intracluster correlation coefficient between trees within clusters and efficiency of cluster sampling as compared to simple random sampling.

Solution. Given, $M = 4$, $N = 103$ and $n = 15$.

Consider the following Table.

Cluster	*Mean* $(\bar{y}_i)$	$\sum_i y_{ij}^2$	$M\bar{y}_i^2$	s_i^2
1	6.71	303.8067	180.0954	41.237
2	14.58	1027.7958	854.3056	59.163
3	5.88	198.0666	138.2976	19.923
4	24.78	2874.5927	2456.1936	139.466

5	9.20	795.0185	338.5600	152.157
6	15.81	1807.5993	999.8244	269.258
7	44.84	834.4251	8042.5024	99.264
8	23.36	2845.1765	2182.7584	220.806
9	33.19	4830.9302	4406.3044	141.542
10	26.91	4287.1930	2896.5924	463.533
11	23.08	2828.4957	2130.7456	232.533
12	14.71	2184.8911	865.5364	439.787
13	15.88	1476.3314	1008.6976	155.878
14	20.44	1795.5999	1671.1744	41.475
15	12.67	1701.9235	642.1156	353.269
Total	292.04			2829.341

Calculations of Cluster Means and Their Variances

Let y_{ij} be the yield of the j^{th} tree in the i^{th} cluster,

$$i = 1, 2, ..., 15; \quad j = 1, 2, 3, 4$$

(*i*) An estimate of the average yield per tree of guava is given as

$$\bar{y}_n = \frac{1}{nM}\sum_{i}^{n}\sum_{j}^{M} y_{ij} = \frac{1}{n}\sum_{i}^{n} \bar{y}_{i.}$$

$$= \frac{292.04}{15} = 19.47 \quad \text{| using the Table}$$

Estimated variance of $\bar{y}_n$ is given by

$$v(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i}^{n}(\bar{y}_{i.} - \bar{y}_n)^2$$

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

$$= \left(\frac{1}{15} - \frac{1}{103}\right)\frac{1}{14}\ [7202.4262 - 15 \times 379.0809]$$

$$= \left(\frac{1}{15} - \frac{1}{103}\right) \times 108.3009 = 6.1686.$$

Therefore, standard error of $\bar{y}_n = \sqrt{6.1686} = 2.48$

An estimate of the efficiency of cluster sampling as compared with simple random sampling can be made in the form of analysis of variance given as below :

NOTES

ANOVA

Source of variation	*df*	*Mean square*
Between clusters	14	108.3 $(= Ms_b^2)$
Within clusters	397	188.6 $(= s_b^2)$
Total	411	249.3 $\left(= \hat{S}^2\right)$

$$RE = \frac{249.32}{4 \times 108.3} = 0.556$$

$$\hat{\rho} = \frac{(1 - 0.576)}{(4 - 1) \times 0.576} = 0.245$$

Example 2. *A survey for the estimation of the number of pepper standards and yield of pepper was conducted in Kerala State (India). The following calculations give the mean square between cluster means (S_b^2) and within mean square (S_w^2) for the number of bearing pepper standards in clusters of various sizes formed by grouping adjacent survey numbers of the seven villages.*

Cluster size	*3*	*5*	*7*	*10*	*15*	*20*
S_b^2	*20.99*	*16.30*	*13.46*	*10.15*	*8.39*	*6.97*
S_w^2	*32.96*	*33.35*	*34.45*	*36.52*	*37.23*	*38.04*

(i) Calculate the intra-class correlation coefficient and efficiencies of various clusters taken as sampling units, compared with that of an element.

(ii) Study the relationship between S_b^2 and size of cluster by fitting the following relations :

(a) $S_b^2 = (S^2 / M^g)$, *where* $S^2 = 42.259$

(b) $S_w^2 = aM^b$

Solution. In order to find the efficiency and intra-cluster correlation, we calculate S^2 using the relation

$$S^2 = \frac{M - 1}{M} S_w^2 + S_b^2,$$

when N is large, the following calculations are prepared to get the estimates of E and ρ by using

$E = (S^2 / MS_b^2)$ and $\rho = (1 - E)/(M - 1)\, E$

Cluster size (M)	3	5	7	10	15	20
S_b^2	20.99	16.30	13.46	10.15	8.39	6.97
S^2	42.96	38.98	42.99	43.02	44.14	43.11
E	0.682	0.478	0.456	0.424	0.343	0.309
ρ	0.230	0.273	0.199	0.151	0.137	0.118

Consider the following Table :

M	$\log M$	S_b^2	$\log S_b^2$	S_w^2	$\log S_w^2$	$\log M \log S_b^2$	$\log M \log S_w^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7) = (2) × (4)	(8) = (2) × (6)
3	0.4771	20.99	1.3221	32.96	1.5180	0.6356	0.7242
5	0.6990	16.30	1.2122	33.35	1.5331	0.8394	1.0646
7	0.8451	13.46	1.1219	34.45	1.5372	0.9417	1.2990
10	1.0000	10.15	1.0021	36.52	1.5625	1.0244	1.8377
15	1.1761	8.39	0.9238	37.23	1.5709	1.0808	1.5709
20	1.3010	6.97	0.8432	38.04	1.5802	1.0918	2.0558

For the relationship $S_b^2 = S^2/M^g$, the estimate of g by the least square technique is given by

$$-g = \frac{\sum^6 \log M \log S_b^2 - \frac{1}{6}\sum^6 \log M \sum^6 \log S_b^2}{\sum^6 (\log M)^2 - \frac{\left(\sum^6 \log M\right)^2}{6}}$$

On substituting values, we have

$$-g = \frac{5.6179 - 5.8946}{5.5062 - 5.0386} = -0.592$$

Thus, the relationship between S_b^2 and M is

$$S_b^2 = \frac{S^2}{M^{0.592}}$$

Similarly, for the relationship $S_w^2 = aM^b$, the values of b and a are given by

$$b = \frac{\sum^6 \log S_w^2 \log M - \frac{\sum \log S_w^2 \sum^6 \log M}{6}}{\sum^6 (\log M)^2 - \frac{\left(\sum^6 \log M\right)^2}{6}}$$

NOTES

$$a = \text{antilog}\ \frac{\left(\sum^{6} \log S_w^2 \sum^{6} \log M\right)}{6}$$

Substituting the values, we have

$$b = \frac{8.5524 - 8.5150}{5.5062 - 5.0386} = \frac{0.0374}{0.4676} = 0.080$$

$$\log a = 1.5487 - 0.080 \times 0.9164 = 1.4754$$

Therefore, $a = 29.88$

Thus, the relationship between S_w^2 and M is

$$S_w^2 = 29.88\ M^{0.080}$$

The values of s_b^2 and s_w^2 from the relationships

$$s_b^2 = S^2\ M^{0.592} \quad \text{and} \quad s_w^2 = 29.88\ M^{0.080},$$

are given below for various values of M along with their true values

M	s_b^2	S_b^2	s_w^2	S_w^2
3	22.03	20.99	32.63	32.96
5	16.28	16.30	33.99	33.35
7	13.34	13.46	34.91	34.45
10	10.80	10.15	35.92	36.52
15	8.49	8.39	37.11	37.23
20	7.16	6.97	38.04	38.04

5.9 SUMMARY

- In cluster sampling, we divide the population into finite number of groups of elements (called clusters).
- The intra-class correlation ρ is given by

$$\rho = \frac{E\left\{\left(y_{i_j} - \bar{\bar{y}}_{N*}\right)\left(y_{i_k} - \bar{\bar{y}}_{N*}\right)\right\}}{E\left(y_{i_j} - \bar{\bar{y}}_{N*}\right)^2}$$

- The relative efficiency (R.E.) of cluster sampling in terms of ρ is given as

$$\text{R.E.} = \frac{1}{1 + (n-1)\rho}$$

5.10 GLOSSARY

- **Cluster.** A group of elements from a finite population is called cluster.
- **Cluster sampling.** The sampling in which every unit is a cluster is known as cluster sampling.

NOTES

5.11 REVIEW QUESTIONS

1. If the NM elements in a population are grouped at random to form N clusters of M elements each, show that a random sample, wor, of n clusters would have the same efficiency as sampling nM elements in random sample, wor.
2. Let there by N clusters of M elements each. A sample of n clusters is taken systematically for estimating the population mean per element. Derive the sampling variance of the estimator in terms of the intracluster correlation coefficient (ρ_c) between pairs of elements in the clusters and (ρ'_c) between pairs of clusters at the samples, assuming N to be a multiple of n.
3. A population consists of N clusters each containing M elements. A simple random sample of n clusters is selected to estimate the population mean per element. Assuming S_w^2, the variance within clusters, to be given by aM^b, $b < 0$ and the cost function to be $C = c_1 nM + c_2 \sqrt{n}$, find the optimum size of the cluster for which the variance of the estimator is minimum, when cost of survey is fixed.
4. For examining the efficiency of sampling households instead of persons for estimating the population of males in a given area, the following assumptions are made

   (*i*) each household consists of 4 persons (husband, wife and two children).

   (*ii*) the sexes of children are binomially distributed.

   Show that the intracluster correlation coefficient is 1/6 and efficiency of sampling households compared to that of sampling persons is 200%. (Sukhatme, 1954).
5. A population consists of N clusters, M_1 being the size of the i^{th} cluster (i = 1, 2, ... N). Clusters are selected one-by-one by pps of clusters, *wr*. The cluster selected at the $(m + 1)$th draw is rejected if the number of distinct clusters selected in the first m draws is a pre-assigned number k. Suppose the i^{th} cluster occurs m_i times in a sample of m draws,

NOTES

$m_i = 0, 1, 2 \ldots, i = 1, 2, \ldots, N$. If $\bar{y}_{i.}$ is the mean of the i^{th} cluster, show that $\bar{y} = \sum_{i}^{N} (m_i/m)\,\bar{y}_{i.}$, is an unbiased estimator of the population mean. Also show that an unbiased estimate of the variance of this estimator is obtained by

$$v(\bar{y}) = \sum_{i}^{N} \frac{m_i (\bar{y}_{i.} - \bar{y})^2}{m(m-1)}.$$

6. For studying the cultivation practices and yield of apple, a pilot sample survey was conducted in a district of Himachal Pradesh (India). The yield (in kilogrammes) of 15 clusters of 4 trees each, selected at random out of 308 bearing trees in a village, are given below :

Cluster / Tree	*1*	*2*	*3*	*4*
1	5.53	4.84	0.69	15.79
2	26.11	10.93	10.08	11.18
3	11.08	0.65	4.21	7.56
4	12.66	32.52	16.92	37.02
5	0.87	3.56	4.81	27.54
6	6.40	11.68	40.05	5.12
7	54.21	34.63	52.55	37.20
8	1.24	35.97	29.54	25.28
9	37.94	47.07	19.64	28.11
10	54.92	17.69	26.24	6.77
11	25.52	38.10	24.74	1.90
12	45.98	5.17	1.17	6.53
13	7.13	34.35	12.18	9.86
14	14.23	16.89	28.93	21.70
15	3.53	40.76	5.15	1.25

(*i*) Estimate the average yield per tree as well as the production of apple in the village and their standard errors.

(*ii*) Estimate the intracluster correlation coefficient between trees within clusters.

(*iii*) Estimate the efficiency of cluster sampling as compared to simple random sampling.

7. A pilot sample survey was conducted to study the management practices and yield of guava in a village in Uttar Pradesh (India). Of the total

NOTES

of 412 bearing trees, 16 clusters of size 4 each were selected and their yield records (kg/tree) are given on next page.

Cluster number	*1st tree*	*2nd tree*	*3rd tree*	*4th tree*
1	5.58	4.84	0.69	15.79
2	26.11	10.93	10.08	11.18
3	11.08	0.65	4.21	7.56
4	12.66	32.52	16.92	37.02
5	0.87	3.56	4.81	23.54
6	6.40	11.68	40.05	5.12
7	54.21	34.63	52.55	37.96
8	1.94	35.97	29.54	25.28
9	37.94	47.07	19.64	29.11
10	56.92	17.69	26.24	1.90
11	27.59	38.10	24.74	6.77
12	45.98	5.17	1.47	6.53
13	7.13	34.35	12.18	9.86
14	14.23	16.89	28.93	21.70
15	3.53	40.76	5.15	1.25
16	5.17	26.11	29.54	19.64

(*i*) Estimate the average yield per tree as well as the total yield of guava in the village along with their standard errors.

(*ii*) Estimate the efficiency of cluster sampling as compared to simple random sampling.

8. A survey on pepper was conducted to estimate the number of pepper standards and production of pepper in Kerala State (India). For this, 3 clusters from 95 were selected by srs, wor. The information on the number of pepper standards recorded is given below :

Cluster number	*Cluster size*	*No. of pepper standards*
1	12	41, 16, 19, 15, 144, 454, 212, 57, 28, 76, 199
2	12	39, 70, 38, 37, 161, 38, 27, 219, 46, 128, 30, 20
3	7	115, 59, 120

NOTES

Estimate the total number of pepper standards along with standard error for the region, given $\overline{M}$ the average cluster size for the population to be 10.

9. Write the intra-class correlation formula for ρ. Use this, to obtain the relative efficiency of cluster sampling in terms of intra-class correlation.
10. Differentiate the following terms :

(*i*) $\bar{\bar{y}}_{n*}$ (*ii*) $\bar{y}_{n*}$ (*iii*) $\bar{\bar{y}}''_{n*}$ (*iv*) $\bar{y}'''_{n*}$.

5.12 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

NOTES

CHAPTER 6 SAMPLING WITH VARYING PROBABILITIES

OBJECTIVES

After going through this chapter, we should be able to :

- know the procedure of selecting a sample.
- with varying probabilities.
- know Hurwitz-Thompson estimate.
- know Midzuno system of sampling.
- know Narain method of sampling.

STRUCTURE

NOTES

6.1 INTRODUCTION

The simple random sampling does not take into account the possible importance of the larger units in the population. There are various ways of using such ancillary information for obtaining more efficient estimates of the population parameters in the sense of giving estimates with smaller S.E. One such method is to assign unequal probabilities of selection to the different units in the population. Thus when the units vary in size and the variable under study is correlated with size, the probabilities of selection may be assigned in proportion to the size of the unit.

Such a sampling procedure in which the units are selected with probabilities proportional to some measure of their size is known as **sampling with problem proportional to size** (p.p.s.)

There is a basic difference between the method of simple random sampling and pps sampling. In the former the probability of drawing a specified unit at any given draw is the same. In the later case, the probability differs from draw to draw. The theory of pps sampling without replacement is more complex than that of simple random sampling. One way of introducing simplification into the theory is to replace a selected unit before another draw is made. This would ensure that the probability of drawing a specified unit at any specific unit at any given draw is the same. In this chapter, we shall introduce the simplified theory which will be appropriate to the procedure of sampling with replacement. Before doing so, we first describe a procedure for selecting a sample with the help of pps sampling.

6.2 PROCEDURE OF SELECTING A SAMPLE WITH VARYING PROBABILITIES

It consists in associating with each unit a set of consecutive natural numbers, the size of the set being proportional to the desired probably. Thus if x_1, x_2, . . . , x_N are +ve integers proportional to the probabilities assigned to the N units in the population respectively, we associate the natural numbers 1 to x_1, with the Ist unit, $x_1 + 1$ to $x_1 + x_2$ with the IInd unit and so on. We draw a number at random from 1 to $S_N = \sum_{i=1}^{N} x_i$, say R, and select that i^{th} unit in the population for which.

$x_0 + x_1 + x_2 + \ldots + x_{i-1} < R \leq x_0 + x_1 + x_2$ in preceions line where x_0 is to be interpreted as zero. It is clear that this procedure of selection gives to the i^{th} unit in the population a probability of selection proportional to x_i. The procedure is to be repeated n times if a sample of size n is required.

The above procedure has been illustrated with the help of following example:

Example : *A village has 10 orchards containing 150, 50, 80, 100, 200, 160, 40, 220, 60 and 140 trees respectively. Select a sample of 4 orchards with replacement and pps sampling.*

Sol. The total number of trees in all the 10 orchards in the village is 1200. The first step in the selection of orchards is to form successive cumulative totals as shown below :

Sr. No.	*Size*	*Cummulative Total*
1	150	150
2	50	200
3	80	280
4	100	380
5	200	580
6	160	740
7	40	780
8	220	1000
9	60	1060
10	140	1200

Now with the help of random number table, we draw a random number not exceeding 1200 to select an orchard. Suppose this number is 600. From the cumulative total of the above table, it is seen that this number is one of the numbers from 581 to 740, associated with the 6th orchard. Thus, the 6th orchard is selected which corresponds to the random number 600. Draw three more random numbers as above. Suppose these numbers are 650, 850 and 300. Then the orchards selected corresponding to these numbers are 6th, 8th and 4th respectively. We observe that in a sample of 4 orchards selected with probability proportional to size and with replacement, the orchards selected are 4th, 6th and 8th. The 6th orchard is selected twice.

Remark : The main drawback of the above procedure is that it involves writing down the successive cumulative totals which is time consuming and tedious especially if the number of units in the population are large. To avoid the drawback of this procedure, there is another method which consists in selecting a pair of random numbers, say (i, j) such that $1 \leq i \leq N$ and $1 \leq j \leq M$, where M is the maximum of the sizes of the N units in the population. If $j \leq x_i$, the i^{th} unit is selected, otherwise it is rejected and another pair of random numbers is chosen. The procedure is repeated n time till we get the required probabilities of selection.

6.3 SAMPLING WITH REPLACEMENT—ESTIMATION OF POPULATION TOTAL AND ITS VARIANCE

NOTES

Consider a population of N units and y_i represent the value of the characteristic under study for the i^{th} unit of the population $i = 1, 2, ..., N$. Suppose further that sampling is done with replacement and p_i be the probability of selecting the i^{th} unit of the population at any given draw.

$$\therefore \quad \sum_{i=1}^{N} p_i = 1$$

Define a variate Z with values $Z_i = \frac{y_i}{Np_i}$ $i = 1, 2, ..., N$ ...(1)

and consider sample mean

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \qquad ...(2)$$

$$\therefore \quad E(\bar{Z}_n) = \frac{1}{n}\sum_{i=1}^{n} E(\bar{Z}_i) = \frac{1}{n} \cdot n\ \bar{y}_N = \bar{y}_N \qquad ...(3)$$

$$E(Z_i)\sum_{i=1}^{N} p_i z_i = \sum_{i=1}^{N} p_i \frac{y_i}{Np_i} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_N$$

Thus, $\bar{Z}_n$ is unbiased estimate of populations mean $\bar{y}_N$.

To obtain the sampling variable, we have

$$V(\bar{Z}_n) = E(\bar{Z}_n^{\,2}) - (E(\bar{Z}_n))^2$$

$$= E\left[\frac{1}{n}\sum^{n} Z_i\right]^2 - \bar{y}_N^{\,2}$$

$$= \frac{1}{n^2} E\left[\sum^{n} Z_i^{\,2} + \sum_{i \neq j}^{n} Z_i\, Z_j\right] - \bar{y}_N^{\,2}$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} E(Z_i^{\,2}) + \sum_{i \neq j}^{n} E(Z_i Z_j)\right] - \bar{y}_N^{\,2} \qquad ...(4)$$

Now $E(Z_i^{\,2}) = \sum_{i=1}^{N} p_i\, Z_i^{\,2}$

Also $E(Z_i Z_j) = E(Z_i)\, E(Z_j)$, since the draws one made with replacement.

$$\therefore \quad E(Z_i Z_j) = \bar{y}_N \cdot \bar{y}_N = \bar{y}_N^{\,2}$$

NOTES

$$\therefore \quad V(\bar{Z}_n) = \frac{1}{n}\left[\sum_{i=1}^{N} p_i Z_i^2 - \bar{y}_N^2\right]$$

$$(\because \quad V(\bar{Z}_n) = \frac{1}{n^2}\left[n \cdot \sum_{i=1}^{N} p_i Z_i^2 + \sum_{i \neq j=1}^{n} \bar{y}_N^2 - n^2 \bar{y}_N^2\right]$$

$$= \frac{1}{n^2}\left[n\sum_{i=1}^{N} p_i Z_i^2 + n(n-1)\,\bar{y}_N^2 - n^2 \bar{y}_N^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{N} p_i Z_i^2 + (n-1)\,\bar{y}_N^2 - n\bar{y}_N^2\right]$$

$$\therefore \quad V(\bar{Z}_n) = \frac{\sigma_Z^2}{n} \qquad ...(5)$$

where $$\sigma_Z^2 = \sum_{i=1}^{N} p_i (Z_i - \bar{Z}_N)^2 \qquad ...(6)$$

In case of sampling with equal probabilities, $p_i = \frac{1}{N}$.

$$\therefore \quad Z_i = y_i \quad \sigma_z^2 = \sigma_y^2 \quad \bar{Z}_n = \bar{y}_n$$

and $$V(\bar{y}_n) = \frac{\sigma_y^2}{n} = \frac{1}{n} \cdot \sum_{i=1}^{N} p_i \left[\frac{y_i}{Np_i} - \bar{y}_N\right]^2$$

$$= \frac{1}{n}\frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - \bar{y}_N)^2 = \frac{1}{n} \cdot \frac{N-1}{N} \cdot S^2 \quad ((\because Np_i = 1))$$

i.e., $$V(\bar{y}_n) = \frac{N-1}{N} \cdot \frac{S^2}{n}. \qquad ...(7)$$

Now when the selection probability is proportional to the value of the variate, in other words, when $p_i \propto y_i$, then the variate Z assumes a constant value $Z_i = \bar{y}_N \ \forall_i$

Thus, $$\sigma_Z^2 = \sum_{i=1}^{N} p_i [Z_i - \bar{y}_N]^2 = 0 \qquad ...(8)$$

Thus, we expect that when p_i is proportional to some measure of size of y_i, then the estimate may be considerably more efficient than that based on simple random sampling.

Now we estimate the sampling variable.

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Z_i - \bar{Z}_n)^2$$

NOTES

$$\therefore \quad E(s_Z^2) = \frac{1}{n-1} E\left[\sum^n Z_i^2 - n\,\overline{Z}_n^2\right]$$

$$= \frac{1}{n-1}\left[\sum^n E(Z_i^2) - n(\overline{Z}_n^2)\right] \qquad ...(9)$$

Now $V(\overline{Z}_n) = E(\overline{Z}_n^2) - \left[E(\overline{Z}_n)\right]^2$

$$\therefore \quad E(\overline{Z}_n^2) = V(\overline{Z}_n) + \bar{y}_N^2 = \frac{\sigma_Z^2}{n} + \bar{y}_N^2$$

Thus, $E[s_Z^2] = \sigma_Z^2$ ...(10) (from (9))

$$E(s_Z^2) = \frac{1}{n-1}\left[\sum_{i=1}^n \left(\sum_{i=1}^n p_i \overline{Z}_i^2\right) - n\left(\frac{\sigma_Z^2}{n} + \bar{y}_N^2\right)\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^n \left[\sum_{i=1}^n p_i Z_i^2 - \bar{y}_N^2\right] - \sigma_Z^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^n \left[\sum_{i=1}^n p_i (Z_i - \bar{y}_N)^2\right] - \sigma_Z^2\right]$$

$$\left| \sum^N p_i (Z_i \bar{y}_N)^2 = \sum^N p_i Z_i^2 + \bar{y}_N^2 - 2\bar{y}_N \bar{y}_N = \sum^N p_i \overline{Z}_i^2 - \bar{y}_N^2 \right.$$

$$= \frac{1}{n-1}\left[n\,\sigma_Z^2 - \sigma_Z^2\right] = \sigma_Z^2$$

Thus, s_Z^2 is an unbiased estimate of σ_Z^2.

It follows that an unbiased estimate of $V(Z_n)$ is given by

$$\text{Est. } V(\overline{Z}_n) = \frac{s_Z^2}{n} = \frac{1}{n(n-1)} \sum_{i-1}^n \left[\frac{y_i}{Np_i} - \frac{\hat{r}}{N}\right]^2 \qquad ...(11)$$

where $\hat{r} = N\,\overline{Z}_n = \frac{1}{n}\sum^n \frac{y_i}{p_i}$ is an unbiased estimate of the population total $N\bar{y}_N$.

Or

$$\text{Est. } V(\overline{Z}_n) = \frac{1}{n(n-1)N^2}\left[\sum_{i=1}^n \left(\frac{y_i}{p_i}\right)^2 - n\,\hat{r}^2\right] \qquad ...(12)$$

which is suitable from for computational purposes.

6.4 PPS SAMPLING WITHOUT REPLACEMENT

NOTES

We shall show that in sampling without replacement when initial probabilities of selection are unequal, the probability of drawing a specified unit of the population at a given draw changes with the draw.

Let p_i be the probability of selecting the i^{th} unit of the population at the first draw $i = 1, 2, ..., N$; $\sum_{i=1}^{N} p_i = 1$ and let p_{i_r} denote the probability of selecting y_i at the r^{th} draw ($r = 1, 2, ..., n$).

We consider sampling without replacement and assume that at any subsequent draw, the probability of selecting a unit from the units available at that draw is $\propto$ to the probability of selecting it at the first draw.

Clearly, $$p_{i_1} = p_i;\ i = 1, 2, ..., N \qquad ...(1)$$

and p_{i_2} = Probability [y_i is not selected at the first draw] × Probability [y_i is selected at the II[nd] draw/y_i is not selected at the I[st] draw.]

$$= \sum_{j \neq i=1}^{N} \text{Probability } [y_{j \neq i} \text{ is selected at the first draw}] \times$$

Probability [y_i is selected at the II[nd] draw/$y_{j \neq i}$ is selected at the I[st] draw.]

If $E_1, E_2, ..., E_n$ are mutually disjoint events and A is a subset of $\bigcup_{i=1}^{n} E_i$, then $P(A) = \sum_{i=1}^{n} P(E_i)\ P(A/E_i)$

$$p_{i_1} = \sum_{j \neq i=1}^{n} p_j \cdot \left(\frac{p_i}{1 - p_j}\right) \qquad \left| P\ (A/B) = \frac{P(A \cap B)}{P(B)} \right.$$

$$p_{i_2} = \left[\sum_{j=1}^{N} \frac{p_j}{1 - p_j} - \frac{p_i}{1 - p_i}\right] p_i$$

$$= \left(S - \frac{p_i}{1 - p_i}\right) p_i \qquad ...(2)$$

where, $$S = \sum_{j=1}^{N} \frac{p_i}{1 - p_j} \qquad ...(3)$$

Clearly, $p_{i_1} \neq p_{i_2}$ $\forall$ $i = 1, 2, . . ., N$ unless $p_i = \frac{1}{N}$. It follows that the expected value of the variate under consideration will change with successive draws.

NOTES

6.5 HORVITZ-THOMPSON ESTIMATE

Let the population consists of N units and y_i be the characteristic under study for the i^{th} unit in the population $i = 1, 2, \ldots, N$. Suppose a sample of size n is drawn without replacement, using aubitrary probabilities of selection at each draw. Thus prior to each succeeding draw there is defined a new probability different for the units available at that draw. The probability different at each draw may or may not depend upon the initial probabilities at the first draw.

Define, a random variable α_i $(i = 1, 2, \ldots, N)$ as follows :

$$\because \quad \alpha_i = \begin{cases} 1, \text{if } y_i \text{ is included in a sample of size } n \\ 0, \text{otherwise} \end{cases} \qquad \ldots(1)$$

and let $$z_i = \frac{ny_i}{NE(\alpha_i)} \qquad \ldots(2)$$

where we assume that every unit has a +ve probability of being included in the sample *i.e.*, $E(\alpha_i) > 0 \ \forall_i$.

Now, $$\bar{z}_n = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{N} \alpha_i z_i \qquad \ldots(3)$$

$$\therefore \quad E(\bar{z}_n) = \frac{1}{n}\sum_{i=1}^{N} z_i E(\alpha_i) = \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{NE(\alpha_i)} E(\alpha_i) = \bar{y}_N \qquad \ldots(4)$$

$\Rightarrow$ Simple Arithmetic Mean of z_i provides an unbiased estimate of the population mean $\bar{y}_N$.

To obtain sampling variable, we have

$$V(\bar{z}_n) = V(\bar{z}_n^{\,2}) - \bar{y}_N^{\,2} \qquad \ldots(5)$$

Now, $$E(\bar{z}_n^{\,2}) = \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i z_i\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i^{\,2} z_i^{\,2} + \sum_{i \neq j=1}^{N} \alpha_i \alpha_j z_i z_j\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} E(\alpha_i) z_i^{\,2} + \sum_{i \neq j=1}^{N} E(\alpha_i \alpha_j)\, z_i z_j\right] \qquad \ldots(6)$$

$$\because E(\alpha_i^{\,2}) = E(a_i)$$

Now, from (4), $$\bar{y}_N^{\,2} = \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i E(\alpha_i)\right]^2$$

NOTES

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N}(E(\alpha_i)^2 z_i^2 + \sum_{i\neq j=1}^{N} E(\alpha_i)E(\alpha_j)z_i z_j\right] \quad ...(7)$$

Using, (7) and (6) in (5), we get

$$V(\bar{z}_n) = \frac{1}{n^2}\left[\sum_{i=1}^{N} E(\alpha_i)\{1-E(\alpha_i)\}z_i^2 + \sum_{i\neq j=1}^{N}\{E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)\}z_i z_j\right] ...(8)$$

In terms of y, (8) becomes

$$V(\bar{z}_n) = \frac{1}{N^2}\left[\sum_{i=1}^{N}\left\{\frac{1-E(\alpha_i)}{E(\alpha_i)}\right\}y_i^2 + \sum_{i\neq j=1}^{N}\left\{\frac{E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)}{E(\alpha_i)E(\alpha_j)}\right\}y_i y_j\right] \quad ...(9)$$

An unbiased estimate of this variable takes the following form.

$$\text{Est. } V(\bar{z}_n) = \frac{1}{N^2}\left[\sum_{i=1}^{n}\frac{1-E(\alpha_i)}{(E(\alpha_i))^2}y_i^2 + \sum_{i\neq j=1}^{n}\frac{E(\alpha_i\alpha_j) - E(\alpha_i)\,E(\alpha_j)}{E(\alpha_i\alpha_j)E(\alpha_i)\,E(\alpha_j)}y_i y_j\right] \quad ...(10)$$

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

$$= \sum_{i=1}^{n}\frac{1-E(\alpha_i)}{(E(\alpha_i))^2}y_i^2 E(\alpha_i^2) + \sum_{i\neq j}^{N}\frac{E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_i\alpha_j)}{E(\alpha_i\alpha_j)\,E(\alpha_i)\,E(\alpha_j)}y_i y_j\,E(\alpha_i\alpha_j)$$

$$= \sum_{i=1}^{n}\frac{1-E(\alpha_i)}{E(\alpha_i)}y_i^2 + \sum_{i\neq j}^{N}\frac{E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)}{E(\alpha_i)\,E(\alpha_j)}y_i y_j \text{ as } E(\alpha_i^2) = E(\alpha_i)$$

There is one drawback of this estimate. It does not reduce to zero when all z_i are equal, a case for which the variance is necessarily zero. Consequently, this estimate may assume –ve values for same samples.

A more elegant expression for the variable of the estimate is given by **Yates and Grundy.** This can be obtained as follows :

Since there are exactly n values of α_i which are 1 and $(N - n)$ value which are zero, thus, we have

$$\sum_{i=1}^{N}\alpha_i = n \quad ...(11)$$

$$\Rightarrow \quad \sum_{i=1}^{N}E(\alpha_i) = n \quad ...(12)$$

Squaring (11), taking expectations again, using (12) and also the fact that $E(\alpha_i^2) = E(\alpha_i)$, we obtain

$$\sum_{i\neq j=1}^{N}E(\alpha_i\alpha_j) = n(n-1) \quad ...(13)$$

Also $\quad E(\alpha_i,\alpha_j) = P(\alpha_i = 1,\ \alpha_j = 1)$

NOTES

$$= P(\alpha_i = 1) \cdot P(\alpha_j = 1/\alpha_i = 1)$$
$$= E(\alpha_i) \cdot E(\alpha_j/\alpha_i = 1) \quad ...(14)$$

$$E(\alpha_i) = \sum_{i=1}^{n} \alpha_i p(\alpha_i) = \alpha_i\, p(\alpha_1) + \alpha\, p(\alpha_2) + \ldots + \alpha_n\, p(\alpha_n) +$$
$$= p(\alpha_1) + p(\alpha_2) + \ldots + p(\alpha_n) = p(\alpha_i = 1)$$

as there are n values of α_i which are equal to 1

$$\text{Thus, } \sum_{j(\neq i)=1}^{N}\left[E(\alpha_i\alpha_j) - E(\alpha_i)\,E(\alpha_j)\right] = E(\alpha_i)\left[\sum_{j(\neq i)=1}^{N}\left\{E(\alpha_j/\alpha_i = 1) - E(\alpha_j)\right\}\right]$$

$$= E(\alpha_i) \sum_{j(\neq i)=1}^{N}\left[E(\alpha_j/\alpha_i = 1) - E(\alpha_j)\right]$$

$$= E(\alpha_i)\,\left[(n-1) - (n - E(\alpha_i)\right]$$

$$\sum_{i\neq j-1}^{N} E(\alpha_i\alpha_j) = n(n-1), \quad \therefore \sum_{i\neq j-1}^{N} E(\alpha_i\alpha_j) = (n-1)$$

$$E(\alpha_j/\alpha_i) = E(\alpha_i/\alpha_j)\ (+\ \alpha_i) = \frac{n(n-1)}{n}$$

$$\text{Also, } \sum_{i=1}^{N} E(\alpha_i) = n$$

$$\Rightarrow \quad E(\alpha_i) + \sum_{j(\neq i)=1}^{N} E(\alpha_j) = n$$

$$\therefore \sum_{j(\neq i)=1}^{N}\left[E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)\right] = -E\,(\alpha_i)\,[1 - E(\alpha_i)] \quad ...(15)$$

Similarly,

$$\sum_{i(\neq j)=1}^{N}\left[E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)\right] = -E(\alpha_j)\,[1 - E(\alpha_j)] \quad ...(16)$$

Using (15) and (16), the expression for variable ($\bar{z}_n$) *i.e.,* (8) can be rewritten as

$$V(\bar{z}_n) = \frac{1}{2n^2}\left[\sum_{i=1}^{N} E(\alpha_i)\,(1 - E(\alpha_i)\,z_i^{\,2} + \sum_{j=1}^{N} E(\alpha_j)\,(1 - E(\alpha_j))\,z_j^{\,2}\right.$$
$$\left. -2\sum_{i\neq j=1}^{N}\left\{E(\alpha_i)\,E(\alpha_j) - E(\alpha_i\alpha_j)\right\}z_i z_j\right]$$

$$= \frac{1}{2n^2}\sum_{i\neq j=1}^{N}\left[E(\alpha_i)\,E(\alpha_j) - E(\alpha_i\alpha_j)\right]\left(z_i^{\,2} + z_j^{\,2} - 2\,z_i\,z_j\right)$$

NOTES

$$= \frac{1}{2n^2} \sum_{i \neq j=1}^{N} \left[E(\alpha_i)\, E(\alpha_j) - E(\alpha_i \alpha_j) \right] \left(z_i - z_j \right)^2 \qquad ...(17)$$

The expressions for $E(\alpha_i)$ and $E(\alpha_i\, \alpha_j)$ can be written explicitly for any given sample size. *e.g.*, let $n = 2$ and assume that at the II[nd] draw, the probability of selecting a unit from the units available is proportional to the probability of selecting it at the I[st] draw.

$\because$ $E(\alpha_i)$ = Probability of including y_j in a sample of two

$$= p_{i_1} p_{i_2}$$

where p_{ir} $(r = 1, 2)$ is the problem of selecting y_i at the s^{th} draw

$$\therefore \quad E(\alpha_i) = p_i \left[s + 1 = \frac{p_i}{l - p_i} \right] \qquad ...(18)$$

$$\left| \quad p_i = p_i\, p_{i_2} = \left(s - \frac{p_i}{l - p_i} \right) p_i \text{ where } s = \sum_{j-1}^{N} \frac{p_j}{1 - p_j} \right.$$

Again, $E(\alpha_i \alpha_j)$ = Problem of including both y_i and y_j in a sample of two

$$= p_{i_1},\ p_{j_2/i} + p_{ji} p_{i_2/J}$$

$$\therefore \quad E(\alpha_i \alpha_j) = p_i \frac{p_j}{1 - p_i} + p_j \frac{p_i}{1 - p_j}$$

$$= p_i\, p_j \left(\frac{1}{1 - p_i} + \frac{1}{1 - p_j} \right) \qquad ...(19)$$

Similar, expressions can be obtained for any sample size.

From (17), it follows that are unbiased estimate of $V(\bar{z}_n)$ is given by

$$\text{Est. } \left[V(\bar{z}_n) \right] = \frac{1}{2n^2} \sum_{i \neq j=1}^{n} \frac{E(\alpha_i) \times E(\alpha_j) - E(\alpha_i\, \alpha_j)}{E(\alpha_i\, \alpha_j)} (z_i - z_j)^2 \qquad ...(20)$$

which is a linear function of the squares of the differences between the z_i in the sample and vanishes when they are equal.

Now, we describe some systems of sampling which have the desirable property that they lead to more efficient estimates than in the case of sampling with replacement.

6.6 MIDZUNO SYSTEM OF SAMPLING

Under this system of selection probabilities, due to Midzuno, the unit at the first draw is selected with unequal probabilities of selection while, at all subsequent draws, the units are selected with equal probability and without replacement.

NOTES

Define a random variable α_i (i = 1, 2, . . . , N) as follows :

$$\alpha_i = \begin{cases} 1, \text{ if } y_i \text{ is included in a sample of size } n \\ 0, \text{ otherwise} \end{cases}$$

Now, $E(\alpha_i) = p_{i_1} + p_{i_2} + \ldots + p_{i_n}$

$= p_i$ + probability that y_i is not selected at the first draw and is selected at any of the subsequent $(n-1)$ draws.

$$= p_i + (1 - p_i)\frac{n-1}{N-1}.$$

$$= \frac{N-n}{N-1}p_i + \frac{n-1}{N-1} \qquad \ldots(21)$$

$E(\alpha_i\alpha_j)$ = Probability that y_i and y_j are both in the sample

= Probability that y_i is selected at the first draw and y_j is selected at any of the subsequent $(n-1)$ draws.

+ Probability that y_j is selected at the first draw and y_i is selected at any of the subsequent $(n-1)$ draws.

+ Probability that neither y_i nor y_j is selected at the first draw but both of them are selected during the subsequent $(n-1)$ draws.

$$= p_i\frac{n-1}{N-1} + p_j\frac{n-1}{N-1} + (1-p_i-p_j)\frac{(n-1)}{(N-1)}\frac{(n-2)}{(N-2)}$$

$$= \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i+p_j) + \frac{n-2}{N-2}\right] \qquad \ldots(22)$$

Similarly, $E(\alpha_i\alpha_j\alpha_k)$ = Probability of including y_i, y_j and y_k in the sample

$$= \frac{(n-1)(n-2)}{(N-1)(N-2)}\left[\frac{(N-n)}{(N-3)}(p_i+p_j+p_k) + \frac{n-3}{N-3}\right] \qquad \ldots(23)$$

By an extension of the above argument, we see that if y_i, y_j, y_k, ..., y_q be the n units, the probability of including there n units is the sample is given by

$$E(\alpha_i\,\alpha_j\,\alpha_k, \ldots, \alpha_q) = \frac{(n-1)(n-2)\ldots(n-\overline{n-1})}{(n-1)(n-2)\ldots(N-\overline{n-1})}$$

$$\left[\frac{N-n}{N-n}\left(p_i+p_j+p_k\ldots+p_q\right) + \frac{N-n}{N-n}\right]$$

NOTES

$$= \frac{1}{{}^{N-1}C_{n-1}}\left[P_i + P_j + P_c \ldots + P_q\right] \qquad \ldots(24)$$

$$\left(\therefore {}^{N-1}C_{n-1} = \frac{(N-1)(N-2)\ldots\overline{(N-n-1)}(N-n)!}{(n-1)(n-2)\ldots\overline{(n-n-1)}\lfloor 0\,(N-n)!}\right)$$

Thus, if p_i are considered proportional to some measure of size of the units in the population, we see that the probability of selecting a specified sample proportional to the total measure of size of the units included in the sample.

Substituting from (21) and (22) in (4) and (17), we obtain the Horvitz Thompson estimate of the population mean and its variance.

In particular, it may be noted that the unbiased estimate of variable of Horvitz Thompson estimate given by (20) reduces to,

$$\frac{N-n}{2(N-1)^2 n^2}\sum_{i\neq j}^{n}\left[(N-n)\,p_i p_j + \frac{n-1}{N-2}(1-p_i-p_j)\right]\frac{(z_i - z_j)^2}{E(\alpha_i \alpha_j)} \qquad \ldots(25)$$

We have from (21) and (22),

$E(\alpha_i)\,E(\alpha_j) - E(\alpha_i\,\alpha_j)$

$$= \left\{p_i + (1-p_i)\frac{n-1}{N-1}\right\}\left\{p_j + (1-p_j)\frac{n-1}{N-1}\right\}$$

$$- p_i\frac{n-1}{N-1} - p_j\frac{n-1}{N-1} - (1-p_i-p_j)\frac{n-1}{N-1}\frac{n-2}{N-2}$$

$$= p_i p_j + \frac{n-1}{N-1}(p_j - p_i\,p_j) + \frac{n-1}{N-1}(p_i - p_i p_j) + \left(\frac{n-1}{N-1}\right)^2(1-p_i)\,(1-p_j)$$

$$- p_i\frac{n-1}{N-1} - p_j\frac{n-1}{N-1} - (1-p_i-p_j)\frac{n-1}{N-1}\frac{n-2}{N-2}$$

$$= p_i p_j\left\{1 - 2\frac{n-1}{N-1} + \left(\frac{n-1}{N-1}\right)^2\right\} + (1-p_i-p_j)\left\{\left(\frac{n-1}{N-1}\right)^2 - \frac{n-1}{N-1}\frac{n-2}{N-2}\right\}$$

$$= p_i p_j\,\frac{(N-1)^2 - 2(n-1)\,(N-1) + (n-1)^2}{(N-1)^2}$$

$$+ \frac{(1-p_i-p_j)}{(N-1)^2}\left\{(n-1)^2 - (n-1)\,(n-2)\,\frac{N-1}{N-2}\right\}$$

NOTES

$$= \frac{1}{(N-1)^2}\left[p_i p_j\left\{N^2+1-2N-2nN+2N+2n-2+n^2+1-2n\right\}\right.$$

$$\left.+(1-p_i-p_j)\left\{\frac{(n^2+1-2n)(N-2)-(n^2-3n+2)(N-1)}{N-2}\right\}\right]$$

$$= \frac{1}{(N-1)^2}\left[p_i p_j (N-n)^2+(1-p_i-p_j)\frac{(n-1)(N-n)}{N-2}\right]$$

$$= \frac{(N-n)}{(N-1)^2}\left[p_i p_j (N-n)+(1-p_i-p_j)\frac{n-1}{N-2}\right]$$

Using in (20), we get (25), which is always +ve, except when the z_i are equal in which case it is zero. Thus, under Midzuno scheme of sampling, the Yates-Grundy estimate of the variance of Horvitz-Thompson estimate is never negative.

The main advantage of this method of sampling is that it is possible to compute a set of revised probabilities of selection p_i'. It means the inclusion probabilities $E(\alpha_i)$ resulting from the revised probabilities p_i' are proportional to the initial probabilities of selection p_i (i = 1, 2, . . ., N).

Now, if p_i' are revised probabilities of selection, we have from (21),

$$E(\alpha_i) = \frac{N-n}{N-1}p_i' + \frac{n-1}{N-1}$$

If this is proportional to the initial probability of selection p_i, we obtain by utilizing the fact that $\sum_{i=1}^{N} E(\alpha_i) = n$,

$$\frac{N-n}{N-1}p_i' + \frac{n-1}{N-1} = np_i$$

$$\sum_{i=1}^{N}\left(\frac{N-n}{N-1}P_i' + \frac{n-1}{N-1}\right) = \sum_{i=1}^{N} np_i = n\sum_{i=1}^{N} p_i = n$$

i.e.,

$$\sum_{i=1}^{N} E(\alpha_i) = n$$

Hence,

$$p_i' = \left(nP_i - \frac{n-1}{N-1}\right)\frac{N-1}{N-n}$$

Since p_i' $(i = 1, 2, \ldots, N)$ must always be +ve, the initial probabilities of selection p_i must satisfy the condition.

$$p_i > \frac{n-1}{n(N-1)}$$

NOTES

Subject to this restriction, it has been shown by Rao that for the sample of size 2, The Horvitz-Thompson estimate of population mean under Midzuno scheme of sampling is always more efficient than the usual estimate in the case of sampling with varying probabilities and with replacement.

6.7 NARAIN METHOD OF SAMPLING

This method does not require any restriction on the initial set of selection probabilities and at the same time leads to a more efficient estimate of the population value than in the case of sampling with replacement. The method consists in constructing revised probabilities of selection p_i^* $(i = 1, 2, \ldots, N)$. The inclusion probabilities $E(\alpha_i)$ resulting from the p_i^* are proportional to the original probabilities of selection p_i. The sampling is done without replacement and the probabilities of selection at the II[nd] and subsequent draws are proportional to the revised probabilities of selection p_i^*at the first draw.

Thus, we have from (18),

$$E(\alpha_i) = p_i^* \left[s^* + 1 - \frac{p_i^*}{1 - p_i^*} \right] = np_i \qquad \ldots(1)$$

where,

$$S^* = \sum_{i=1}^{N} \frac{p_i^*}{1 - p_i^*}. \qquad \ldots(2)$$

Methods of solving the set of equation (1) to obtain the revised probabilities of selection p_i^* have been given by Narain and Yates and Grundy. The computations are tedious for $n > 2$ and the method eventually breaks down.

STUDENT ACTIVITY 1

1. Write in details the procedure of selecting a sample with varying probability.

STUDENT ACTIVITY 2

1. Mention the difference between sampling with replacement and sampling without replacement with varying probabilities.

NOTES

6.8 ILLUSTRATIVE EXAMPLES

Example 1. *A village has 10 holdings consisting of 50, 30, 45, 25, 40, 26, 24, 35, 28 and 27 fields, respectively. Select a sample of four holdings with the replacement method and with probability proportional to the number of fields in the holding.*

Solution. The first step in the selection of holdings is to form cumulative totals as shown in the following Table :

S.No.	*Size* (X_i)	*Cumulative size*	*Numbers associated*
1	50	50	1–50
2	30	80	51–80
3	45	125	81–125
4	25	150	126–150
5	40	190	151–190
6	26	216	191–216
7	44	260	217–260
8	35	295	261–295
9	28	323	296–323
10	27	350	324–350

For selecting a holding, a random number not exceeding 350 is drawn with the help of a random number table. Let the random number thus selected is 272. It can be seen from the cumulative totals that the number is associated with the group 261–295, *i.e.*, the 8^{th} holding is selected corresponding to the random number 272. Similarly, we can select three more random numbers. Let these numbers are 346, 165 and 094. Then the holdings selected corresponding to these random numbers are the 10^{th}, 5^{th} and 3^{rd}, respectively. Hence, a sample of 4 holdings selected with probability proportional to size will contain the 8^{th}, 10^{th}, 5^{th} and 3^{rd} holdings.

Example 2. *If the yields (in 10 kg) of the 8 orchards in Example. 2, are 60, 35, 30, 44, 30, 50, 22 and 40 respectively, estimate, along with the standard error, the total production of the 8 orchards, from samples selected in the example, by using Horvitz-Thompson estimator.*

Solution. Since the sample selected in Example. 2, includes the orchards at serial numbers 5 and 7, the yield (in 10 kg) of two selected orchards are 30 and 22, respectively.

We have the following results from example 2.

S.No.	No. of trees (X_i)	Yield (Y_i)	$p_i = (X_i/X)$
1	50	60	0.185
2	30	35	0.111
3	25	30	0.093
4	40	44	0.148
5	26	30	0.096
6	44	50	0.163
7	20	22	0.074
8	35	40	0.130
Total	290	311	1.000

For the sample selected, we have $x_1 = 26$, $x_2 = 20$, $y_1 = 30$, $y_2 = 22$, $p_1 = 0.096$ and $p_2 = 0.074$.

The Horvitz-Thompson estimator for population total is given by

$$\hat{Y}_{HT} = \sum_{i}^{n} \frac{y_i}{\pi_i}$$

with $$v\left(\hat{Y}_{HT}\right) = \sum_{i}^{n} \sum_{j>i}^{n} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Now, we have $$S = \sum_{i}^{8} \frac{p_i}{1 - p_i} = 1.157$$

For the selected units,

$$\pi_1 = p_1 \left[S + 1 - \frac{p_1}{1 - p_1}\right] = 0.1969$$

$$\pi_2 = p_2 \left[S + 1 - \frac{p_2}{1 - p_2}\right] = 0.1538$$

$$\pi_1 = p_1 p_2 \left[\frac{1}{1 - p_1} + \frac{1}{1 - p_2}\right] = 0.0155$$

Thus, $$\hat{Y}_{HT} = \frac{30}{0.1969} + \frac{22}{0.1538} = 295.403 \text{ (in 10 kg units)}$$

and $$v\left(\hat{Y}_{HT}\right) = \left[\frac{0.1969 \times 0.1538 - 0.0155}{0.0155}\right]\left[\frac{30}{0.1969} - \frac{22}{0.1538}\right]^2$$

$$= 79.91$$

$\therefore$ Standard error of $\hat{Y}_{HT} = \sqrt{v\left(\hat{Y}_{HT}\right)} = \sqrt{79.91} = 8.93$

NOTES

Example 3. *From the population of 8 orchards in Example 2 select a sample of 2 orchards using :*

(i) Randomised probability proportional to size sampling systematic sampling

(ii) Random group method and

(iii) Probability proportional to size

and estimate the average production along with their respective standard errors.

Solution. (*i*) Randomised probability proportional to size sampling **Systematic** Sampling.

For selecting a sample of size 2 using randomised probability proportional to size sampling systematic sampling, we arrange the population units at random and the arrangement thus obtained is :

S.No.	*No. of trees* (X_i)	*Cumulative total*
1	30	30
2	50	80
3	26	106
4	25	131
5	40	171
6	20	191
7	44	235
8	35	270

The value of k, the sampling interval in the above example, is

$$\frac{\sum_{i}^{N} X_i}{n} = \frac{270}{2} = 135$$

Selecting a random number between 1 and 135, we get 120. Another random number is then obtained by adding 135 to the random number selected. The units corresponding to the cumulative totals 120 and 255 are at serial numbers 4 and 8, respectively. Thus, the orchards selected in the sample are at serial numbers 4 and 8 in the original table with the number of trees being 25 and 35, respectively.

The estimate and the estimated variance of the estimate of average production, are

$$\hat{\bar{Y}} = \frac{1}{nN}\left[\frac{y_1}{p_1} + \frac{y_2}{p_2}\right]$$

and $$v\left(\hat{\bar{Y}}\right) = \frac{1}{N^2 n^2 (n-1)}\left[1 - n(p_1 + p_2) + n\sum_i^N p_i^2\right]\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

In our case,

$$N = 8,\ y_1 = 30,\ x_1 = 25,\ p_1 = \frac{25}{270} = 0.093$$

$$n = 2,\ y_2 = 40,\ x_2 = 35,\ p_2 = \frac{35}{270} = 0.130$$

Thus, $$\hat{\bar{Y}} = \frac{1}{16}\left[\frac{30}{0.093} + \frac{40}{0.130}\right] = 39.39$$

and $$v\left(\hat{\bar{Y}}\right) = \frac{1}{4 \times 1}\ [1 - 2\ (0.093 + 0.130) + 2 \times 0.13526]$$

$$\times\left(\frac{30}{8 \times 0.093} - \frac{40}{8 \times 0.130}\right)^2 = 0.1783$$

$\therefore$ Standard error of $\hat{\bar{Y}} = \sqrt{v\left(\hat{\bar{Y}}\right)} = \sqrt{0.1783} = 0.4224$

(*ii*) Random Group Method

As the units are randomised in (*i*), we take the first 4 units as forming group Ist and the remaining 4 units as group IInd, respectively. Now, from each group, one unit is to be selected by probability proportional to the number of trees. For this, let us prepare the following table :

	Group Ist			*Group IInd*	
S.No.	*No. of trees*	*Probability*	*S.No.*	*No. of trees*	*Probability*
1	30	0.111	5	40	0.148
2	50	0.185	6	20	0.074
3	26	0.096	7	44	0.163
4	25	0.093	8	35	0.130
Total	131	0.485	Total	139	0.515

Using Lahiri's method of selection and referring to the table of random numbers, the admissible pairs selected for group Ist and group IInd are (2, 41) and (3, 38). Hence, the units selected are 1 and 6 in the original table.

The Rao, Hartley and Cochran estimate and its estimate of the variance are

$$\hat{\bar{Y}} = \frac{1}{N}\left[\frac{y_1}{p_1/\pi_1} + \frac{y_2}{p_2/\pi_2}\right]$$

NOTES

and $$v\left(\hat{\bar{Y}}\right) = \left(1-\frac{2}{N}\right)\left[\pi_1\left(\frac{y_1}{Np_1}-\hat{\bar{Y}}\right)^2+\pi_2\left(\frac{y_2}{Np_2}-\hat{\bar{Y}}\right)^2\right]$$

For the units selected, we have

$$N = 8,\quad y_1 = 60,\quad x_1 = 50,\quad p_1 = \frac{50}{270} = 0.185$$

$$n = 2,\quad y_2 = 50,\quad x_2 = 44,\quad p_2 = \frac{44}{270} = 0.163$$

$$\pi_1 = (p_1 + p_2 + p_3 + p_4)$$
$$= 0.185 + 0.111 + 0.093 + 0.096 = 0.485$$

$$\pi_2 = (p_5 + p_6 + p_7 + p_8)$$
$$= 0.148 + 0.163 + 0.074 + 0.130 = 0.515$$

Thus, $$\hat{\bar{Y}} = \frac{1}{8}\left[\frac{60\times 0.485}{0.185}+\frac{54\times 0.515}{0.163}\right] = 39.41$$

and $$v\left(\hat{\bar{Y}}\right) = (1-2/8)\left[0.485\left(\frac{60}{8\times 0.185}-39.41\right)^2\right.$$
$$\left.+0.515\left(\frac{50}{8\times 0.163}-39.41\right)^2\right]$$
$$= 0.0802$$

$\therefore$ Standard error of $\hat{\bar{Y}} = 0.2832$

(*iii*) **Probability Proportional to Total Size** *(Midzuno Sampling Procedure)*

In this method, the first unit is selected by probability proportional to the number of trees and the other by simple random sampling, without replacement, from the remaining units. Using Lahiri's method of selection and referring to the random number table, the pair selected is (6, 18). As the size of the 6th unit is greater than the second number of the random pair, *i.e.*, 18, the unit at serial number 6 is selected. The remaining units are now numbered from 1 to 7 and, thereafter, one random number from 1 to 7 is drawn to select the second unit in the sample. The random number from the table happens to be 6 and, therefore, the unit at serial number 6 in the new arrangement is selected. Thus, the units selected by this method are at serial numbers, 6 and 7 in the original table.

The Horvitz-Thompson estimator for average and the Yates-Grundy estimate of the variance for the estimate under this selection procedure are

$$\hat{\bar{Y}} = \frac{1}{N}\left[\frac{y_1}{\pi_1}+\frac{y_2}{\pi_2}\right]$$

and $$v\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\left(\frac{\pi_1\pi_2 - \pi_{12}}{\pi_{12}}\right)\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2$$

We have, $N = 8,\ y_1 = 50,\ x_1 = 44$

$n = 2,\ y_2 = 22,\ x_2 = 20,$

$$\pi_1 = \frac{8-2}{8-1} \times 0.163 + \frac{2-1}{8-1} = 0.282$$

$$\pi_2 = \frac{8-2}{8-1} \times 0.074 + \frac{2-1}{8-1} = 0.206$$

$$\pi_{12} = \frac{2-1}{8-1}\left[\frac{(8-2)}{(8-2)}(0.163 + 0.074) + \frac{(2-2)}{(8-2)}\right] = .033$$

Therefore, $$\hat{\bar{Y}} = \frac{1}{8}\ (50/0.282 + 22/0.206] = 35.51$$

and $$v\left(\hat{\bar{Y}}\right) = \frac{1}{64}\ \frac{(0.282 \times 0.206 - 0.033)}{0.033}\ (50/0.282 - 22/0.206)^2$$

$$= 58.2596$$

$\therefore$ Estimated standard error of $\hat{\bar{Y}} = \sqrt{58.2596} = 7.6328$

For this selection procedure, the expressions for inclusion probabilities are given by

6.9 SUMMARY

- A sampling procedure in which the units are selected with probabilities proportional to some measure of their size is known as p.p.s. sampling.
- When the units vary in size and the variable under study is correlated with size, the probabilities of selection can be assigned in proportion to the size of the unit.
- Under Midzyno scheme of sampling, the Yates-Grundy estimate of the variance of Hurwitz-Thompson estimate is never negative.
- The Hurwitz-Thompson estimate of population mean under Midzuno scheme of sampling is more efficient as in case of sampling with varying probabilities and with replacement.

6.10 GLOSSARY

- **Sampling with problem proportional to size.** It is also known as p.p.s. sampling.

NOTES

6.11 REVIEW QUESTIONS

1. From a population of N units, a sample of two distinct units is drawn. The first unit of the sample with probabilities p_i, $(i = 1, ..., N)$, $\sum_i p_i = 1$, based on measures of size X_i. After dropping the unit already selected, the second unit is selected with probability proportionate to q_i, $\sum q_i = 1$, where q_i are obtained by

$$p_i = \sum_{j \neq i} p_j q_i/(1 - q_i), \quad \text{for} \quad i = 1, 2, ..., N$$

The probability of the unit U_i, included in the sample, equals $2\ p_i$. Prove that the variance of the estimator $\sum (y_i/2p_i)$ of this sampling scheme is less than or equal to the variance of the same estimator in the case of sampling, *wr*.

2. In sampling with unequal probabilities, wor, a sample of size 2 is drawn. The first unit is selected with probability proportional to size sampling and the second with probability proportional to size sampling of the remaining units. Show that Yates and Grundy's variance estimator is always positive for this sampling system.

3. If the first unit in the sample is selected with probability proportional to size, the second with probability proportional to size of the remaining units, while the remaining $(n - 2)$ units are selected with equal probabilities without replacement. Prove that Yates and Grundy's variance estimator would be positive for this sampling system.

4. The results of a sample survey on the number of bearing lime trees and the area reported under lime, in each of the twenty-two villages growing lime in one of the tehsils of Nellore district, are given below :

S. No. of village	*Area under lime (in acres)*	*No. of bearing lime trees*
1	32.77	2328
2	7.97	754
3	0.62	105
4	15.61	949
5	42.85	3091
6	40.03	1736
7	9.39	840

8	6.33	311
9	5.05	0
10	94.55	3044
11	53.71	2483
12	0.67	128
13	0.82	102
14	2.15	60
15	0.43	0
16	123.36	11799
17	0.29	26
18	3.00	317
19	4.00	190
20	2.00	180
21	6.21	752
22	45.85	3091

From this population,

(*i*) select five samples of 10 villages each with probabilities proportional to area under lime, with replacement by Lahiri's method.

(*ii*) from one of these samples, estimate the total number of bearing lime trees in the above tehsil together with its standard error.

(*iii*) compare the efficiency of probability proportional to size sampling wr with that of simple random sampling.

5. A pilot scheme for the study of cultivation practices and yield of guava was carried out in Uttar Pradesh during 1980-81. The villages were selected with probability proportional to area under fresh fruit. The data concerning the selected villages from the selected tehsil are given below :

S. No. of the selected village	*Area under fresh fruit (in acres)*	*Area under guava crop (in acres)*
1	127.00	166.15
2	32.00	24.73
3	75.00	100.77
4	57.00	87.14
5	68.00	116.28
6	61.00	60.22

NOTES

7	6.00	13.59
8	27.00	41.70
9	68.00	10.52
10	13.17	13.85
11	8.00	12.92
12	8.00	10.73
13	22.00	38.64
14	10.00	15.92
15	30.00	9.09
16	99.00	155.51
17	7.38	10.34
18	63.00	95.16
19	13.00	22.40
20	14.00	10.97
21	30.00	39.07
22	10.00	13.70
23	13.00	26.40
24	7.00	1.50
25	8.00	12.57
26	24.00	2.00
27	3.00	6.72
28	24.00	20.75
29	38.00	51.65
30	13.00	16.42
31	4.00	3.90
32	12.00	2.44
33	1.00	3.90
34	5.00	15.31
35	1.00	1.44
36	13.00	14.88
37	19.00	23.01
38	5.00	3.44
39	12.00	14.32
40	12.00	24.39
41	7.38	9.88
42	7.00	17.66
43	2.00	3.26
44	8.00	6.89
45	1.00	0.84
46	10.00	13.02
47	22.00	32.85

(*i*) Considering the above 47 villages as constituting the population, divide it into three classes, one having area under fresh fruit less than or equal to 25 acres, the second between 25 and 60 acres and the third over 60 acres. From each of these classes, select a sample of two villages with probability proportional to the area under fresh fruit and without replacement, using the random numbers in the order given below :

9978 20873 15319 5736 19978 28865

(*ii*) Estimate the total area under guava and its standard error using

(*a*) Horvitz-Thompson estimator

(*b*) Des Raj ordered estimator

(*c*) Unordered Murthy estimator

(*iii*) Divide the villages in each class into two equal groups at random and select one villages from each group with probability proportional to the area under fresh fruits, with the help of the random numbers given above. Estimate the total area under guava with its standard error.

6. The data given below relate to 15 villages of one stratum, collected in a survey conducted in Andhra Pradesh for estimating the number of cashewnut trees. The area under cashewnut in different villages, as reported by Patwaris, and the actual number of cashewnut trees in different villages are given in Columns (2) and (3) of Table given below :

(1)	*(2)*	*(3)*
1	71.45	2614
2	46.28	1927
3	116.27	5162
4	178.11	6248
5	276.98	10842
6	22.72	1045
7	67.34	2359
8	26.21	1228
9	82.85	3107
10	0.36	1082
11	97.16	3247
12	178.16	8039
13	17.83	1013
14	521.46	19489
15	17.34	557

NOTES

(*i*) Draw a probability proportional to size sample of size 2 without replacement from this stratum and obtain an unbiased estimate of the total number of trees and the variance of the estimate by adopting Horvitz-Thompson estimator.

(*ii*) Divide the stratum into two groups at random and select one village from each group with probability to area under cashewnut trees. Estimate the total number of trees with its standard error. You may use the random number 20153 and 44227 for selecting units.

7. Data on the number of boats landing (x) and the catch of fish (y) at a particular centre in the Malabar Coast of India on a particular day of 12 hours (6:00 am to 6:00 pm) are given below :

Hours	1	2	3	4	5	6	7	8	9	10	11
x (in number)	42	52	19	6	23	56	36	59	14	14	26
y (in maunds)	568	887	223	88	352	1295	934	1265	486	443	980

Select a sample of size 3 with probability proportional to the number of boats landing, by the following methods :

(*i*) Hartley-Rao's probability proportional to size systematic sampling

(*ii*) Rao, Hartley and Cochran's random group method (RHS method)

From the selected samples, estimate the total catch of fish in a day along with respective standard errors.

8. To obtain an idea about the relative performance of various sampling procedures for samples of size 2, two populations of size $N = 4$ were studied by Yates and Grunds, and Des Raj, which are given below :

S. No. of the unit	p_i	*Population A* (y_i)	*Population B* (y_i)
1	0.1	0.5	0.8
2	0.2	1.2	1.4
3	0.3	2.1	1.8
4	0.4	3.2	2.0

Obtain the sampling variances of the following for sample size 2 :

(*i*) Probability proportional to size sampling sampling with replacement

(*ii*) Ordered Des Raj estimator

(*iii*) Unordered Murthy estimator

(*iv*) Horvitz-Thompson estimator

(*v*) Midzuno system of sampling

(*vi*) Probability proportional to size sampling systematic sampling

(*vii*) Rao, Hartley and Cochran method

Compare the efficiency of each method with that of sampling with replacement.

9. A population consists of 7 units of sizes 10, 20, 30, 40, 50, 60, and 70. A sample of 2 units is to be drawn by probability proportional to size wor. Find the probabilities of inclusion in the sample for (*i*) each unit and (*ii*) each pair of units.

10. A sample of 10 villages was drawn from a tehsil with probability proportional to size wr, size being 1951 census population (X = 415, 149).

The relevant data are given below :

Village	*1951 census population (x)*	*Cultivated area (y) in acres*	*Village*	*1951 census population (x)*	*Cultivated area (y) in acres*
1	5511	4824	6	7357	5505
2	865	924	7	5131	4051
3	2535	1948	8	4654	4060
4	3523	3013	9	1146	809
5	8368	7678	10	1165	1014

(*i*) Estimate the total cultivated area and its standard error.

(*ii*) Obtain the sample size required to ensure the relative standard error of 2 percent.

11. Write a short note on

(*a*) Hurwitz-Thompson Estimate

(*b*) Midzuno Scheme of Sampling

(*c*) Narain Scheme of Sampling

6.12 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

NOTES

CHAPTER 7 TWO STAGE SAMPLING

OBJECTIVES

After going through this chapter, we should be able to :

- know about two stage sampling, multi-stage sampling.
- estimation of the variance of the sample mean in case of two stage sampling.
- about comparison of two stages sampling and one stage sampling.

STRUCTURE

NOTES

7.1 INTRODUCTION

We have, seen that the larger the cluster, the less efficient it will be. It is thus logical to expect that for a given number of elements, greater precision will be attained by distributing them over a large number of clusters and then sampling a larger number of elements from each of them or completely enumerating them.

The procedure of first selecting clusters and then choosing a specified number of elements from each selected cluster is known as **sub-sampling** or **two-stage sampling**. The clusters which form the units of sampling at the first stage are called the **first stage units** and the elements groups of elements within clusters which form the units of sampling at the second stage are called **sub-units** or **second stage units.** The procedure can be easily generalized to three or more stages and is termed as **multi-stage sampling.**

7.2 TWO STAGE SAMPLING, EQUAL FIRST STAGE UNITS: ESTIMATE OF THE POPULATION MEAN AND ITS VARIANCE

We first consider the case of equal clusters and assure that the population is composed of NM elements grouped into N first stage units of M second stage units each. Let n denote the number of first stage units in the sample and m be number of second stage units to be drawn from each selected first stage units. Further we assume that units at each stage are selected with equal probability.

Now let,

y_{ij} : the value of the j^{th} second stage unit in the i^{th} first stage unit $J = 1, 2, ..., m$, $i = 1, 2, ..., n$.

$\bar{y}_{i*}$ = the mean per second stage unit in the i^{th} first stage unit in the population ; $i = 1, 2, ..., n$.

$\bar{y}_{**} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}$ = the mean per second stage unit in the population

$\bar{y}_{im} = \frac{1}{m} \sum_{j=1}^{m} y_{ij}$ = the mean per second stage unit of the i^{th} first stage unit in the sample.

and

$\bar{y}_{nm} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{im}$ = the mean per second stage unit in the sample.

NOTES

We shall now show that $\bar{y}_{nm}$ is an unbiased estimate of the population mean $\bar{y}_{**}$ and then proceed to derive its variance.

Since, the sample is selected in two stages, the expected value is also appropriately worked out in two stages, first overall possible sample of m from each of n fixed first stage units and then overall possible samples of n. Thus, we have.

$$E(\bar{y}_{nm}) = E\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{im}\right] E\left[\frac{1}{N}\sum_{im}^{n}E(\bar{y}_{im}/i)\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{im}\right] = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i*} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{M}\sum_{j=1}^{M}y_{ij}\right] = \bar{y}_{**} \quad ...(1)$$

Thus, sample mean of all the elements in the sample gives an unbiased estimate of the population mean.

Now, to obtain the variable of the estimate, we have

$$V(\bar{y}_{nm}) = V[E(\bar{y}_{nm}/n)] + E[V(\bar{y}_{nm}/n)]$$

$$= V\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i*}\right) + E\left[\frac{1}{n^2}\sum_{i=1}^{n}V(\bar{y}_{im}/i)\right]$$

$$= V\left[\frac{1}{n}\sum^{n}\bar{y}_{i*}\right] + E\left[\frac{1}{n^2}\sum^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2\right]$$

$$= V\left(\frac{1}{n}\sum^{n}\bar{y}_{i*}\right) + \frac{1}{n}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^{\,2} + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_{\omega}^{\,2} \qquad \left|\ \bar{S}_w^{\,2} = \frac{1}{N}\Sigma S_i^2\right.$$

...(2)

where, $S_b^{\,2} = \dfrac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i*}-\bar{y}_{**})^2$

and $\bar{S}_w^{\,2} = \dfrac{1}{N}\sum_{i=1}^{N}S_i^{\,2} = \dfrac{1}{N}\sum_{i=1}^{N}\dfrac{1}{M-1}(y_{ij}-\bar{y}_{i*})^2$

The variance (2) is made up of two compounents. If the selected first stage units had been completely enumerated *i.e.*, m = M, the variance of the sample mean would be the first component only. Actually, each selected first stage unit is only partially enumerated by means of a sample of second stage units drawn from it. The second term thus represents the contribution to the sampling variance arising from sub–sampling the selected first stage units.

NOTES

Now when $n = N$, *i.e.*, every first stage unit in the population is sampled, we are left with the second stage component. This case corresponds to stratified sampling with first stage units as strata and simple rendom sample of m drawn from each of several strata. We can thus look upon a sub-sampling design as a case of incomplete stratified, the first component representing the additional contribution to the variance of the estimate of mean arising from incomplete stratification.

When N is large relative to n, so that $\frac{N-n}{N}$ can be taken

$$V(\bar{y}_{nm}) = \frac{S_b^2}{n} + \left(\frac{1}{m} - \frac{1}{M}\right)\frac{\bar{S}_w^2}{n} \qquad ...(3)$$

When M is large relative to m, then $\frac{M-m}{M} = 1$, then

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \frac{\bar{S}_w^2}{nm} \qquad ...(4)$$

And when finite multipliers at both stages can be taken as unity,

than $$V(\bar{y}_{nm}) = \frac{S_b^2}{n} + \frac{\bar{S}_w^2}{nm} \qquad ...(5)$$

7.3 TWO STAGE SAMPLING, EQUAL FIRST STAGE UNITS : ESTIMATION OF THE VARIANCE OF THE SAMPLE MEAN

The estimation of variance of the sample mean in two stage sampling involves the estimates of S_b^2 and $\bar{S}_w^2$.

Let the mean square between first stage unit means in the sample be

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{im} - \bar{y}_{nm})^2 \qquad ...(1)$$

Also let s_i^2 denote the mean square between second stage units drawn from the i^{th} first stage units in the sample defined by

$$s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{im})^2 \qquad ...(2)$$

From (1),

$$(n-1)\,s_b^2 = \sum_{i=1}^{n}\bar{y}_{im}^{\,2} - n\,\bar{y}_{nm}^{\,2}$$

NOTES

$$\therefore \quad (n-1)\,E\,(s_b^{\,2}) = E\left(\sum_{i=1}^{n} \bar{y}_{im}^{\,2}\right) - n\,E\,(\bar{y}_{nm}^{\,2}) \qquad \ldots(3)$$

Now $$E\left(\sum_{i=1}^{n} \bar{y}_{im}^{\,2}\right) = E\left[\sum_{i=1}^{n} E(\bar{y}_{im}^{\,2}/i)\right]$$

$$= E\left[\sum_{i=1}^{n}\left\{\bar{y}_{i*}^{\,2} + \left(\frac{1}{m} - \frac{1}{M}\right) S_i^{\,2}\right\}\right]$$

$$V(\bar{y}_{im}) = E(\bar{y}_{im}^{\,2}) - \left[E(\bar{y}_{im})\right]^2$$

$$i.e.,\ E(\bar{y}_{im}^{\,2}) = V(\bar{y}_{im}) + \left(E(\bar{y}_{im})\right)^2$$

$$\therefore \quad E\left(\sum_{i=1}^{n} \bar{y}_{im}^{\,2}\right) = \frac{n}{N}\left[\sum_{i=1}^{N} \bar{y}_{i*}^{2} + N\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2}\right] \qquad \ldots(4)$$

$$\left(\because\ p = \frac{n}{N}\right)$$

Again $$\because\ V(\bar{y}_{nm}) = E(\bar{y}_{nm}^{\,2}) - \bar{y}_{**}^{\,2}$$

$$\therefore \quad n\,E\left[\bar{y}_{nm}^{\,2}\right] = \left(1 - \frac{n}{N}\right) S_b^{\,2} + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2} + n\,\bar{y}_{**}^{\,2} \qquad \ldots(5)$$

Using (5) and (4) in (3), we get

$$E(s_b^{\,2}) = S_b^{\,2} + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2} \qquad \ldots(6)$$

$$(n-1)\,E(s_b^{\,2}) = \frac{n}{N}\left[\sum_{i=1}^{N} \bar{y}_{i*}^{2} + N\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2}\right]$$

$$-\left(1 - \frac{n}{N}\right) S_b^{\,2} - \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2} - n\,\bar{y}_{**}^{2}$$

$$\therefore \quad E(S_b)^2 = \frac{1}{n-1}\left[\frac{n}{N}\sum_{i=1}^{N} \bar{y}_{i*}^{2} - \left(1 - \frac{n}{N}\right) S_b^{\,2} - n\,\bar{y}_{**}^{2}\right] + \frac{(n-1)}{n-1}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2}$$

$$= \frac{1}{n-1}\left[\frac{n}{N}\sum_{i=1}^{N} (\bar{y}_{i*}^{2} - \bar{y}_{**}^{2}) - \left(1 - \frac{n}{N}\right) S_b^{\,2}\right] + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2}$$

$$= \frac{1}{n-1}\left[\frac{n}{N}\sum_{i=1}^{N} (\bar{y}_{i*} - \bar{y}_{**})^2 - \left(1 - \frac{n}{N}\right) S_b^{\,2}\right] + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2}$$

$$= \frac{1}{n-1}\left[\frac{n}{N}(N-1)\,S_b^{\,2} - \left(1 - \frac{n}{N}\right) S_b^{\,2}\right] + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^{\,2}$$

$$= S_b^{\,2} + \left(\frac{1}{m} + \frac{1}{M}\right)\bar{S}_w^{\,2}$$

NOTES

Now, we know that for fixed i, $E(s_i^2) = S_i^2$.

Also for vaying i,

$$E\left(\frac{1}{n}\cdot\sum_{i=1}^{n}S_i^2\right) = \frac{1}{N}\sum_{i=1}^{N}S_i^2 = \overline{S}_w^2$$

$$\text{Thus, } \left[\frac{1}{n}\cdot\sum_{i=1}^{n}s_i^2\right] = \overline{S}_w^2 \qquad ...(7)$$

$$E\left[\frac{1}{n}\left(S_1^2+S_2^2+...+S_n^2\right)\right] = \overline{S}_w^2 \Rightarrow \frac{1}{n}\left[E(S_1^2)+E(S_2^2)+...+E(S_n^2)\right] = \overline{S}_w^2$$

$$\therefore \Rightarrow \frac{1}{n}[S_1^2+S_2^2+....+S_n^2] = \overline{S}_w^2$$

$$\Rightarrow \frac{1}{n}\left[E(S_1^2)+E(S_2^2)+...+E(S_n^2)\right] = \overline{S}_w^2$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}E(s_i^2) = \overline{S}_w^2 \Rightarrow E\left[\frac{1}{n}\Sigma s_i^2\right] = \overline{S}_w^2$$

$$\because \quad \text{Est. } \overline{S}_\omega^2 = \overline{s}_w^2 = \frac{1}{n}\sum_{i=1}^{n}s_i^2 = \frac{\sum_{i=1}^{n}(m-1)\,s_i^2}{n(m-1)} \qquad ...(8)$$

= Mean square within first stage units in the ANOVA of the sample

Also from (6),

$$\text{Est. } S_b^2 = s_b^2 - \left[\frac{1}{m}-\frac{1}{M}\right]\overline{s}_w^2 \qquad ...(9)$$

When, $m = M$, we have Est. $S_b^2 = s_b^2$ which is known result for one stage sampling without sub-sampling. Thus we see that in two stage sampling, the estimate of S_b^2 is less than the corresponding mean square between the first stage unit means in the sample.

$$\text{Now,} \quad V(\overline{y}_{nm}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_\omega^2$$

$$\therefore \quad \text{Est. } V(\overline{y}_{nm}) = \left(\frac{1}{n}-\frac{1}{N}\right)\text{Est. } S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\text{Est. } \overline{S}_w^2$$

$$\therefore \quad \text{Est. } V(\overline{y}_{nm}) = \left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + \frac{1}{N}\left(\frac{1}{m}-\frac{1}{M}\right)\overline{s}_w^2 \qquad ...(10)$$

As $N \to \infty$,

$$\text{Est. } V(\overline{y}_{nm}) = \frac{s_b^2}{n} = \frac{m\,s_b^2}{mn} = \frac{\text{Mean square between first stage units in the ANOVA of the sample}}{\text{nm}}$$

NOTES

Again for M large, $\frac{M-m}{M} = 1$,

$$\therefore \quad \text{Est. } S_b^2 = s_b^2 - \frac{\bar{s}_w^2}{m} \qquad (\because M \text{ is large})$$

$$\therefore \quad \text{Est } V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{Nm}\bar{s}_w^2$$

7.4 ALLOCATION OF SAMPLE TO THE TWO STAGES : EQUAL FIRST STAGE UNITS

We know that,

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 \qquad ...(1)$$

The expression (1) shows that the precision of a two stage sample, apart from the values of S_b^2 and $\bar{S}_w^2$, depends upon the distribution of the sample between the two stages *i.e.,* on n and m individually. Thus the cost of surveying a two stage sample will depend upon n and m. Here we will consider the problem of choosing n and m so that the variable of the sample mean is minimized for a given cost. Alternatively, we can choose n and m so as to provide an estimate of the desired precision for the minimum cost.

We shall first consider the case in which the cost of the survey is proportional to the size of the sample, so that

$$C = cnm \qquad ...(2)$$

where, C is the total cost of the survey and c is a constants. Let the total cost of the survey be fixed at, say $C = C_0$. Then (2) becames

$$m = \frac{C_0}{cn} \qquad ...(3)$$

Substituting (3) in (1), we get

$$V(\bar{y}_{nm}) = \left(S_b^2 - \frac{\bar{S}_w^2}{M}\right)\frac{1}{n} - \frac{S_b^2}{N} + \frac{c\,\bar{S}_w^2}{C_0} \qquad ...(4)$$

which is monotonically decreasing function of n if $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0$, reachin its minimum when n assumes the maximum. value, namely $\hat{n} = \frac{C_0}{c}$ for $\hat{m} = 1$.

If $\left(S_b^2 - \frac{\overline{S}_w^2}{M}\right) < 0$, which for large N, is equivalent to stating that if the intra-class correlation is –ve,then R.H.S of (4) becomes a monotonically increasing function of n, reaching its minimum when n is minimum given by $\hat{n} = \frac{C_0}{cm}$ or in other words there is no sub-sampling.

The alternative approach of minimizing the cost of estimating the population mean with the desired precision leads to the same solution for m. For, let V_0 be the value of the variable with which it is desired to estimate the population mean, so that

$$V_0 = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\frac{\overline{S}_w^2}{n} \qquad ...(5)$$

Solving for n, we get

$$n = \frac{S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\overline{S}_w^2}{V_0 + \frac{S_b^2}{N}} \qquad ...(6)$$

Substituting (6) in (2), we get

$$C = cm\left\{\frac{S_b^2 - \frac{\overline{S}_w^2}{M}}{V_0 + \frac{S_b^2}{N}}\right\} + \frac{C\,\overline{S}_w^2}{V_0 + \frac{S_b^2}{N}} \qquad ...(7)$$

Clearly, If $S_b^2 - \frac{\overline{S}_w^2}{M} > 0$, C attains a minimum value when m assumes the smallest integral value, namely 1 and if.

$S_b^2 - \frac{\overline{S}_w^2}{M} < 0$, the minimum is attained when $\hat{m}$ = M.

We shall now consider a more general case, when the cost of the survey is represented by

$$C = c_1 n + c_2 nm \qquad ...(8)$$

where c_1, c_2 are +ve constants.
Using (1) and (8), we obtain

$$C\left[V(\overline{y}_{nm}) + \frac{S_b^2}{N}\right] = c_1\left(S_b^2 - \frac{1}{M}\overline{S}_w^2\right) + c_2\,\overline{S}_\omega^2 + mc_2\left(S_b^2 - \frac{1}{M}\overline{S}_w^2\right) + \frac{C_1\,\overline{S}_w^2}{m} \qquad ...(9)$$

Clearly, the minimum value of (9) will provide the optimum allocation for both the cases, when either C or V is fixed is advance and V or C minimized.

NOTES

Now, for $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0$, (9) can be written as

$$C\left(V(\bar{y}_{nm}) + \frac{S_b^2}{M}\right) = \left(\sqrt{c_1\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right)} + \sqrt{c_2\,\bar{S}_w^2}\right)^2 + \left(\sqrt{mc_2\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right)} - \sqrt{\frac{c_1\,\bar{S}_w^2}{m}}\right)^2 \quad ...(10)$$

The equation (10) is minimum when second term on R.H.S is equated to zero

$$i.e., \qquad m = \sqrt{\frac{c_1}{c_2}\frac{\bar{S}_w^2}{S_b^2 - \frac{\bar{S}_w^2}{M}}} \cong \sqrt{\frac{c_1}{c_2}\left(\frac{1}{\rho} - 1\right)} \quad ...(11)$$

where ρ is the intra-class correlation within first stage units.

$$\frac{\bar{S}_w^2}{S_b^2 - \frac{\bar{S}_w^2}{M}} = \frac{\frac{(NM-1)}{NM}S^2(1-\rho)}{\frac{NM-1}{M(N-1)}\frac{S^2}{M}[1+(M-1)\rho] - \frac{1}{M}\frac{NM-1}{NM}S^2(1-\rho)}$$

$$= \frac{\frac{1-\rho}{N}}{\frac{1}{M(N-1)}[1+(M-1)\rho] - \frac{1-\rho}{NM}}$$

$$= \frac{(1-\rho)/N}{\frac{N+MN\rho-N\rho-N+N\rho+1-\rho}{M(N-1)}} = \frac{(1-\rho)/N}{\frac{\rho(NM-1)+1}{NM(N-1)}} = \frac{(1-\rho)}{\frac{\rho(NM-1)+1}{M(N-1)}}$$

$$= \frac{(1-\rho)}{\frac{\rho\left(M-\frac{1}{N}\right)+\frac{1}{N}}{M\left(1-\frac{1}{N}\right)}} \cong \frac{1-\rho}{\rho}, \text{ for large } N,$$

For $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) \le 0$, the expression on R.H.S of (10) is minimum when m is maximum attainable integral value. If the total cost fixed for the survey, namely, C_0, exceeds $(c_1 + c_2 M)$, this is given by $\hat{m} = M$ and n is the greatest integer not exceeding $\frac{C_0}{c_1 + c_2 M}$.

If C_0 is less than $(c_1 + c_2 M)$, m is greatest integer not exceeding $\frac{C_0 - c_1}{c_2}$ and $\hat{n}$ is 1.

Now, $\hat{m}$ is dependent upon the magnitude of c_1, c_2 as well as on ρ. In general, if $S_b^2 - \frac{\overline{S}_w^2}{M} > 0$, the optimum value for n will be small if (*i*) the travel cost between first stage units and other cost which go to make up C, are cheap (*ii*) the cost of collecting sub surplus from the selected first stage units is large and (*iii*) the intraclass correlation is large.

NOTES

7.5 COMPARISON OF TWO STAGE WITH ONE STAGE SAMPLING

One stage sampling procedures comparable with two stage sampling will involve either

(*i*) Sampling nm elements in one stage or

(*ii*) Sampling $\frac{nm}{M}$ first stage units as clusters, without further sub-sampling,

The variance of mean of a simple random sample of nm elements selected by procedure (*i*) is given by

$$V(\bar{y}_{nm}) = \left(\frac{1}{nm} - \frac{1}{NM}\right)S^2 \qquad ...(1)$$

To examine how this compares with the variance of a two stage sample, it is convenient to express the latter in terms of the intra-class correlation between elements of the first stage units.

Now, $$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{M}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\overline{S}_w^2$$

Also, $$S_b^2 = \frac{NM-1}{M(N-1)}\frac{S^2}{M}\left[1 + (M-1)\rho\right]$$

and

$$\overline{S}_w^2 = \frac{NM-1}{NM}S^2(1-\rho)$$

$$\therefore V(\bar{y}_{nm})_{\text{two stage}} = \frac{NM-1}{NM}\frac{S^2}{nm}\left[1 - \frac{m(n-1)}{M(N-1)} + \rho\left\{\frac{N-n}{N-1}\frac{m}{M}(M-1) - \frac{M-m}{M}\right\}\right]$$

...(2)

NOTES

When sub sampling rate $\frac{m}{M}$ is small *i.e.*, when M is large, (2) can be written as

$$V(\bar{y}_{nm})_{\text{two stage}} \cong \frac{S^2}{nm}\left[1+\rho\left(\frac{N-n}{N-1}m-1\right)\right] \quad ...(3)$$

Comparing (3) with (1) for large M, we see that the relative change in variable using sub-sampling in given by

$$\rho\left(\frac{N-n}{N-1}m-1\right)$$

The relative efficiency is given by

$$\text{R.E} \cong \frac{\frac{1}{nm}S^2}{\frac{S^2}{nm}\left[1+\rho\left\{\left(\frac{N-n}{N-1}m\right)-1\right\}\right]} = \frac{1}{1+\rho\left\{\frac{N-n}{N-1}m-1\right\}} \quad ...(4)$$

The R.E. of cluster sampling is

$$\text{R.E.} \cong \frac{1}{1+(M-1)\rho} \quad ...(5)$$

Comparing (4) and (5), we see that the Relative Efficiency of sub-sampling compared to unrestricted sampling of elements is approximately equal to that of sampling clusters of size $m\left(\frac{N-n}{N-1}\right)$.

7.6 EFFECT OF CHANGE IN SIZE OF FIRST STAGE UNITS ON THE VARIANCE

We know that the variance of mean of two stage sample consisting of n first stage units with m second stage units is

$$V(\bar{y}_{nm}) = \frac{NM-1}{NM}\frac{S^2}{nm}\left[1-\frac{m(n-1)}{M(N-1)}+\rho_1\left\{\frac{(N-n)m}{(N-1)M}(M-1)-\frac{M-m}{M}\right\}\right] \quad ...(1)$$

where, ρ_1 represents intra-class correlation within first stage units of size M.

We now suppose that the first stage units are combined to give N/c new first stage units with CM second stage units each. Thus new variance is given by

$$V'(\bar{y}_{nm}) = \frac{NM-1}{NM}\frac{S^2}{nm}\left[1-\frac{m(n-1)}{M(N-C)}+\rho_2\right.$$

$$\left.\left\{\frac{N-nC}{N-C}\cdot\frac{m}{MC}(MC-1)-\frac{MC-m}{MC}\right\}\right] \quad ...(2)$$

NOTES

where, ρ_2 will now represent the intra-class correlation within first stage units of size MC.

The difference between two variances can be expressed as

$$V(\bar{y}_{nm}) - V'(\bar{y}_{nm}) = \frac{NM-1}{NM}\frac{S^2}{nm}\left[\frac{m(n-1)}{M(N-1)}\frac{C-1}{N-C}+\alpha_1\rho_1-\alpha_2\rho_2\right] \quad ...(3)$$

where

$$\alpha_1 = \frac{N-n}{N-1}\frac{m}{M}(M-1)-\frac{M-m}{M}$$

$$\alpha_2 = \frac{N-nc}{N-C}\frac{m}{MC}(MC-1)-\frac{MC-m}{MC}$$

Since, $$\alpha_1-\alpha_2 = \frac{m}{M}\left[\frac{(C-1)(n-1)(NM-1)}{(N-1)(N-C)}\right] \geq 0$$

and the first term inside the bracket in (3) is non–negative, we conclude that

$$V(\bar{y}_{nm}) - V'(\bar{y}_{nm}) \geq 0.$$

whenever $\rho_1 > \rho_2$ provided that ρ_1 and ρ_2 are +ve.

In other words, again in precision is brought about by enlarging first stage units whenever the intra-class correlation (*i*) is +ve and (*ii*) decreases as the size of the first stage unit increases.

It also follows that smaller the value of ρ_2, the larger is the gain, so that by choosing for consolidation those first stage units which are as different as possible the gain can be increased.

7.7 ESTIMATION OF THE POPULATION MEAN

Now we give the theory appropriate for first stage units of unequal size when simple random sampling is employed at each stage.

Let,

M_i = the number of second stage units in the i^{th} first stage units i
$i = 1, 2, ..., N.$

NOTES

$M_0 = \sum_{i=1}^{N} M_i$, the total number of second stage units in the population.

m_i = the number of second stage units to be selected from the i^{th} first stage unit, if it is in the sample.

$m_0 = \sum_{i=1}^{n} m_i$, the number of second stage units in the sample.

$$\bar{y}_{i*} = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} \qquad ...(1)$$

$$\bar{\bar{y}}_{N*} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i*} \qquad ...(2)$$

$$\bar{y}_{**} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{y}_{i*}}{N\bar{M}} = \frac{1}{N}\sum_{i=1}^{N} u_i \bar{y}_{i*} \qquad ...(3)$$

where, $$N\bar{M} = \sum_{i=1}^{N} M_i \quad \text{and} \quad u_i = \frac{M_i}{\bar{M}} \qquad ...(4)$$

Several estimates of the population mean $\bar{y}_{**}$ can be formed. The simplest is the mean of the first stage unit means in the sample given by

$$\bar{y}_{s_2} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i(m_i)} \qquad ...(5)$$

where the summation runs over the first stage units in the sample and $\bar{y}_{i(m_i)}$ represents the A.M. of the m_i selected second stage units in the i^{th} first stage unit.

A second estimate to be formed is based on first stage unit totals and is given by

$$\bar{y}'_{s_2} = \frac{1}{n\bar{M}}\sum^{n} M_i \bar{y}_{i(m_i)} = \frac{1}{n}\sum^{n} u_i \bar{y}_{i(m_i)} \qquad ...(6)$$

A third estimate is the ratio estimate given by

$$\bar{y}''_{s_2} = \frac{\sum^{n} M_i \bar{y}_{i(m_i)}}{\sum^{n} M_i} = \frac{\sum^{n} u_i \bar{y}_{i(m_i)}}{\sum^{n} u_i} = \frac{\bar{y}'_{s_2}}{\bar{u}_n} \qquad ...(7)$$

NOTES

More generally, we may consider a ratio estimate of the population mean by letteing x_{ij} be the value of an auxillary variable x corresponding to the value y_{ij} of y, the variable under study. Let

$$\bar{x}_{**} = \frac{1}{N\bar{M}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} x_{ij} = \frac{1}{N} \sum_{i=1}^{N} u_i \bar{x}_{i*} \quad ...(8)$$

and

$$\bar{x}'_{s_2} = \frac{1}{n} \sum^{n} u_i \bar{x}_{i(m_i)} \quad ...(9)$$

Then the general ratio estimate of the population mean $\bar{y}_{**}$ is defined as

$$\bar{y}_{RS_2} = \frac{\bar{y}'_{s_2}}{\bar{x}'_{s_2}} \bar{x}_{**} \quad ...(10)$$

If $x_{ij} = 1 \; \forall \; i, j$ then $\bar{y}_{Rs_2} = \bar{y}''_{s_2}$.

$$\bar{y}_{Rs_2} = \frac{\bar{y}_{s_2}}{\bar{x}'_{s_2}} \bar{x}_{**} \text{ . Now } \bar{x}_{**} = \frac{1}{N\bar{M}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} 1 = \frac{1}{N\bar{M}} NM_i = \frac{M_i}{\bar{M}} = u_i$$

Also, $$\bar{x}_{i(m_i)} = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij} = \frac{1}{m_i} \sum_{j=1}^{m_i} 1 = 1$$

$$\therefore \quad \bar{x}_{s_2} = \frac{1}{n} \sum_{i=1}^{n} u_i .1 = \bar{u}_n$$

$$\therefore \quad \bar{y}_{Rs_2} = \frac{\bar{y}_{s_2}}{\bar{u}_n} u_i = \frac{\bar{y}_{s_2}}{\bar{u}_n} .1 = \bar{y}''_{s_2}$$

Since, $$u_i = \frac{M_i}{\bar{M}}$$

$$\Rightarrow \quad \sum_{i=1}^{N} u_i = \frac{\sum_{i=1}^{N} M_i}{\bar{M}} = N$$

$$\Rightarrow \sum_{i=1}^{N} (u_i - 1) = 0$$

$$\Rightarrow \quad u_i = 1 \; \forall_i$$

7.8 TWO STAGE SAMPLING : UNEQUAL FIRST STAGE UNITS

NOTES

Expected values and variances of the different estimates

(I) First Estimate $\bar{y}_{s_2}$

We have, $$E(\bar{y}_{s_2}) = E\left[\frac{1}{n}\sum^{n}\bar{y}_{i(m_i)}\right] = E\left[\frac{1}{n}\sum^{n}E(\bar{y}_{i(m_i)}/i)\right]$$

$$= E\left[\frac{1}{n}\sum^{n}\bar{y}_{i^*}\right] = \frac{1}{N}\cdot\sum_{i=1}^{N}\bar{y}_{i^*} = \bar{y}_{N^*} \neq \bar{y}_{**} \qquad ...(1)$$

Thus, $\bar{y}_{s_2}$ is biased estimate of $\bar{y}_{**}$.

Now, $$V(\bar{y}_{s_2}) = V\left[E(\bar{y}_{s_2}/n)\right] + E\left[V(\bar{y}_{s_2}/n)\right]$$

$$= V\left[\frac{1}{n}\sum^{n}\bar{y}_{i^*}\right] + E\left[\frac{1}{n^2}\sum^{n}V(\bar{y}_{i(m_i)}/i)\right]$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^{\,2} + E\left[\frac{1}{n^2}\sum^{n}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^{\,2}\right]$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^{\,2} + \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^{\,2} \qquad ...(2)$$

where, $$S_b^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{y}_{i^*}-\bar{\bar{y}}_{N^*}\right)^2 \qquad ...(3)$$

and

$$S_i^{\,2} = \frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(y_{ij}-\bar{y}_{i^*}\right)^2 \qquad ...(4)$$

Thus, M.S.E $(\bar{y}_{s_2}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^{\,2} + \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^{\,2} + (\bar{\bar{y}}_{N^*}-\bar{y}_{**})^2$ ...(5)

The bias of the estimate is

$$\bar{\bar{y}}_{N^*} - \bar{y}_{**} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i^*} - \frac{1}{N\bar{M}} \cdot \sum_{i=1}^{N} M_i \bar{y}_{i^*}$$

$$= -\frac{1}{N\bar{M}}\left[\sum_{i=1}^{N} M_i \bar{y}_{i^*} - \bar{M}\left(\sum_{i=1}^{N} \bar{y}_{i^*}\right)\right]$$

$$= -\frac{1}{NM}\sum_{i=1}^{N} \left(M_i - \bar{M}\right)\left(\bar{y}_{i^*} - \bar{\bar{y}}_{N^*}\right) \qquad ...(6)$$

An unbiased estimate of the bias is provided by

$$\text{Est. (Bias in } \bar{y}_{x_2}) = -\frac{N-1}{N\bar{M}(n-1)}\sum_{i=1}^{n}(M_i - \bar{M}_n)(\bar{y}_{i(mi)} - \bar{y}_{s_2}) \qquad ...(7)$$

E [Est. (Bias in $\bar{y}_{s_2}$)]

$$= -\frac{N-1}{N\bar{M}} E\left[\frac{1}{n-1}\sum_{i=1}^{n} E\left\{(M_i - \bar{M}_n)(\bar{y}_{i(m_i)} - \bar{y}_{s_2}) / n\right\}\right.$$

$$= -\frac{N-1}{N\bar{M}} E\left[\frac{1}{n-1}\sum_{i=1}^{n} (M_i - \bar{M}_n)(\bar{y}_{i^*} - \bar{\bar{y}}_{n^*})\right]$$

$$= -\frac{1}{N\bar{M}}\sum_{i=1}^{N} (M_i - \bar{M})(\bar{y}_{i^*} - \bar{\bar{y}}_{N^*}) = (\bar{\bar{y}}_{N^*} - \bar{y}_{**})$$

Thus it follows that an unbiased estimate of the population mean is given by

$$\bar{y}_{s_2} + \frac{N-1}{N\bar{M}}\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{M}_n)(\bar{y}_{i(m_i)} - \bar{y}_{s_2}) \qquad ...(8)$$

The bias arises because the inequality of sizes of the first stage units causes the probabilities of selection of the second stage units to vary from one first stage unit to another.

(II) Second Estimate $\bar{y}_{s_2}$

This is an unbiased estimate of population mean

$$\therefore \qquad E(\bar{y}'_{s_2}) = E\left[\frac{1}{n}\sum_{i=1}^{n} u_i \bar{y}_{i(m_i)}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} u_i E(\bar{y}_{i(m_i)} / i)\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} u_i\,\bar{y}_{i*}\right] = \frac{1}{n}\sum_{i=1}^{N} u_i\bar{y}_{i*} = \bar{y}_{**} \qquad ...(9)$$

NOTES

The variance of estimate is given by

$$V(\bar{y}'_{s_2}) = V\left[E\left(\bar{y}'_{s_2}/n\right)\right] + E\left[V\left(\bar{y}'_{s_2}/n\right)\right]$$

$$= V\left[\frac{1}{n}\sum_{i=1}^{n} u_i\bar{y}_{i*}\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n} u_i^2 V(\bar{y}_{i(m_i)}/i)\right]$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)S_b'^2 + E\left[\frac{1}{n^2}\sum_{i=1}^{n} u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2\right]$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)S_b'^2 + \frac{1}{nN}\sum_{i=1}^{N} u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2 \qquad ...(10)$$

where, $$S_b'^2 = \frac{1}{N-1}\sum_{i=1}^{n}(u_i\bar{y}_{i*} - \bar{y}_{**})^2$$

and $$S_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{y}_{i*})^2$$

It is seen that first component in (10) depends upon the variation between the cluster totals and it is larger than the corresponding component in (2), provided the correlation between the cluster size and the cluster mean is +ve and the bias compoent $(\bar{\bar{y}}_{N*} - \bar{y}_{**})$ is negligible. The second component of (10) is also likely to be larger than the corresponding component in (2). Thus unless the bias in $\bar{y}_{s_2}$ is therefore likely to be serious, the estimate $\bar{y}'_{s_2}$ may not be preferred to $\bar{y}_{s_2}$.

7.9 CONCEPT OF MULTISTAGE SAMPLING

Suppose that each unit in the population can be divided into a number of smaller units, or subunits. A sample on n units has been selected. If subunits within a selected unit give similar results, we can select and measure a sample of the subunits in any chosen unit to avoid uneconomically measure

NOTES

them all. This technique is called *subsampling,* since the unit is measured by its samples, or *multi-stage sampling,* since the sample is taken in moire than one step. The firstly selected sample units are called the *primary units,* and then select a sample of subunits from these primary units.

Multistage sampling is a complex form of cluster sampling. Using all the sample elements in all the selected clusters may be prohibitively expensive or not necessary. Under these circumstances, multistage cluster sampling becomes useful. Instead of using all the elements contained in the selected clusters. The researcher randomly selects elements from each cluster. Constructing the clusters is the first stage. Deciding what elements within the cluster to use is the second stage. The technique is used frequently when a complete list of all members of the population does not exist and is inappropriate.

In some cases, several levels of cluster selection may be applied before the final sample elements are reached. For example, household surveys conducted by the Australian Bureau of Statistics begin by dividing metropolitan regions into 'collection districts', and selecting some of these collection districts (first stage). The selected collection districts are then divided into blocks, and blocks are chosen from within each selected collection district (second stage). Next, dwellings are listed within each selected block, and some of these dwellings are selected (third stage). This method mean that it is not necessary to create a list of every dwelling in the region, only for selected blocks. In remote areas, an additional stage of clustering is used, in order to reduce travel requirements.

Although cluster sampling and stratified sampling bear some superficial similarities, they are substantially different. In stratified sampling, a random sample is drawn from all the strata, where in cluster sampling only the selected clusters are studied, either in single stage or multi stage.

The prime stimulus for multi-stage sampling is administrative convenience. It is more flexible than one stage sampling. It reduces to one stage sampling, unless this is the best choice of smaple size of subsample. We have chance of selecting smaller value which appears moire efficient.

In two-stage sampling are sampling plan gives first method for selecting n units. Then for each stages unit, a method is given for selecting the specified number of subunits from it. In finding the mean and variance of an estimate average must be taken over all samples that can be generated by this first stage process.

STUDENT ACTIVITY 1

1. Write a short note on two-stage sampling.

STUDENT ACTIVITY 2

1. Show that $\bar{y}_{nm}$ is an unbiased estimage of the population mean $\bar{y}_{**}$.

NOTES

7.10 ILLUSTRATIVE EXAMPLES

Example 1. *At an experimental station, there were 100 fields sown with wheat. Each field was divided into 16 plots of equal size (1.16th hectare). Out of 100 fields, 10 were selected by simple random sampling, wor. From each selected field, 4 plots were chosen by random sampling, wor. The yields in kg/plot are given below :*

Selected field	Plots			
	1	2	3	4
1	4.32	4.84	3.96	4.04
2	4.16	4.36	3.50	5.00
3	3.06	4.24	4.76	3.12
4	4.00	4.84	4.32	3.72
5	4.12	4.68	3.46	4.02
6	4.08	3.96	3.42	3.08
7	5.16	4.24	4.96	3.84
8	4.40	4.72	4.04	3.98
9	4.20	4.66	3.64	5.00
10	4.28	4.36	3.00	3.52

(i) Estimate the wheat yield per hectare for the experimental station along with its standard error.

(ii) How can an estimate obtained from a simple random sample of 40 plots be compared with the estimate obtained above, in (i) ?

(iii) Obtain optimum n and m under cost function 100 = 4n + nm.

Solution. Here, $N = 100$, $M = 16$, $n = 10$ and $m = 4$.

consider the following Table :

S. No.	$\sum_{j}^{4} \bar{y}_{ij}$	$\bar{y}_{i.}$	$(\bar{y}_{i.} - \bar{y})^2$	$\sum_{j=1}^{4} y_{ij}^2$	$\bar{y}_{i.}^2$	$\left(\sum_{j}^{4} \bar{y}_{ij}^2 - m\bar{y}_{i.}^2\right)$
1	17.16	4.290	0.0267	74.091	18.404	0.475
2	17.02	4.255	0.0165	73.565	18.105	1.145
3	15.18	3.795	0.4469	59.733	14.402	2.125
4	16.88	4.220	0.0087	71.925	71.808	0.694
5	16.28	4.070	0.0143	67.009	16.545	0.749
6	14.54	3.635	0.2586	53.511	13.213	0.659

7	18.20	4.550	0.1794	83.950	20.703	1.138
8	17.14	4.285	0.0251	73.800	17.361	0.356
9	17.50	4.375	0.0618	77.605	19.141	1.041
10	15.16	3.790	0.4402	58.718	14.364	1.262
Total		41.265	1.4782	693.908		9.644

(*i*) An estimate of the average wheat yield, with usual notations, is given by

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} \bar{y}_{i.} = \frac{41.265}{10} = 4.1265$$

The estimated variance of $\bar{y}$ is

$$v(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) s_w^2$$

On simplifications, we get

$$s_b^2 = \frac{1}{n-1}\sum_{i}^{n} (\bar{y}_{i.} - \bar{y})^2 = \frac{1.4782}{9} = 0.1642$$

and $$s_w^2 = \frac{1}{n(m-1)}\sum_{i}^{n}\sum_{j}^{m} (y'_{ij} - \bar{y}_{i.})^2 = \frac{9.644}{30} = 0.3215$$

Therefore, $$v(\bar{y}) = \left(\frac{1}{10} - \frac{1}{100}\right) 0.1642 + \frac{1}{100}\left(\frac{1}{4} - \frac{1}{16}\right) 0.3215$$

$$= 0.0145$$

and standard error of $\bar{y} = \sqrt{0.0145} = 0.120$

(*ii*) In simple random sampling, the estimate of variance is given by

$$v_{ran}(\bar{y}) = \left(\frac{1}{nm} - \frac{1}{NM}\right) s^2$$

The estimate of S^2, using two stage sampling design, can be written as

$$s^2 = \frac{1}{NM-1}\left[M(N-1)s_b^2 + \left\{N(M-1) - (M-m)\frac{(N-1)}{m}\right\} s_w^2\right]$$

$$= \frac{1}{1600-1}\left[16 \times 99 \times 0.1642 + \left\{100 \times 15 - \frac{99 \times 12}{4}\right\} 0.3215\right]$$

$$= 0.4045$$

Thus, $$v_{ran}(\bar{y}) = \left(\frac{1}{40} - \frac{1}{1600}\right) 0.4045 = 0.0099$$

(*iii*) The given cost function is of the form $C = c_1 n + c_2\, nm$ with $c_1 = 4$, $c_2 = 1$, and $C = 100$. The optimum value of m is given by

NOTES

$$m_{opt} = \left[\frac{c_1}{c_2}\,\frac{s_w^2}{s_b^2 - \frac{s_w^2}{m}}\right]^{1/2}$$

$$= \left[\frac{4}{1} \times \frac{0.3215}{0.1642 - (0.3215/4)}\right]^{1/2} = 4$$

Substituting the value of m in the given cost function, the optimum value of n is given by

$$n_{opt} = \frac{100}{8} \cong 12.5$$

Example 2. *For study of feeding and rearing practices of sheep and yield of wool in the Rajasthan State, during the year 1980-81, two stage sampling design with tehsils as first stage units and villages in the tehsil as second stage units was adopted. The data given below are the stationary sheep population in the selected villages in each of 4 tehsils selected from 12 tehsils of a certain Division, as counted in the survey along with the number of villages in the teshil.*

Selected tehsil	*Number of villages in the tehsil (M_i)*	*Stationary sheep population in the selected villages*
X	*102*	*266, 890, 311, 46, 174, 31, 17, 186, 224, 31, 102, 46, 31, 109, 275, 128, 125, 267, 153, 152, 84, 21, 52, 10, 0 48, 94, 123, 87, 109, 0, 310, 3*
Y	*105*	*129, 57, 64, 11, 163, 77, 278, 50, 26, 127, 252, 194, 350, 0, 572, 149, 275, 114, 387, 53, 34, 150, 224, 185, 157, 244, 466, 203, 354, 816, 242, 140, 66, 590, 747, 147*
Z	*200*	*247, 622, 225, 278, 181, 132, 659, 403, 281, 236, 595, 265, 431, 190, 348, 232, 88, 1165, 831, 120, 987, 938, 197, 614, 187, 896, 330, 485, 60, 60, 1051, 651,] 552, 968, 987*
T	*88*	*347, 362, 34, 11, 133, 36, 34, 61, 249, 170, 112, 42, 161, 75, 68, 0, 247, 186, 473, 0, 143, 198, 65, 0, 308, 122, 345, 0, 223, 302, 219, 120, 199, 35, 0, 0*

NOTES

Solution. Estimate the mean stationary sheep population in the Division during the year 1980–81, together with its standard error when $\overline{M} = 124$. Here we have

$$\overline{M} = 124, \quad M_1 = 102, \quad M_2 = 105, \quad M_3 = 200, \quad M_4 = 88$$

$$n = 4, \quad m_1 = 34, \quad m_2 = 36, \quad m_3 = 35, \quad m_4 = 36$$

$$\overline{y}_{1.} = 135, \quad \overline{y}_{2.} = 225, \quad \overline{y}_{3.} = 471, \quad \overline{y}_{4.} = 141$$

$$s_{w1}^2 = \frac{1}{33}(1427522 - 619650) = 24481$$

$$s_{w2}^2 = \frac{1}{35}(3219435 - 1822500) = 39912$$

$$s_{w3}^2 = \frac{1}{34}(11451530 - 7764435) = 105346$$

and $$s_{w4}^2 = \frac{1}{35}(1274076 - 715716) = 15953$$

(I) **First Estimate.** An unbiased estimate of the mean of the sheep population is given by

$$\overline{y} = \frac{1}{n\overline{M}} \sum_i^n M_i \overline{y}_{i.}$$

$$\therefore \quad \overline{y} = \frac{(102 \times 135) + (105 \times 225) + (200 \times 471) + (88 \times 141)}{4 \times 124}$$

$$= \frac{144003}{496} \cong 290$$

The estimate of $V(\overline{y})$ is given by

$$v(\overline{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{nN} \sum_i^n W_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{wi}^2$$

where, $$s_b^2 = \frac{1}{n-1} \sum_i^n (W_i \overline{y}_{i.} - \overline{y})^2$$

On simplifying,

$$s_b^2 = \frac{1}{4-1} [\{0.67 \times (135)^2 + 0.72 \times (225)^2$$

$$+ 2.60 \times (471)^2 + 0.50 \times (141)^2\} - 4 \times (290)^2]$$

$$= \frac{163054.79}{3} = 54351.60$$

NOTES

and $$\frac{1}{nN}\sum_{i}^{n} W_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{wi}^2$$

$$= \frac{1}{48}\left[\left(\frac{102}{124}\right)^2 \times 473 + \left(\frac{105}{124}\right)^2 \times 729 + \left(\frac{200}{124}\right)^2 \times 2483 + \left(\frac{88}{124}\right)^2 \times 262\right]$$

$$= \frac{7433.99}{48} = 154.87$$

Thus,

$$v(\bar{y}) = \frac{54351.60}{6} + 154.87 = 9213.47$$

$\therefore$ Standard crror of $\bar{y} = \sqrt{9213.47} = 94.94$

(II) **Second Estimate.** Another estimate of the mean stationary sheep population is

$$\bar{y}_1 = \frac{1}{n}\sum_{i}^{n} \bar{y}_{i.}$$

$$= \frac{1}{4}\left[\frac{4594}{34} + \frac{8093}{36} + \frac{16492}{35} + \frac{5080}{36}\right] = \frac{972}{4} = 243$$

An estimate of the variance of $\bar{y}_1$ is given by

$$v(\bar{y}_1) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{nN}\sum_{i}^{n} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{wi}^2$$

where, $$s_{wi}^2 = \frac{1}{m_i - 1}\sum_{j}^{mi} \left(y_{ij} - \bar{y}_i\right)^2$$

and $$s_b^2 = \frac{1}{n-1}\sum_{i}^{n} \left(\bar{y}_{i.} - \bar{y}_1\right)^2$$

$\therefore$ $$s_b^2 = \frac{1}{3}\ (310572 - 236196) = 24792$$

and $s_{w_1}^2, s_{w_2}^2, s_{w_3}^2$ and $s_{w_4}^2$, are as given above

Thus,

$$v(\bar{y}_1) = \left(\frac{1}{4} - \frac{1}{12}\right) \times 24792 + \frac{1}{4 \times 12}\left[\left(\frac{1}{34} - \frac{1}{102}\right) \times 24481\right.$$

$$+ \left(\frac{1}{36} - \frac{1}{105}\right) \times 39912 + \left(\frac{1}{35} - \frac{1}{200}\right) \times 105346$$

$$\left. + \left(\frac{1}{36} - \frac{1}{88}\right) \times 15953\right]$$

$$= 4214$$

$\therefore$ Standard error of $\bar{y}_1 = \sqrt{4214} = 64.92$

(III) **Third Estimate.** We have $\bar{y}_2$, also an estimate of population mean, which is

$$\bar{y}_2 = \frac{\sum_i^n M_i \bar{y}_{i.}}{\sum_i^n M_i} = \frac{\frac{1}{n}\sum_i^n u_i \bar{y}_{i.}}{\frac{1}{n}\sum_i^n u_i} = \frac{\bar{y}}{\bar{u}}$$

We calculate, $\bar{u} = \frac{\sum_i^n M_i}{n\bar{M}} = \frac{1}{4}(0.82 + 0.85 + 1.61 + 0.71)$

$= 0.998$

Hence, $\bar{y}_2 = \frac{\bar{y}}{\bar{u}} \cong 291$

Also, estimate of $V(\bar{y}_2)$ is

$$v(\bar{y}_2) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b''^2 + \frac{1}{nN}\sum_i^n u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{wi}^2$$

where $s_b''^2 = \sum_i^n \frac{u_i^2 (\bar{y}_{i.} - \bar{y}_2)^2}{(n-1)}$

$\therefore$ $s_b''^2 = \frac{1}{(4-1)}$ [0.67 $(135 - 291)^2$ + 0.72 $(225 - 291)^2$

+ 2.60 $(471 - 291)^2$ + 0.05 $(141 - 291)^2$]

$= 38401.87$

and the second part of the variance is as in (*i*).

Thus, $v(\bar{y}_2) = \frac{38401.87}{6} + 154.87$

$= 6555.19$

$\therefore$ Standard error $\bar{y}_2 = \sqrt{6555.19} = 80.86$

7.11 SUMMARY

- A sampling in which clusters are selected first and then a specified number from the selected cluster is known as sub-sampling or two stage sampling.

NOTES

NOTES

7.12 GLOSSARY

- **Sub-sampling.** The procedures of first selecting cluster, and there choosing a specified number of elements from selected cluster is known as sub-sampling or two stage sampling.

7.13 REVIEW QUESTIONS

1. In two-stage sampling, n fsu's are selected with pps, wr. If the i^{th} fsu occurs r_i times in the sample, one of the following procedures may be adopted for second stage sampling :

   (*i*) r_i m_i ssu's are selected with simple random sample, wor;

   (*ii*) r_i independent samples of m_i ssu's (drawn with simple random sample, wor) are taken; and

   (*iii*) m_i units are selected without replacement and observations are weighted by r_i.

   Obtain unbiased estimators of the population total Y and their sampling variances for all these cases. If V_1, V_2 and V_3 are the variances of the estimators, show that, for the same expected sample size,

   $$V_1 \leq V_2 \leq V_3$$

2. The following sampling schemes for estimating the population mean of a characteristic were considered :

   (*i*) The population is divided into N clusters of M units each and two-stage sampling is adopted where n clusters and m units from each sampled cluster are selected with simple random sampling, *wr*, and

   (*ii*) The population is divided into clusters of m' units each and a sample of n' such clusters is selected with simple random sampling, *wr*.

   Show that, in both the cases, the sample mean is an unbiased estimator of the population mean and derive the variances in both cases. Derive the condition that the efficiencies of these two schemes be the same when $nm = n'\ m'$. (Singh. D., 1956)

3. In a sample survey for estimating the number of standards of pepper in a Tehsil with 72 villages, a sample of 12 villages was selected with srs, wor, and from each selected village 5 clusters of 20 fields each were drawn with srs, wor. Data on the number of clusters in the sample

villages and on the number of standards in the sample clusters are given below :

Sample village	*No. of clusters*	*Number of standards in sample clusters*				
		1	*2*	*3*	*4*	*5*
1	27	430	402	363	975	389
2	24	586	1234	100	368	344
3	14	1164	546	3060	1724	1274
4	116	693	218	836	1218	575
5	25	191	270	4502	4184	243
6	118	1036	1333	1179	728	1957
7	147	1555	254	950	382	355
8	36	910	452	129	122	243
9	91	340	0	92	28	340
10	171	57	59	0	0	21
11	86	159	45	242	1075	539
12	88	84	462	147	16	10

Estimate unbiasedly the total number of standards in the Tehsil and also obtain its standard error.

4. Raw wool contains varying amounts of grease, dirt and other impurities and its quality is measured by the percentage by weight of clean wool, termed clean content. To estimate the clean content, an electrical core-boring machine is used which takes cores of about 1/4 1b from a bale, which are then subjected to laboratory analysis. In an experiment, 6 bales were drawn from a large lot with equal probability and from each bale 4 cores were taken at random and the clean contents determined. The results of this experiment are given on next page.

Core	*Sample bales*					
	1	*2*	*3*	*4*	*5*	*6*
1	54.3	57.0	54.6	54.9	59.9	57.8
2	56.2	58.7	57.5	60.1	57.8	59.7
3	58.9	58.2	59.3	58.7	60.9	59.6
4	55.5	57.1	57.5	55.6	57.5	58.1

(*i*) Estimate the average clean content of wool for the lot and also obtain an estimate of its standard error.

(*ii*) Obtain the efficiency of sampling 12 bales and 2 cores from each bale as compared to that of the above scheme.

NOTES

5. To estimate the total yield of paddy in a district, a stratified two-stage sampling design was adopted where 4 villages were selected from each stratum, with pps, wr, size being geographical area. From each sample village 4 plots were drawn circular systematically for ascertaining the yield of paddy. The following data give the yield of paddy for the sample plots :

Stratum	*Sample village*	*Inverse of probability*	*Total number of plots*	*Yield of paddy (kg)*			
				1	*2*	*3*	*4*
1	1	440.21	28	104	182	148	87
	2	660.43	84	108	64	132	156
	3	31.50	240	100	115	50	172
	4	113.38	76	346	350	157	119
2	1	21.00	256	124	111	135	216
	2	16.80	288	123	177	106	138
	3	24.76	222	264	78	144	55
	4	49.99	69	300	114	68	111
3	1	67.68	189	110	281	120	114
	2	339.14	42	80	61	118	124
	3	100.00	134	121	212	174	106
	4	68.07	161	243	116	314	129

Estimate the total yield of paddy and obtain an estimate of its standard error.

6. A crop cutting survey by the method of stratified multi-stage random sampling was carried out in one district, on jute crop, for estimating the average yield of green weight of jute for the district, with its three administrative sub-divisions constituting the strata. In each administrative sub-division, a specified number of villages was selected at random within each selected village. Three fields under jute were choosen at random out of the total number of fields under jute in the village. In each field, a plot of 1/160th acres was located, harvested and the green weight of jute recorded in kg. The data obtained are shown below :

NOTES

Sub-division	Total area under jute in acres	Yields of green weight of jute in kg per plot for villages and fields selected
1	5089	86, 85, 57, 81, 71, 92, 72, 37, 51, 81, 50, 43, 78, 71, 79
2	4133	86, 45, 81, 55, 56, 55, 91, 70, 64, 19, 62, 41
3	3007	81, 8, 43, 67, 48, 47, 35, 34, 37

Estimate the average yield of green weight of jute in kg per acre for the district and calculate its standard error.

7. Show that the relative efficiency of sub-sampling (for unrestricted case) is approximately equal to that of sampling cluster of size $m\left(\frac{N-n}{N-1}\right)$.

8. What is the effect of change in size of first stage units on the variable.

9. Obtain different estimates of the population mean in two stage sampling with unequal first stage units.

7.14 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

NOTES

CHAPTER 8 RATIO METHOD OF ESTIMATION

OBJECTIVES

After going through this chapter, we should be able to :

- compute expected value and variance of the ratio estimate
- compute estimate of the variance of the ratio estimate
- know optimum property and efficiency of the ratio estimate

STRUCTURE

8.1 INTRODUCTION

In the ratio method, we obtain an auxilialry variate x_i correlated with y_i, when the population total X of the x_i is known. But in practice, x_i is oftenly

taken as the value of y_i at some precious time when a complete census was taken.

In developing the theory of simple random sampling, we have considered only estimates based on simple arithmetic means of the observed values in the sample. In ratio method of estimation, we make use of the ancillary infromation which under certain conditions, give more reliable estimates of the population values than those based upon the simple averages. We shall first deal with the theory for sampling with equal probability of selection.

8.2 NOTATIONS

Let us denote by

y_i = the value of the characteristic under study for the i^{th} unit of the population.

x_i = the value of the ancillary characteristic on the same unit.

Y = the total of y values in the population.

X = the total of x values in the population.

$r_i = \frac{y_i}{x_i}$: the ratio of y to x for the i^{th} unit.

$\bar{r}_N = \frac{1}{N}\sum_{i=1}^{N} r_i$: the simple A.M. of the ratios for the units in the population.

$\bar{r}_n = \frac{1}{n}\sum_{i=1}^{n} r_i$: the simple A.M. of the ratios for the units in the sample.

$R_N = \frac{\bar{y}_N}{\bar{x}_N} = \frac{Y}{X}$: the ratio of the population mean of y to the population mean of x.

$R_n = \frac{\bar{y}_n}{\bar{x}_n} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$: the corresponding ratio for the sample.

Ratio Estimate

R_n provides an estimate of the population ratio R_N and the product of R_N and $\bar{x}_N$ given by $\bar{y}_R = R_N \bar{x}_N$ gives an estimate of the mean value $\bar{y}_N$ in the population. This estimate is known as the **ratio estimate** of the population mean.

8.3 BIAS OF THE RATIO ESTIMATOR

NOTES

We first show that R_n is biased estimate of R_N.

Since $\bar{y}_n$ and $\bar{x}_n$ are unbiased estimates of $\bar{y}_N$ and $\bar{x}_N$ respectively.

We may write

$$R_N = \frac{\bar{y}_N}{\bar{x}_N} = \frac{E(\bar{y}_n)}{E(\bar{x}_n)} = \frac{E(R_n\bar{x}_n)}{E(\bar{x}_n)} \quad ...(1)$$

$$\therefore \quad \text{Bias in } R_n = E(R_n) - R_N = E(R_n) - \frac{E(R_n\bar{x}_n)}{E(\bar{x}_n)}.$$

$$= \frac{E(R_n)E(\bar{x}_n) - E(R_n\bar{x}_n)}{E(\bar{x}_n)} = -\frac{Cov(R_n, \bar{x}_n)}{\bar{x}_N} \quad ...(2)$$

The above result enables us to obtain as upper bound to the bias in the ratio estimate. We see that

$$|\text{Bias in } R_n| \le \frac{\sigma_{R_n}.\sigma_{\bar{x}_n}}{\bar{x}_N}$$

$$= \sigma_{R_n}\sqrt{\frac{N-n}{Nn}}\,C_x \quad ...(3)$$

$$Cov(R_n, \bar{x}_n) = r\sigma_{R_n}\sigma_{\bar{x}_n} \le \sigma_{R_n}\sigma_{\bar{x}_n}$$

$$\because \; V(\bar{x}_n) = \frac{N-n}{N}.\frac{S^2}{n}$$

where $C_x = \dfrac{S_x}{\bar{x}_N}$ is the coefficient of variation of x.

Equation (3) implies that if n is sufficiently large, the bias in ratio estimate R_n is negligible as compared is S.D.

In order to obtain various approximations to the bias, we proceed as follows.

Let

$$y_i = \bar{y}_N + \varepsilon_i$$

so that

$$\bar{y}_n = \bar{y}_N + \bar{\varepsilon}_n \quad ...(4)$$

where,

$$E(\bar{\varepsilon}_n) = 0 \quad \text{and} \quad E(\bar{\varepsilon}_n^{\,2}) = \frac{N-n}{N}\frac{S_y^{\,2}}{n} \quad ...(5)$$

Similarly, let $x_i = \bar{x}_N + \varepsilon_i'$ so that

$$\bar{x}_n = \bar{x}_N + \bar{\varepsilon}_n' \quad ...(6)$$

where

$$E(\bar{\varepsilon}_n') = 0 \quad \text{and} \quad E(\bar{\varepsilon}_n'^{\,2}) = \frac{N-n}{N}\frac{S_x^{\,2}}{n} \quad ...(7)$$

NOTES

Now, $$R_N = \frac{\bar{y}_n}{\bar{x}_n} = \frac{\bar{y}_N\left[1+\frac{\bar{\varepsilon}_n}{\bar{y}_N}\right]}{\bar{x}_N\left[1+\frac{\bar{\varepsilon}_n}{\bar{x}_N}\right]} \quad ...(8)$$

It is understood that $\bar{x}_n \neq 0, \bar{x}_N \neq 0$ and $\bar{y}_N \neq 0$. We now assume that $|\bar{\varepsilon}_v' / \bar{x}_N| < 1$ so that we may expand $\left(1+\frac{\bar{\varepsilon}_n'}{\bar{x}_N}\right)$ as a series in powers of $\bar{\varepsilon}_n'$.

Expanding and multiplying, we get

$$R_n = R_N\left(1+\frac{\bar{\varepsilon}_n}{\bar{y}_N}\right)\left[1+\frac{\bar{\varepsilon}_n'}{\bar{x}_N}\right]^{-1}$$

$$= R_N\left(1+\frac{\bar{\varepsilon}_n}{\bar{y}_N}\right)\left[1-\frac{\bar{\varepsilon}_n'}{\bar{x}_N}+\left(\frac{\bar{\varepsilon}_n'}{\bar{x}_N}\right)^2-\left(\frac{\bar{\varepsilon}_n'}{\bar{x}_N}\right)^3+\left(\frac{\bar{\varepsilon}_n'}{\bar{x}_N}\right)^4-\ldots\right]$$

$$= R_N\left[1+\frac{\bar{\varepsilon}_n}{\bar{y}_N}-\frac{\bar{\varepsilon}_n'}{\bar{x}_N}+\frac{\bar{\varepsilon}_{n1}^{\,2}}{\bar{x}_N^{\,2}}-\frac{\bar{\varepsilon}_n\bar{\varepsilon}_n'}{\bar{y}_N\bar{x}_N}+\frac{\bar{\varepsilon}_n\bar{\varepsilon}_{n1}'^{\,2}}{\bar{y}_N\bar{x}_N^{\,2}}-\frac{\bar{\varepsilon}_n'^{\,3}}{\bar{x}_N^{\,3}}+\right.$$

$$\left.\frac{\bar{\varepsilon}_n'^{\,4}}{\bar{x}_N^{\,4}}-\frac{\bar{\varepsilon}_n\bar{\varepsilon}_n'^{\,3}}{\bar{y}_N\bar{x}_N^{\,3}}+\ldots\right] \quad ...(9)$$

Taking expectations term by term, we get

$$E(R_n) = R_N + R_N\,E\left[\frac{\bar{\varepsilon}_n'^{\,2}}{\bar{x}_N^{\,2}}-\frac{\bar{\varepsilon}_n\bar{\varepsilon}_n'}{\bar{y}_N\bar{x}_N}+\frac{\bar{\varepsilon}_n\bar{\varepsilon}_n'^{2}}{\bar{y}_N\bar{x}_N^{\,2}}\right.$$

$$\left.-\frac{\bar{\varepsilon}_n'^{\,3}}{\bar{x}_N^{\,3}}+\frac{\bar{\varepsilon}_n'^{\,4}}{\bar{x}_N^{\,4}}-\frac{\bar{\varepsilon}_n\bar{\varepsilon}_n'^{\,3}}{\bar{y}_N\bar{x}_N^{\,3}}+\ldots\right] \quad ...(10)$$

Various approximation to the bias in the ratio estimate R_n can now be obtained. First we shall assume that the contribution of terms involving powers in $\bar{\varepsilon}_n$ and $\bar{\varepsilon}_n'$ higher than second to the value of $E(R_n)$ is negligible, being of the order of $\frac{1}{n^v}$, $v > 1$.

Denoting to a first approximate the expected value of R_n by E_1, (R_n), We may write

$$E_1(R_n) = R_N + R_N\left[\frac{E(\bar{\varepsilon}_n'^{\,2})}{\bar{x}_N^{\,2}}-\frac{E(\bar{\varepsilon}_n\bar{\varepsilon}_n')}{\bar{y}_n\bar{x}_N}\right] \quad ...(11)$$

NOTES

Now, $$E(\bar{\varepsilon}_n \bar{\varepsilon}_n') = \frac{1}{n^2} E\left[\left(\sum_{i=1}^{n} \varepsilon_i\right)\left(\sum_{i=j}^{n} \varepsilon_i'\right)\right]$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} \varepsilon_i \varepsilon_i' + \sum_{i \neq j}^{n} \varepsilon_i \varepsilon_i'\right]$$

$$= \frac{1}{n^2}\left[\frac{n}{N}\sum_{i=1}^{N} \varepsilon_i \varepsilon_i' + \frac{n(n-1)}{N(N-1)} \sum_{i \neq j=1}^{N} \varepsilon_i \varepsilon_j'\right]$$

$$E\left(\sum^{n} \varepsilon_i \varepsilon_i'\right) = \sum_{i=1}^{N} (\varepsilon_i \varepsilon_i')p. = \frac{n}{N}\sum_{i=1}^{N} \varepsilon_i \varepsilon_i'$$

where $\frac{n}{N}$ is the probability that the specified unit is included in the sample.

Also, we know that $$\frac{1}{n(n-1)} E\left[\sum_{i \neq j}^{n} y_i y_j\right] = \frac{n(n-1)}{N(N-1)} \sum_{i \neq j=1}^{N} y_i y_j$$

$$\therefore \quad E(\bar{\varepsilon}_n \bar{\varepsilon}_n') = \frac{1}{nN}\sum_{i=1}^{N} \varepsilon_i \varepsilon_i' + \frac{(n-1)}{nN(N-1)}\left[\left(\sum_{i=1}^{N} \varepsilon_i\right)\left(\sum_{i=1}^{N} \varepsilon_i'\right) - \sum_{i=1}^{N} \varepsilon_i \varepsilon_i'\right]$$

$$= \frac{N-n}{Nn}\frac{1}{N-1}\left(\sum_{i=1}^{N} \varepsilon_i \varepsilon_i'\right)$$

$$y_i = \bar{y}_N + \varepsilon_i$$

$$\therefore y_i = \sum_{i=1}^{N} \bar{y}_N + \sum_{i=1}^{N} \varepsilon_i \Rightarrow \sum_{i=1}^{N} y_i = N\bar{y}_N + \sum_{i=1}^{N} \varepsilon_i$$

$$\therefore \bar{y}_N = \bar{y}_N + \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \Rightarrow \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i = 0$$

$$\Rightarrow \Sigma \varepsilon_i = 0 \text{ Also } \sum^{N} \varepsilon_i' = 0$$

$$= \frac{N-n}{N}\frac{1}{n} \cdot \rho S_y S_x \qquad \ldots(12)$$

where ρ is the correlation coefficient between y and x, given by

$$\rho = \frac{E\left[(y_i - \bar{y}_N)(x_i - \bar{x}_N)\right]}{\sqrt{E(y_i - \bar{y}_N)^2 \, E(x_i - \bar{x}_N)^2}} \qquad \ldots(13)$$

NOTES

$$\rho = \frac{E(\varepsilon_i \varepsilon_i')}{S_y S_x}$$

$$\Rightarrow \quad \rho S_y S_x = E(\varepsilon_i \varepsilon_i') = \sum_{i=1}^{N} (\varepsilon_i \varepsilon_i') \cdot p = \sum_{i=1}^{N} (\varepsilon_i \varepsilon_i') \frac{1}{N-1}$$

$\because$ probability of selecting any unit from the available units at the second draw is $\frac{1}{N-1}$ as x_i is ancillary characterstic and is selected in IInd draw.

Substituting from (7) and (12) in (11), we get

$$E_1(R_n) = R_n \left[1 + \frac{N-n}{Nn}\left(\frac{S_x^2}{\bar{x}_N^2} - \frac{\rho S_y S_x}{\bar{y}_N \bar{x}_N}\right)\right]$$

$$= R_N \left[1 + \frac{N-n}{Nn}\left(C_x^2 - \rho C_y C_x\right)\right] \quad ...(14)$$

where, $C_x = \frac{S_x}{\bar{x}_N}$ and $C_y = \frac{S_y}{\bar{y}_N}$ ...(15)

Now, $\because$ $\bar{y}_R = R_n \bar{x}_N$ is the ratio estimate of the population mean, it follows that if $E_1(\bar{y}_R)$ denotes a first approximate to the expected value of the ratio estimate of the population mean, then

$$E_1(\bar{y}_R) = \bar{y}_N \left[1 + \frac{N-n}{Nn}\left(C_x^2 - \rho C_y C_x\right)\right] \quad ...(16)$$

$R_N = \frac{\bar{y}_N}{\bar{x}_N}$

Alternatively, Koop has shown that

$$R_n = \frac{\bar{y}_n}{\bar{x}_n} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{N} y_i - \sum_{i}^{N-n} y_i}{\sum_{i=1}^{N} x_i - \sum_{i=1}^{N-n} x_i} = \frac{N\bar{y}_N - (N-n)\bar{y}_{N-n}}{N\bar{x}_N - (N-n)\bar{x}_{N-n}}$$

$$= R_N \left(1 - \frac{N-n}{N}\frac{\bar{y}_{N-n}}{\bar{y}_N}\right)\left(1 - \frac{N-n}{N}\frac{\bar{x}_{N-n}}{\bar{x}_N}\right)^{-1} \quad ...(17)$$

when all the x_i' are of the same sign, $\left|\frac{N-n}{N}\frac{\bar{x}_{N-n}}{\bar{x}_N}\right| < 1$ and the expansion of IInd term on RHS is valid, Taking expectation term by term, we reach the same expression as that of (14).

Denoting to a first approximate the relative bias in a ratio estimate by B_1, We have

$$B_1 = \frac{E(R_n) - R_N}{R_N} = \frac{N-n}{N}\frac{1}{n}(C_x^2 - \rho C_y C_x) \quad ...(18)$$

NOTES

When $C_y = C_x = C$, the expression for B_1 simplifies and we get for large N,

$$B_1 \cong \frac{C^2}{n}(1-\rho) \qquad ...(19)$$

The bias decreases as n increases and it will be seen that the ratio estimate is consistent.

For n large and ρ sufficiently high, the bias will usually be negligible.

The bias vanishes altogether when

$$C_x^2 = \rho\, C_y\, C_x \qquad ...(20)$$

$$i.e.,\ \bar{y}_N = \rho\, \frac{S_y}{S_x}\, \bar{x}_N \qquad ...(21)$$

In other words, the bias to a first approximate vanishes when the regression of y on x is a straight line through the origin. This result is in fact more general. For, let the regression of y on x be represented by

$$E(y/x) = \beta x \qquad ...(22)$$

or
$$y = \beta x + \varepsilon \qquad ...(23)$$

where
$$E(\varepsilon/x) = 0 \qquad ...(24)$$

Summing both sides of (23) overall the units in the population, we obtain

$$\bar{y}_N = \beta \bar{x}_N \quad \Big| \quad \sum^N y_i = \beta \sum^N x_i + \sum^N \varepsilon_i$$

or
$$\beta = R_N \quad \Big| \quad \therefore\ \bar{y}_N = \beta \bar{x}_N + \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i = \beta \bar{x}_N + E(\varepsilon_i) = \beta\ \bar{x}_N + 0$$

Again summing both sides of (23) overall the units in the sample,

We get

$$\bar{y}_n = \beta \bar{x}_n + \bar{\varepsilon}_n$$

where,
$$\bar{\varepsilon}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} \varepsilon_i$$

Thus, we have

$$E(R_n) = E[E(R_n/x_i)] = E\left[E\left(\frac{\bar{y}_n}{\bar{x}_n} / x_i\right)\right]$$

$$= E\left[E\left(\frac{\beta \bar{x}_n + \bar{\varepsilon}_n}{\bar{x}_n} / x_i\right)\right]$$

$$= E\left[E\left(\beta + \frac{\bar{\varepsilon}_n}{\bar{x}_n}\right) / x_i\right]$$

$$= E\left[E(\beta) + E\left(\frac{\bar{\varepsilon}_n}{\bar{x}_n}\right)/x_i\right] = E[\beta + 0] = \beta$$

$$= R_N$$

Thus, $E(R_n) = R_N$

8.4 COMPARISON OF THE RATIO ESTIMATED

By different

$$V(R_n) = \mathrm{E}[R_n - E(R_n)]^2 \qquad ...(1)$$

From equation (9), we write to a first approximation

$$R_n \cong R_N + R_N\left(\frac{\bar{\varepsilon}_n}{\bar{y}_N} - \frac{\bar{\varepsilon}'_n}{\bar{x}_N}\right) + R_N\left(\frac{\bar{\varepsilon}'^2_n}{\bar{x}^2_N} - \frac{\bar{\varepsilon}_n\bar{\varepsilon}'_n}{\bar{y}_N\bar{x}_N}\right) \qquad ...(2)$$

$\because \quad E(R_n) = R_N$

$$\therefore \quad E[R_n - E(R_n)]^2 \cong E\left[R_N + R_N\left(\frac{\bar{\varepsilon}_n}{\bar{y}_N} - \frac{\bar{\varepsilon}'_n}{\bar{x}_N}\right) + R_N\left(\frac{\bar{\varepsilon}'^2_n}{\bar{x}_N{}^2} - \frac{\bar{\varepsilon}_n\bar{\varepsilon}'_n}{\bar{y}_N\bar{x}_N}\right) - R_N\right]^2$$

$$\cong R_N{}^2 E\left(\frac{\bar{\varepsilon}_n}{\bar{y}_N} - \frac{\bar{\varepsilon}'_n}{\bar{x}_N}\right)^2 \qquad (\because \text{ neglecting IIIrd and higher degree terms})$$

$$= R_N^2 E\left[\frac{\bar{\varepsilon}_n{}^2}{\bar{y}_N{}^2} + \frac{\bar{\varepsilon}'^2_n}{\bar{x}_N{}^2} - \frac{2\bar{\varepsilon}_n\bar{\varepsilon}_n}{\bar{y}_N\bar{x}_N}\right] \qquad ...(3)$$

$$= R_N^2\left[\frac{V(\bar{y}_n)}{\bar{y}_N{}^2} + \frac{V(\bar{x}_n)}{\bar{x}_N{}^2} - \frac{2\mathrm{COV}(\bar{y}_n, \bar{x}_n)}{\bar{y}_N\bar{x}_N}\right]$$

$$\bar{y}_n = \bar{y}_N + \bar{\varepsilon}_n$$

$$\bar{\varepsilon}_n{}^2 = (\bar{y}_n - \bar{y}_N)^2$$

$$= \bar{y}_n{}^2 + \bar{y}_N{}^2 - 2\bar{y}_n\ \bar{y}_N$$

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

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

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

$$= R_N^2\frac{N-n}{Nn}\left[\frac{S_y{}^2}{\bar{y}_N{}^2} + \frac{S_x{}^2}{\bar{x}_N{}^2} - \frac{2\rho S_y S_x}{\bar{y}_N\bar{x}_N}\right]$$

NOTES

$$\therefore \quad V_1(R_n) = R_N^2 \frac{N-n}{Nn}\left[C_y^{\,2} + C_x^{\,2} - 2\rho C_y C_x\right] \qquad ...(4)$$

where, V_1 denote the variance of a ratio estimate to a first approximation.

or Relative variance is given by

$$V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn}\left[C_y^{\,2} + C_x^{\,2} - 2\rho C_x C_y\right] \qquad ...(5)$$

When, $C_y = C_x = C$, equation (5) becomes

$$V\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn} 2C^2(1-\rho) \qquad ...(6)$$

For large N,

$$V_1\left(\frac{R_n}{R_N}\right) \cong \frac{2C^2}{n}(1-\rho) = 2 \text{ (relative bias in } B_1) \qquad ...(7)$$

Now, to obtain the variance of the estimate of the mean *i.e.*, $\bar{y}_R$, we have

$$\bar{y}_R = R_n \, \bar{x}_n$$

$$\therefore \quad V_1(\bar{y}_R) = \bar{x}_N^{\,2} V(R_N)$$

$$= \bar{x}_N^{\,2} R_N^{\,2} \frac{N-n}{Nn}\left[C_y^{\,2} + C_x^{\,2} - 2\rho C_y C_x\right] \qquad \text{(from(4)}$$

$$\therefore \quad V_1(\bar{y}_R) = \bar{y}_N^{\,2} \frac{N-n}{Nn}\left[C_y^{\,2} + C_x^{\,2} - 2\rho C_y C_x\right] \qquad ...(8)$$

$$\Big| \quad \bar{x}_N R_N = \bar{y}_N$$

$$= \frac{N-n}{Nn}\left[S_y^{\,2} + R_N^{\,2} S_x^{\,2} - 2\rho R_N S_y S_x\right] \qquad ...(9)$$

$$\Big| \quad C_x = \frac{S_x}{\bar{x}_N} \quad \text{and} \quad C_y = \frac{S_y}{\bar{y}_N}$$

$$= \frac{N-n}{Nn}\left[S_y^{\,2} + R_N^{\,2} S_x^{\,2} - 2R_N S_{yx}\right] \qquad ...(10)$$

where, $S_{yx} = cov.\,(\bar{y}_n, \bar{x}_n)$.

Also, we have

$$B_1 = \frac{N-n}{Nn}\left(C_x^{\,2} - \rho C_y C_x\right) \qquad ...(11)$$

Comparing (8) with (11), we see that both the bias and the variance of a ratio estimate are of the order $\frac{1}{n}$. Hence for n sufficiently large, the bias is negligible as compared with S.D which is of the order $\frac{1}{\sqrt{n}}$.

Also, it can be seen that

$$\frac{S_y^2}{\bar{y}_N^2} + \frac{S_x^2}{\bar{x}_N^2} - \frac{2S_{yx}}{\bar{y}_N \bar{x}_N} = \frac{1}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} (y_i - R_N x_i)^2 \quad ...(12)$$

$$\because \text{ R.H.S} = \frac{1}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} (y_i - R_N x_i)^2$$

$$= \frac{1}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} [(y_i - \bar{y}_N) + (\bar{y}_N - R_N x_i)]^2$$

$$= \frac{1}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} (y_i - \bar{y}_N)^2 + \frac{1}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} (\bar{y}_N - R_N x_i)^2$$

$$+ \frac{2}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} (y_i - \bar{y}_N)(\bar{y}_N - R_N x_v)$$

$$= \frac{S_y^2}{\bar{y}_N^2} + \frac{1}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} \left[\bar{y}_N \left(1 - \frac{x_i}{\bar{x}_N}\right)\right]^2$$

$$+ \frac{2}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} (y_i - \bar{y}_N)\bar{y}_N \left(1 - \frac{x_i}{\bar{x}_N}\right)$$

$$= \frac{S_y^2}{\bar{y}_N^2} + \frac{1}{(N-1)} \frac{1}{\bar{x}_N^2} \sum_{i=1}^{N} (x_i - \bar{x}_N)^2 - \frac{2}{(N-1)\bar{y}_N \bar{x}_N} \sum_{i=1}^{N} (y_i - \bar{y}_N)(x_i - \bar{x}_N)$$

$$= \frac{S_y^2}{\bar{y}_N^2} + \frac{S_x^2}{\bar{x}_N^2} - \frac{2S_{yx}}{\bar{y}_N \bar{x}_N} = \text{L.H.S.}$$

It follows from (4) that

$$V_1(R_n) = \frac{N-n}{N-1} \frac{1}{n} \frac{1}{\bar{x}_N^2} \frac{1}{N} \sum_{i=1}^{N} (y_i - R_N x_i)^2 \quad ...(13)$$

If the population is regarded as divided into k classes with the N_i units in the i^{th} class having the value x_i each, the variance given by (13) becomes

$$V_1(R_n) = \frac{N-n}{N-1} \frac{1}{n} \frac{1}{\bar{x}_N^2} \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - R_N x_i)^2 \quad ...(14)$$

Obviously, the term under the summation sign is proportional to the variance of y for a fixed value of x when the regression of y on x is linear and the regression line passes through the origin.

NOTES

$$\therefore \quad E(y_{ij}/x_i) = R_N x_i \qquad \left| \; E(y/x) = \beta x = R_N x \;\because\; \beta = R_N \right.$$

$$\Rightarrow \quad \sum_{j=1}^{N_i} (y_{ij} - R_N x_i)^2 = \sum_{j=1}^{N_i} \left[y_{ij} - E(y_{ij}/x_i) \right]^2 = N_i \, V(y_{ij}/x_i)$$

Thus,

$$V_1(R_n) = \frac{N-n}{N-1} \frac{1}{n} \frac{1}{\bar{x}_N^2} \frac{1}{N} \sum_{i=1}^{k} N_i V(y_{ij}/x_i)$$

$$V_1(\bar{y}_R) = \frac{N-n}{nN(N-1)} \sum_{i=1}^{k} N_i V(y_{ij}/x_i) \qquad ...(15)$$

Thus, we see that variance of a ratio estimate depends upon the relationship between the variance of y for a fixed x and the value of x.

In usual practice, we take

$$(i) \quad V(y/x) = \text{constant} = \gamma$$

$$(ii) \quad V(y/x) = \gamma x \qquad ...(16)$$

$$(iii) \quad V(y/x) = \gamma x^2$$

Thus, the approximate expressions for the variance of ratio estimate becomes

$$(i) \quad V_1(\bar{y}_R) = \frac{N-n}{N-1} \cdot \frac{\gamma}{n} \qquad \left| \; \sum_{i=1}^{k} N_i = N \right.$$

$$(ii) \quad V_1(\bar{y}_R) = \frac{N-n}{N-1} \frac{\gamma}{n} \bar{x}_N \qquad \left| \; \sum_{i=1}^{k} N_i V(y/x) = \sum_{i=1}^{k} N_i V x_i = VN. \frac{1}{N} \sum N_i x_i \right.$$

$$(iii) \quad V_1(\bar{y}_R) = \frac{N-n}{N(N-1)} \frac{\gamma}{n} \sum_{i=1}^{k} N_i x_i^2 \qquad \left| \; = VN. \; \bar{x}_N \right.$$

8.5 ESTIMATE OF THE VARIANCE OF THE RATIO ESTIMATE

We know that $\bar{y}_n, \bar{x}_n, s_y^2, s_x^2$ are unbiased estimates of the corresponding population values.

Similarly,

$$s_{yx} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{n-1}$$

is an unbiased estimate of the corresponding population value

$$S_{yx} = \rho S_y S_x.$$

The consistant estimate of the relative variable of ratio estimate in given by

$$\text{Est. } V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn}\left[\frac{s_y^2}{\bar{y}_n^2} + \frac{s_x^2}{\bar{x}_n^2} - \frac{2s_{yx}}{\bar{y}_n \bar{x}_n}\right]$$

$$= \frac{N-n}{Nn}\frac{1}{\bar{y}_n^2}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - R_n x_i)^2$$

Thus, the estimate of the variances of R_n and $\bar{y}_R$ are

$$\text{Est. } V_1(R_n) = \frac{N-n}{Nn}\frac{1}{\bar{x}_n^2}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - R_n x_i)^2$$

and

$$\text{Est. } V_1(\bar{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - R_n x_i)^2 \text{ respectively}$$

These are biased estimates but the bias will be negligible if the coefficient of variation of y and of x are small *i.e.*, $\frac{s_y}{\bar{y}_n}$ and $\frac{s_x}{\bar{x}_n}$ are small.

Particular Case . It is the case of a weighted mean in which the weights are in the nature of ancillary information, varying from one sampling unit to another, with y_i of the form $w_i \eta_i$ and $x_i = w_i$

$$\therefore \qquad R_n = \bar{\eta}_w = \frac{\sum_{}^{n} w_i \eta_i}{\sum_{}^{n} w_i}$$

The sample estimate of the variance for this case is given by

$$\text{Est. } V_1(\bar{\eta}_w) = \frac{N-n}{Nn}\frac{1}{\bar{w}_n^2}\frac{1}{n-1}\sum_{i=1}^{n} w_i^2(\eta_i - \bar{\eta}_w)^2$$

and

$$\text{Est. } V_1(\bar{y}_R) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n} w_i^2(\eta_i - \bar{\eta}_w)^2$$

8.6 CONDITIONS FOR OPTIMUM RATIO ESTIMATE

We assume that (i) the population of size N can be regarded as divided into k classes with N_i units in the i^{th} class having the value x_i each with

NOTES

$\sum_{i=1}^{k} N_i = N$ and that in repeated samples of size n drawn from the population with equal probability and without replacement, the number of units having the value x_i is fixed, say n_i, $i = 1, 2, \ldots, k$.

(*ii*) N_i are sufficiently large so that $\frac{N_i - 1}{N_i} \cong 1$ and $\frac{N_i - n_i}{n_i - 1} \cong 1$.

(*iii*) $E\ (y_{ij}/x_i) = \beta x_i$ and

(*iv*) $V\ (y_{ij}/x_i) = \gamma x_i$

Now, let us consider a subclass of linear estimates of the type $T = \sum_{i=1}^{k} \lambda_i n_i \bar{y}_{n_i}$ where all the y observations in the i^{th} class are assumed to have the same weight λ_i.

We shall prove that the ratio estimate in this subclass is optimum *i.e.*, it is unbiased and has minimum variance.

Since, the estimate is to be unbiased, we have

$$E[T/n_1, n_2, \ldots, n_k] = E\left[\sum_{i=1}^{k} \lambda_i n_i \bar{y}_{ni} | n_1, n_2 \ldots, n_k\right]$$

$$= \sum_{i=1}^{k} \lambda_i n_i E[\bar{y}_{ni}] = \sum_{i=1}^{k} \lambda_i n_i \bar{y}_{ni} \qquad (\because\ E(\bar{y}_n) = \bar{y}_n)$$

$$= \sum_{i=1}^{k} \lambda_i\ n_i\ E(y_{ij}\ /\ x_i)$$

$$= \sum_{i=1}^{k} \lambda_i\ n_i\ \beta x_i \qquad \ldots(1)$$

$$= \bar{y}_N \qquad (\because\ T \text{ is unbiased } \therefore\ E(T) = \bar{y}_N)$$

$$= E(\bar{y}_n) = \sum_{i=1}^{k} \bar{y}_{N_i} \cdot p = \sum_{i=1}^{k} \frac{N_i}{N} \bar{y}_{N_i}$$

$$= \sum_{i=1}^{k} \frac{N_i}{N} \beta x_i \qquad \ldots(2)$$

From (1) and (2), we have

$$\sum_{i=1}^{k} \left(n_i\ \lambda_i - \frac{N_i}{N}\right) x_i = 0 \qquad \ldots(3)$$

Now, the variance of T for fixed values of $n_1, n_2, \ldots, n_k$ is given by

$$V(T/n_1, n_2, \ldots, n_k) = \sum_{i=1}^{k} \lambda_i^{\ 2}\ n_i^{\ 2} V(\bar{y}_{n_i}\ /\ n_i)$$

$$= \sum_{i=1}^{k} \lambda_i^2 n_i \frac{N_i - n_i}{N_i - 1} \gamma x_i$$

$$\cong \gamma \sum_{i=1}^{k} \lambda_i^2 n_i x_i \quad ...(4)$$

$$V(\bar{y}_{ni}) = \frac{N_i - n_i}{N - 1} \frac{S^2}{n_i}$$

where S^2 = population variable.

$$= \frac{N_i - n_i}{N - 1} \frac{V(y_{ij})}{n_i}$$

We shall now determine the λ_i so that the variance given by (4) is minimum subject to the condition (3)

Consider

$$\phi = \gamma \sum_{i=1}^{k} \lambda_i^2 n_i x_i - m \sum_{i=1}^{k} \left[\left(n_i \lambda_i - \frac{N_i}{N} \right) x_i \right]$$

where μ is some constant

or
$$\phi = \sum_{i=1}^{k} n_i x_i \left[\lambda_i \sqrt{\gamma} - \frac{\mu}{2\sqrt{\gamma}} \right]^2 + (\text{terms independent of } \lambda_i) \quad ...(5)$$

Then ϕ, is minimum when $\lambda_i = \frac{\mu}{2\gamma}$; $i = 1, 2, ..., k$ ...(6)

$$\frac{d\phi}{d\lambda_i} = 2 \sum n_i x_i \left[\lambda_i \sqrt{\gamma} - \frac{\mu}{2\gamma} \right] \sqrt{\gamma}$$

$$= 0 \Rightarrow \lambda_i = \frac{\mu}{2\gamma}$$

Also $\frac{d^2\phi}{d\lambda_i^2} > 0.$

To evaluate μ, we have from (3),

$$\mu = 2\gamma \frac{\bar{x}_N}{n\bar{x}_n} \quad ...(7)$$

$$\frac{\mu}{2\gamma} \sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \frac{N_i}{N} x_i$$

$$\Rightarrow \quad \frac{\mu}{2\gamma} n \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i} = \frac{1}{N} \sum_{i=1}^{k} N_i x_i \qquad (\because n = \sum_{i=1}^{k} n_i)$$

$$\Rightarrow \quad \frac{\mu}{2\gamma} n . \bar{x}_n = \bar{x}_N \qquad (\because \sum_{i=1}^{k} N_i = N)$$

$$\Rightarrow \quad \mu = 2\gamma \frac{\bar{x}_N}{n\bar{x}_n}))$$

NOTES

Thus, from (6), we have

$$\lambda_i = \frac{\bar{x}_N}{n\bar{x}_n} \qquad ...(8)$$

Using the optimum value of λ_i given by (8), we see that the estimate T reduces to the usual ratio estimate, namely,

$$\bar{y}_R = \frac{\bar{y}_n}{\bar{x}_n}\bar{x}_N$$

with minimum variance given by

$$V(\bar{y}_R / n_1, n_2, ..., n_k) = \gamma \sum_{i=1}^{k} \lambda_i^2 n_i x_i = \gamma \sum_{i=1}^{k} \frac{\bar{x}_N^2}{n^2 \bar{x}_n^2} n_i x_i$$

$$= \gamma \frac{\bar{x}_N^2}{n} \frac{1}{\bar{x}_n^2} \frac{\sum_{i=1}^{k} n_i x_i}{n}$$

$$= \gamma \frac{\bar{x}_N^2}{n} \frac{1}{\bar{x}_n^2} . \bar{x}_n \qquad (\because n = \sum n_i)$$

$$= \gamma \frac{\bar{x}_N^2}{n \bar{x}_n} \qquad ...(9)$$

showing that the ratio estimate under consideration is unbiased and has minimum variance.

8.7 EFFICIENCY OF THE RATIO ESTIMATE

We know that variance of estimate of population mean based on A.M is given by

$$V(\bar{y}_n) = \frac{N-n}{N} \frac{S_y^2}{n}.$$

Also, we know that variance to the first approximation of the estimate of population mean based on the ratio method is

$$V(\bar{y}_R) = \frac{N-n}{Nn}\left[S_y^2 + R_N^2 S_x^2 - 2R_N \rho S_y S_x\right]$$

Now, the relative efficiency of an estimate B compared to that of another estimate A based on a sample of equal size is defined as ratio of the inverse of their variances.

$$i.e., \quad E = \frac{V(A)}{V(B)}$$

Thus, $$\text{Efficiency} = \frac{S_y^2}{S_y^2 + R_N^2 S_x^2 - 2\rho R_N S_y S_x}$$

$$= \frac{1}{1 + \left(\frac{C_x^2}{C_y}\right) - 2\rho\left(\frac{C_x}{C_y}\right)}$$

$R_N = \frac{\bar{y}_N}{x_{\bar{N}}}$

$C_y^2 = S_y^2 / \bar{y}_N^2$

$C_x^2 = S_x^2 / \bar{x}_N^2$)

It follows that, in large samples, the ratio estimate will be more efficient than the corresponding sample estimate based on the simple A.M if the denominator is less than 1.

i.e., $$\left(\frac{C_x}{C_y}\right)^2 < 2\rho\left(\frac{C_x}{C_y}\right)$$

i.e., $$\rho\frac{C_y}{C_x} > \frac{1}{2}.$$

In case $C_y = C_x$, then $\rho > \frac{1}{2}$.

8.8 UNBIASED RATIO TYPE ESTIMATE

Another ratio–type estimate of the population mean may be defined as

$$\bar{y}'_R = \bar{r}_n \bar{x}_N \qquad ...(1)$$

where $$\bar{r}_n = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}\sum\frac{y_i}{x_i} \qquad ...(2)$$

Now, $$E(\bar{y}'_R) = \bar{x}_N E(\bar{r}_n) = \bar{x}_N \bar{r}_N \neq \bar{y}_N$$

$\Rightarrow$ the estimate $\bar{y}'_R$ is biased.

Its bias is given by

$$\text{Bias in } \bar{y}'_R = \bar{r}_N \bar{x}_N - \bar{y}_N$$

$$= -\frac{1}{N}\left[\sum_{i=1}^{N} y_i - N\bar{r}_N \bar{x}_N\right]$$

$$= -\frac{1}{N}\left[\sum_{i=1}^{N} r_i x_i - N\bar{r}_N \bar{x}_N\right] \qquad ((\because r_i = \frac{y_i}{x_i}))$$

NOTES

$$= -\frac{N-1}{N} S_{rx} \qquad \ldots(3)$$

$$S_{rx} = \frac{1}{N-1}\sum_{i=1}^{N}(r_i - \bar{r}_N)(x_i - \bar{x}_N)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}[(r_i x_i - \bar{r}_N\, x_i - \bar{x}_N r_i + \bar{x}_N \bar{r}_N)]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} r_i x_i - \bar{r}_N\left(\sum_{i=1}^{N} x_i\right) - \bar{x}_N\left(\sum_{i=1}^{N} r_i\right) + N\,\bar{x}_N\,\bar{r}_N\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} r_i x_i - \bar{r}_N\, N\,\bar{x}_N - \bar{x}_N N\,\bar{r}_N + N\,\bar{x}_N\,\bar{r}_N\right]$$

$$\therefore \quad (N-1)\, S_{rx} = \sum_{i=1}^{N} r_i x_i - N\,\bar{r}_N\,\bar{x}_N$$

Since, $s_{rx} = \dfrac{1}{n-1}\sum_{i=1}^{N}(r_i - \bar{r}_n)(x_i - \bar{x}_n)$ is an unbiased estimate of S_{rx}, it follows that an unbiased ratio–type estimate of the population mean is given by

$$\bar{y}_R'' = \bar{r}_n \bar{x}_N + \frac{N-1}{N} s_{rx} = \bar{r}_n \bar{x}_N + \frac{n(N-1)}{N(n-1)}[\bar{y}_n - \bar{r}_n \bar{x}_n] \qquad \ldots(4)$$

$$s_{rx} = \frac{1}{n-1}\left[\sum^{n} r_i x_i - n\bar{r}_n \bar{x}_n\right]$$

$$= \frac{1}{n-1}\left[\sum^{n} y_i - n\,\bar{r}_n \bar{x}_n\right] = \frac{n}{n-1}(\bar{y}_n - \bar{r}_n \bar{x}_n)$$

which is the result given by Hartley and Ross.

Now we know that the variance of $\bar{y}_R$ is given by

$$V(\bar{y}_R) = \frac{N-n}{Nn}\left\{S_y^{\,2} + R_N^{\,2}\, S_x^{\,2} - 2\, R_N\, S_{yx}\right\}$$

Thus for large N, we have

$$V(\bar{y}_R) \cong \frac{1}{n}\left\{S_y^{\,2} + R_N^{\,2}\, S_n^{\,2} - 2\, R_N\, S_{yx}\right\} \qquad \ldots(5)$$

Similarly, we can prove that the variance of $\bar{y}_R''$ for large sample, is

$$V(\bar{y}_R'') \cong \frac{1}{n}\left\{S_y^{\,2} + \bar{r}_N^{\,2}\, S_x^{\,2} - 2\,\bar{r}_N\, S_{yx}\right\} \qquad \ldots(6)$$

NOTES

Thus,

$$V(\bar{y}_R) - V(\bar{y}_R'') = \frac{S_x^2}{n}\left[(R_N - \beta)^2 - (\bar{r}_N - \beta)^2\right] \quad ...(7)$$

where, $\beta = \frac{S_{yx}}{S_x^2}$ is the regression coefficient of y on x.

From equation (7) we see that for large sample, $\bar{y}_R''$ will be more efficient then $\bar{y}_R$ if the regression co-efficient β is nearer to $\bar{r}_N$ them to $\bar{R}_N$.

In that case $V(\bar{y}_R) - V(\bar{y}_R'') > 0 \Rightarrow V(\bar{y}_R'') < V(\bar{y}_R)$

If $\bar{r}_N = R_N$, then the two variances are equal.

Now we give a method for reducing the bias in ratio estimate.

Let a random sample of size $(2n)$ be split at random into two sub-classes each of size n. Denote by $\bar{y}_R^{(1)}, \bar{y}_R^{(2)}$ and $\bar{y}_R^{(3)}$ the usual ratio estimates based on the sub-classes of size n and the entire sample of size $(2n)$ respectively. Consider the weighted estimate

$$\bar{y}_w = w_1\bar{y}_R^{(1)} + w_2\bar{y}_R^{(2)} + (1 - w_1 - w_2)\,\bar{y}_R^{(3)} \quad ...(8)$$

Since, $\bar{y}_R^{(1)}$ and $\bar{y}_R^{(2)}$ are based on the same sample size, it follows that

$$w_1 = w_2 = w$$

Thus,
$$\bar{y}_w = w\,\bar{y}_R^{(1)} + w\,\bar{y}_R^{(2)} + (1 - 2w)\,\bar{y}_R^{(3)} \quad ...(9)$$

Now we shall determine the weight w so that the bias in the estimate $\bar{y}_w$ to the Ist degree of approximate is zero.

Now we know that $E_1(\bar{y}_R)$ *i.e,* first approximate, to the expected value of the ratio estimate of the population mean is

$$E_1(\bar{y}_R) = \bar{y}_N\left[1 + \frac{N-n}{Nn}\left(C_x^2 - \rho\, C_y C_x\right)\right]$$

Thus,
$$E_1\left[\bar{y}_R^{(1)}\right] = E_1\left[\bar{y}_R^{(2)}\right] = \bar{y}_N\left[1 + \frac{N-n}{Nn}\left(C_x^2 - \rho\, C_y C_x\right)\right] \quad ...(10)$$

and
$$E_1\left[\bar{y}_R^{(3)}\right] = \bar{y}_N\left[1 + \frac{N-2n}{2Nn}\left(C_x^2 - \rho\, C_y C_x\right)\right] \quad ...(11)$$

Thus,
$$E_1[\bar{y}_w] = w\,E(\bar{y}_R^{(1)}) + w\,E(\bar{y}_R^{(2)}) + (1 - w2)\,E(\bar{y}_R^{(3)})$$

$$= \bar{y}_N + \bar{y}_N\left(C_x^2 - \rho\, C_y C_x\right)\left[\frac{2w\,(N-n)}{Nn} + \frac{(1-2w)\,(N-2n)}{2Nn}\right] \quad ...(12)$$

NOTES

The bias in the estimate $\bar{y}_w$, to the first degree of approximation will be zero if

$$\frac{2w(N-n)}{Nn} + \frac{(1-2w)(N-2n)}{2Nn} = 0$$

$$\Rightarrow \quad w = -\frac{N-2n}{2N} \qquad ...(13)$$

Hence an almost unbiased ratio type estimate, to the Ist degree of approximation is given by

$$\bar{y}_w = \frac{(2N-2n)}{N}\bar{y}_R^{(3)} - \frac{(N-2n)}{2N}\bar{y}_R^{(1)} - \frac{(N-2n)}{2N}\bar{y}_R^{(2)}$$

It finite correction factors can be considered to be negligible, then

$$\bar{y}_w \cong 2\bar{y}_R^{(3)} - \frac{1}{2}\bar{y}_R^{(1)} - \frac{1}{2}\bar{y}_R^{(2)} \qquad ...(14)$$

Now, we obtain M.S.E. of the estimate $\bar{y}_w$:

$$\text{M.S.E}(\bar{y}_w) = E[\bar{y}_w - \bar{y}_N]^2$$

$$\cong E\left[2(\bar{y}_R^{(3)} - \bar{y}_N) - \frac{1}{2}(\bar{y}_R^{(1)} - \bar{y}_N) - \frac{1}{2}(\bar{y}_R^{(2)} - \bar{y}_N)\right]^2$$

$$= 4E(\bar{y}_R^{(3)} - \bar{y}_N)^2 + \frac{1}{4}E(\bar{y}_R^{(1)} - \bar{y}_N)^2 + \frac{1}{4}E(\bar{y}_R^{(2)} - \bar{y}_N)^2$$

$$-2E\left\{(\bar{y}_R^{(3)} - \bar{y}_N)(\bar{y}_R^{(1)} - \bar{y}_N)\right\} + \frac{1}{2}E\left\{(\bar{y}_R^{(1)} - \bar{y}_N)(\bar{y}_R^{(2)} - \bar{y}_N)\right.$$

$$\left. - 2E\left\{(\bar{y}_R^{(3)} - \bar{y}_N)(\bar{y}_R^{(2)} - \bar{y}_N)\right\}\right. . \qquad ...(15)$$

Now, we have to the Ist degree of approximation,

$$E[\bar{y}_R - \bar{y}_N]^2 = V(\bar{y}_R) = \frac{N-n}{Nn}\bar{y}_N^2(C_y^2 + C_x^2 - 2\rho C_y C_x)$$

Thus,

$$E[\bar{y}_R^{(1)} - \bar{y}_N]^2 = E[\bar{y}_R^{(2)} - \bar{y}_N]^2 \cong \frac{N-n}{Nn}\bar{y}_N^2(C_y^2 + C_x^2 - 2\rho C_y C_x) \qquad ...(16)$$

and

$$E[\bar{y}_R^{(3)} - \bar{y}_N]^2 \cong \frac{N-2n}{2Nn}\bar{y}_N^2(C_y^2 + C_x^2 - 2\rho C_y C_x) \qquad ...(17)$$

Now, we will show that, to the same degree of approximation,

$$E[(\bar{y}_R^{(3)} - \bar{y}_N)(\bar{y}_R^{(1)} - \bar{y}_N)] = E[(\bar{y}_R^{(3)} - \bar{y}_N)(\bar{y}_R^{(2)} - \bar{y}_N)]$$

$$\cong \frac{N-2n}{N}\frac{\bar{y}_N^2}{2n}(C_y^2 + C_x^2 - 2\rho C_x C_y) \qquad ...(18)$$

and

$$E\left[(\bar{y}_R^{(1)} - \bar{y}_N)(\bar{y}_R^{(2)} - \bar{y}_N)\right] \cong -\frac{\bar{y}_N^2}{N}\left[C_y^2 + C_x^2 - 2\rho C_x C_y\right] \quad ...(19)$$

NOTES

For this purpose, let

$$\begin{aligned} \bar{y}_n^{(i)} &= \bar{y}_N + \varepsilon_i \\ \bar{x}_n^{(i)} &= \bar{x}_N + \varepsilon_i', \quad i = 1, 2. \\ \bar{y}_{2n} &= \bar{y}_N + \varepsilon_3 \\ \bar{x}_{2n} &= \bar{x}_N + \varepsilon_3' \end{aligned} \quad ...(20)$$

where $\bar{y}_n^{(1)}$, $\bar{y}_n^{(2)}$ are the means of the y observations based on the two sub–samples and $\bar{x}_n^{(1)}, \bar{x}_n^{(2)}$ the corresponding means of the x observation, and

$$\begin{aligned} E(\varepsilon_1) &= E(\varepsilon_2) = E(\varepsilon_3) = 0 \\ E(\varepsilon_1') &= E(\varepsilon_2') = E(\varepsilon_3') = 0 \end{aligned} \quad ...(21)$$

Now, $$\bar{y}_n^{(i)} = \bar{y}_N + \varepsilon_i = \bar{y}_N\left(1 + \frac{\varepsilon_i}{\bar{y}_N}\right)$$

and $$\bar{x}_n^{(i)} = \bar{x}_N + \varepsilon_i' = \bar{x}_N\left(1 + \frac{\varepsilon_i'}{\bar{x}_N}\right)$$

$$\therefore \quad R_n = \frac{\bar{y}_n^{(i)}}{\bar{x}_n^{(i)}} = \frac{\bar{y}_N}{\bar{x}_N}\left(1 + \frac{\varepsilon_i}{\bar{y}_N}\right)\left(1 + \frac{\varepsilon_i'}{\bar{x}_N}\right)^{-1}$$

$$\because \bar{y}_R = R_n \bar{x}_N$$

$$\therefore \quad \bar{y}_R^{(i)} = R_n \bar{x}_N = \bar{y}_N\left(1 + \frac{\varepsilon_i}{\bar{y}_N}\right)\left(1 + \frac{\varepsilon_i'}{\bar{x}_N}\right)^{-1}$$

$$\therefore \quad \bar{y}_R^{(i)} - \bar{y}_N = \bar{y}_N\left(1 + \frac{\varepsilon_i}{\bar{y}_N}\right)\left(1 + \frac{\varepsilon_i'}{\bar{x}_N}\right)^{-1} - \bar{y}_N$$

Assuming $\left(\frac{\varepsilon_i'}{\bar{x}_N}\right) < 1$, $i = 1, 2, 3, \ldots$ we get

$$\bar{y}_R^{(i)} - \bar{y}_N = \bar{y}_N\left[\frac{\varepsilon_i}{\bar{y}_N} - \frac{\varepsilon_i'}{\bar{x}_N} + \frac{\varepsilon_i'^2}{\bar{x}_N^2} - \frac{\varepsilon_i \varepsilon_i'}{\bar{y}_N \bar{x}_N} + \ldots\right] \quad ...(22)$$

Using (22) and ignoring terms in ε of order higher than two, we have on taking expectation,

NOTES

$$E\left[(\bar{y}_R^{(i)} - \bar{y}_N)(\bar{y}_R^{(3)} - \bar{y}_N)\right]$$

$$= E\left[\bar{y}_N\left\{\frac{\varepsilon_i}{\bar{y}_N} - \frac{\varepsilon_i'}{\bar{x}_N} + \frac{\varepsilon_i'^2}{\bar{x}_N^2} - \frac{\varepsilon_i\varepsilon_i'}{\bar{y}_N\bar{x}_N}\right\}\ \bar{y}_N\left\{\frac{\varepsilon_3}{\bar{y}_N} - \frac{\varepsilon_3'}{\bar{x}_N} + \frac{\varepsilon_3'^2}{\bar{x}_N^2} - \frac{\varepsilon_3\varepsilon_3'}{\bar{y}_N\bar{x}_N} + \ldots\right\}\right]$$

$$= \bar{y}_N^2 E\left[\frac{\varepsilon_i\varepsilon_3}{\bar{y}_N^2} + \frac{\varepsilon_i'\varepsilon_3'}{\bar{x}_N^2} - \frac{\varepsilon_i\varepsilon_3'}{\bar{x}_N\bar{y}_N} - \frac{\varepsilon_i'\varepsilon_3}{\bar{x}_N\bar{y}_N}\right]$$

$$= \bar{y}_N^2\left[\frac{cov.(\bar{y}_n^{(i)}, \bar{y}_{2n})}{\bar{y}_N^2} + \frac{cov.(\bar{x}_n^{(i)}, \bar{x}_{2n})}{\bar{x}_N^2}\right.$$

$$\left. - \frac{cov.(\bar{y}_n^{(i)}, \bar{x}_{2n})}{\bar{x}_N\bar{y}_N} - \frac{cov.(\bar{x}_n^{(i)}, \bar{y}_{2n})}{\bar{x}_N\bar{y}_N}\right] \quad ; i = 1, 2. \qquad ...(23)$$

Now, $cov.\ (\bar{y}_n^{(i)}, \bar{y}_{2n}) = E\left[\bar{y}_n^{(i)}\ \bar{y}_{2n}\right] - E(\bar{y}_n^{(i)})\ E(\bar{y}_{2n})$

$$= E\left[E\ (\bar{y}_n^{(i)}\ \bar{y}_{2n} / 2n)\right] - \bar{y}_n^2$$

$$= E\left[\bar{y}_{2n}\ \bar{y}_{2n}\right] - \bar{y}_n^2$$

$$= E\ (\bar{y}_{2n}^2) - (E(\bar{y}_{2n}))^2$$

$$= V\ (\bar{y}_{2n}) = \left(\frac{1}{2n} - \frac{1}{N}\right) S_y^2 \qquad ...(24)$$

$$\left| \begin{array}{l} (\because\ V\ (\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2) \\ \qquad i = 1,2. \end{array} \right.$$

Similarly,

$$cov.\left(\bar{x}_n^{(i)}, \bar{x}_{2n}\right) = \left(\frac{1}{2n} - \frac{1}{N}\right) S_x^2 \ ; i = 1, 2.$$

$$cov.\left(\bar{y}_n^{(i)}, \bar{x}_{2n}\right) = \left(\frac{1}{2n} - \frac{1}{N}\right) S_{yx} \ ; i = 1, 2. \qquad ...(25)$$

$$cov.\left(\bar{x}_n^{(i)}, \bar{y}_{2n}\right) = \left(\frac{1}{2n} - \frac{1}{N}\right) S_{yx} \ ; i = 1, 2.$$

Using (24), (25) in (23), we get (18).

Proceeding in a similar way and ignoring the terms in ε of order higher than two, we have, on taking expectation,

$$E\left[(\bar{y}_R^{(1)} - \bar{y}_N)(\bar{y}_R^{(2)} - \bar{y}_N)\right] = \bar{y}_N^2\left[\frac{cov.(\bar{y}_n^{(1)}, \bar{y}_n^{(2)})}{\bar{y}_N^2} + \frac{cov.(\bar{x}_n^{(1)}, \bar{x}_n^{(2)})}{\bar{x}_N^2}\right.$$

$$\left. - \frac{cov.(\bar{y}_n^{(1)}, \bar{x}_n^{(2)})}{\bar{y}_N\ \bar{x}_N} - \frac{cov.(\bar{x}_n^{(1)}, \bar{y}_n^{(2)})}{\bar{y}_N\ \bar{x}_N}\right] \qquad ...(26)$$

Now $cov.(\bar{y}_n^{(1)}, \bar{x}_n^{(2)}) = E\left[\bar{y}_n^{(1)}, \bar{x}_n^{(2)}\right] - E\left[\bar{y}_n^{(1)}\right] E\left[\bar{x}_n^{(2)}\right]$

$$= E\left[\bar{y}_n^{(1)}\left\{\frac{2n\,\bar{x}_{2n} - n\,\bar{x}_n^{(1)}}{n}\right\}\right] - \bar{y}_N\,\bar{x}_N$$

$$\bar{x}_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{x}_n^{(2)} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{x}_{2n} = \frac{1}{2n}\sum_{i=1}^{2n} x_i$$

$$\therefore \quad \sum_{i=1}^{2n} x_i - \sum_{i=1}^{n} x_i = 2n\bar{x}_{2n} - n\bar{x}_n^{(1)}$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i = 2n\bar{x}_{2n} - n\bar{x}_n^{(1)} \Rightarrow \bar{x}_n^{(2)} = \frac{2n\bar{x}_{2n} - n\bar{x}_n'}{n}$$

$$\therefore \quad cov.(\bar{y}_n^{(1)}, \bar{x}_n^{(2)}) = 2E\left[\bar{y}_n^{(1)}\bar{x}_{2n}\right] - E\left[\bar{y}_n^{(1)}\bar{x}_n^{(1)}\right] - \bar{x}_N\bar{y}_N$$

$$= 2E\left[\bar{y}_n^{(1)}\,\bar{x}_{2n}\right] - 2\,\bar{x}_N\bar{y}_N - E\left[\bar{y}_n^{(1)}\bar{x}_n^{(1)}\right] + \bar{x}_N\bar{y}_N$$

$$= 2cov.\left[\bar{y}_n^{(1)}\,\bar{x}_{2n}\right] - cov.\left(\bar{y}_n^{(1)}, \bar{x}_n^{(1)}\right) \qquad ...(27)$$

Using (25), (27) gives

$$cov.(\bar{y}_n^{(1)}, \bar{x}_n^{(2)}) = 2\left(\frac{1}{2n} - \frac{1}{N}\right)S_{yx} - \left(\frac{1}{n} - \frac{1}{N}\right)S_{yx} = -\frac{1}{N}S_{yx} \qquad ...(28)$$

Similarly,

$$cov.(\bar{x}_n^{(1)}, \bar{y}_n^{(2)}) = -\frac{1}{N}S_{yx} \qquad ...(29)$$

Putting $x = y$ in (29), we have

$$cov.(\bar{x}_n^{(1)}, \bar{y}_n^{(2)}) = -\frac{1}{N}S_y^{\,2}$$

$$cov.(\bar{x}_n^{(1)}, \bar{y}_n^{(2)}) = -\frac{1}{N}S_x^{\,2} \qquad ...(30)$$

Using (28), (29), (30) in (26), are get (19).

Finally using (16), (17), (18), (19), in (15), we get

$$\text{M.S.E. } (\bar{y}_w) \cong \left(\frac{1}{2n} - \frac{1}{N}\right)\bar{y}_N^{\,2}\left[C_y^{\,2} + C_x^{\,2} - 2\rho\,C_x\,C_y\right] \qquad ...(31)$$

Also, to same degree of approximation

$$\text{M.S.E. } (\bar{y}_R^{(3)}) \cong \left(\frac{1}{2n} - \frac{1}{N}\right)\bar{y}_N^{\,2}\left[C_y^{\,2} + C_x^{\,2} - 2\rho\,C_x\,C_y\right] \qquad ...(32)$$

Since, the estimate $\bar{y}_w$ is almost unbiased to the Ist degree of approximation, while $\bar{y}_R^{(3)}$ is not, the former may be preferred to the later.

NOTES

8.9 RATIO ESTIMATE IN STRATIFIED SAMPLING

Let N_i denote the number of units in the t-th stratum and n_i be the sample size to be selected therefrom, so that

$$\sum_{t=1}^{k} N_t = N \quad \text{and} \quad \sum_{t=1}^{k} n_t = n \qquad ...(1)$$

Let us denote by R_{nt}, the estimate of the population ratio given by

$$R_{Nt} = \frac{\bar{y}_{Nt}}{\bar{x}_{Nt}}$$

and by $\bar{y}_{Rt}$, the ratio estimate of the population mean $\bar{y}_{Nt}$ for the t-th stratum.

Then the ratio estimate of the population mean

$$\bar{y}_N = \sum_{t=1}^{k} \frac{Nt}{N} \quad \bar{y}_{Nt} = \sum_{t=1}^{k} pt\, \bar{y}_{Nt} \qquad ...(2)$$

is given by

$$\bar{y}_{Rs} = \sum_{t=1}^{k} \frac{Nt}{N} \quad \bar{y}_{Rt} = \sum_{t=1}^{k} pt\, \bar{y}_{Rt} \qquad ...(3)$$

where $pt = \frac{N_t}{N}$; $t = 1, 2, ..., k$

Now, we know that the ratio estimate to the first approximation is given by

$$E_1(\bar{y}_{Rs}) = \sum_{t=1}^{k} p_t\, \bar{y}_{Nt} \left\{1 + \frac{N_t - n_t}{N_t n_t}\left(C_{tx}^2 - \rho_t C_{tx} C_{ty}\right)\right\} \qquad ...(4)$$

where $$C_{tx} = \frac{S_{tx}}{\bar{x}_{Nt}} \quad \text{and} \quad C_{ty} = \frac{S_{ty}}{\bar{y}_{Nt}} \qquad ...(5)$$

It follows that $\bar{Y}_{Rs}$ is a biased but consistent estimate of the population mean $\bar{Y}_N$. To obtain an idea of how the bias diminishes with the sample size, we assume that the finite multiplier in each stratum approximates to unity, $n_t = \frac{n}{k}$ and that C_{tx}, C_{ty} and ρ_t are the same over all strata, say, C_x, C_y and ρ respectively. Thus, the relative bias in the estimate $\bar{y}_{Rs}$ is given by

$$\text{Bias} = \frac{k}{n}\left(C_x^2 - \rho\, C_x C_y\right) \qquad ...(6)$$

Now, in order that $\bar{y}_{Rs}$ should provide a satisfactory estimate of the population mean, the sample size in each stratum should be sufficiently large.

NOTES

We know that the variance of ratio estimate to the first approximation is given by

$$V_1(\bar{y}_R) = \frac{N-n}{Nn}\left(S_y^2 + R_N^2 S_x^2 - 2\rho R_N S_y S_x\right) \quad ...(7)$$

Using equation (7), the variance of $\bar{y}_{Rs}$ to the first approximation is given by

$$V_1(\bar{y}_{Rs}) = \frac{1}{N}\sum_{t=1}^{k} p_t \frac{N_t - n_t}{n_t}\left[S_{ty}^2 + R_{Nt}^2 S_{tx}^2 - 2R_{Nt}\rho_t S_{tx} S_{ty}\right] \quad ...(8)$$

The equation (8) can be rewritten as

$$V_1(\bar{y}_{Rs}) = \frac{1}{N}\sum_{t=1}^{k} \frac{p_t(N_t - n_t)}{n_t(N_t - 1)} \sum_{i=1}^{N_t}\{y_{ti} - R_{Nt}\, x_{ti}\}^2 \quad ...(9)$$

the above results have been proved under the assumption that the sample size n_t is large. Of course, this cannot be always true. Taking into consideration this difficulty, Hansen, Hurwitz and Gurney (1946) suggested another estimate known as combined ratio estimate, given by the following formula :

$$\bar{y}_{Rc} = \frac{\sum_{t=1}^{k} p_t \bar{y}_{nt}}{\sum_{t=1}^{k} p_t \bar{x}_{nt}} \bar{x}_N \quad ...(10)$$

The above estimate (10) is consistent estimate of the population mean. To obtain its expected value and variance we proceed as follows :

Let,
$$\sum_{t=1}^{k} p_t \bar{y}_{nt} = \bar{y}_N + \varepsilon_1 \quad ...(11)$$

$$\sum_{t=1}^{k} p_t \bar{x}_{nt} = \bar{x}_N + \varepsilon_2$$

where,

$$E(\varepsilon_1) = E(\varepsilon_2) = 0$$

$$E(\varepsilon_1^2) = \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} p_t^2 S_{ty}^2 \text{ and } E(\varepsilon_2^3) = \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} p_t^2 S_{tx}^2 \quad ...(12)$$

Proceeding as done in the case of ratio estimate, the first approximation gives the following result :

$$E_1(\bar{y}_{Rc}) = \bar{y}_N\left\{1 + \frac{E(\varepsilon_2^2)}{\bar{x}_N^2} - \frac{E(\varepsilon_1\varepsilon_2)}{\bar{x}_N\,\bar{y}_N}\right\} \quad ...(13)$$

NOTES

and the relative bias in $\bar{y}_{Rc}$ is given by

$$\text{Bias} = \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} p_t^2 \left\{ \frac{S_{tx}^2}{\bar{x}_N^2} - \frac{\rho_t S_{tx} S_{ty}}{\bar{x}_N \bar{y}_N} \right\} \quad ...(14)$$

Next, to show that this bias diminishes with the increase in sample size n_t, we shall assume that $n_t \propto N_t$ and also that $\frac{S_{tx}}{\bar{x}_N}$, $\frac{S_{ty}}{\bar{y}_N}$ and ρ_t are same over all strata, say C_x, C_y and ρ respectively.

Thus the relative bias in the combined ratio estimate is given by

$$\text{Bias } (\bar{y}_{Rc}) = \frac{N-n}{N} \frac{1}{n} \left(C_x^2 - \rho C_x C_y \right) \quad ...(15)$$

Thus it follows that even if the sample size in each stratum is small, a combined ratio estimate gives a satisfactory estimate of the population mean provided that the total sample size is sufficiently large.

Similarly, to the first approximation, we have

$$V_1 (\bar{y}_{Rc}) = \bar{y}_N^2 \left\{ \frac{E(a_1^2)}{\bar{y}_N^2} + \frac{E(a_2^2)}{\bar{x}_N^2} - \frac{2E(a_1 a_2)}{\bar{y}_N \bar{x}_N} \right\}$$

$$= \frac{1}{N} \sum_{t=1}^{k} p_t \frac{N_t - n_t}{n_t} \left\{ S_{ty}^2 + R_N^2 S_{tx}^2 - 2R_N \rho_t S_{ty} S_{tx} \right\} \quad ...(16)$$

Comparing (8) and (16), we see that the variance of the combined ratio estimate has the same form as that ratio estimate based on separate strata.

The difference between two sampling variance is given by

$$V_1 (\bar{y}_{Rc}) - V_1 (\bar{y}_{Rs}) = \frac{1}{N} \sum_{t=1}^{k} \frac{p_t (N_t - n_t)}{n_t} \left[S_{tx}^2 \{ R_N - R_{Nt} \}^2 \right.$$

$$\left. + 2(R_N - R_{Nt}) \{ R_{Nt} S_{tx}^2 - \rho_t S_{ty} S_{tx} \} \right] \quad ...(17)$$

From equation (17), it is obvious that the difference in the two variances depend upon two factors :

(*i*) the magnitude of variation between the strata ratios and

(*ii*) the value of $\{R_{Nt} S_{tx}^2 - \rho_t S_{ty} S_{tx}\}$

The II[nd] function is usually small and even it vanishes when the regression of y on x is a straight line passing through the origin within each stratum. Thus it follows that the combined estimate will have a lower precision than that based upon separate strata. On the other hand side, the bias in the former case will be smaller than the later case. Thus, it is concluded that unless the population ratios in the different strata vary considerably, the combined estimate will provide an estimate with negligible bias and will give a precision as high as that of the estimate based upon separate ratios.

STUDENT ACTIVITY 1

1. Show that R_n is not an unbiased estimate of R_N.

STUDENT ACTIVITY 2

1. What is optimum property of ratio estimate ? Discuss the efficiency of the ratio estimate.

8.10 ILLUSTRATIVE EXAMPLES

Example 1. *For studying milk yield, feeding and management practices of milch animals in the year 1977-78, the whole of Panjab State was divided into 4 zones according to agro-climatic conditions. The total number of milch animals in 17 randomly selected villages (in 1977-78) of zone A, along with their livestock census data in 1976, are shown below :*

S. No. of village	1	2	3	4	5	6	7
Number of milch animals in survey (Y)	1129	1144	1125	1138	1137	1127	1163
Number of milch animals in census (X)	1141	1144	1127	1153	1117	1140	1153
S. No. of village	8	9	10	11	12	13	14
Number of milch animals in survey (Y)	1153	1164	1130	1153	1125	1116	1115
Number of milch animals in census (X)	1146	1189	1137	1170	1115	1130	1118
S. No. of village	15	16	17				
Number of milch animals in survey (Y)	1112	1112	1123				
Number of milch animals in ceusus (X)	1122	1113	1166				

Estimate the total number of milch animals in 117 villages of zone A

(i) by Ratio Method and

(ii) by simple mean per unit method

Also compare its precision, given the total number of milch animals in the census = 143968

Solution.

Given $N = 117, \quad n = 17, \quad X = 143968,$

$$\sum x_i = 19381, \quad \sum y_i = 19266$$

$$\bar{x} = 1140.06, \quad \bar{y} = 1133.29$$

$$\hat{R} = 0.994$$

$$s_y^2 = 287.85, \quad s_x^2 = 458.56, \quad s_{xy} = 262.86$$

(I) **Ratio Estimate.** The total number of milch animals by the ratio method of estimation is given by

$$\hat{Y}_R = \hat{R}\,X = 0.994 \times 143968 = 143120$$

NOTES

and an estimate of the variance of $\hat{Y}_R$ is given as

$$v\left(\hat{Y}_R\right) = \frac{(1-f)N^2}{n}\left[s_y^2 + R^2 s_x^2 - 2R\, s_{xy}\right]$$

$$= \frac{100 \times 117}{17}\,[287.85 + 453.61 - 522.62]$$

$$= 150{,}304$$

(II) **Mean Per Unit Estimate.** The total number of milch animals by mean per unit estimate is given by

$$\hat{Y} = N\bar{y} = 117 \times 1133.29 = 132595$$

and an estimate of variance of $\hat{Y}$ is given by

$$v\left(\hat{Y}\right) = \frac{(1-f)N^2}{n} s_y^2 = \frac{100 \times 117}{17} \times 287.85$$

$$= 198{,}108$$

Hence, the relative precision of ratio estimate is given as

$$R.P. = \frac{v\left(\hat{Y}\right) - v\left(\hat{Y}_R\right)}{v\left(\hat{Y}_R\right)} \times 100 = 32$$

Example 2. *The following data were collected in a pilot survey for estimating the extent of cultivation and production of fresh fruits in three districts of Uttar Pradesh in the year 1976-77.*

Stratum number	*Total no. of villages* (N_m)	*Total area (in hect.) under orchard* (X_m)	*No. of villages in sample* (n_m)	*Area under orchards in ha.* (x_m)	*Total no. of trees* (y_m)
1	*985*	*11253*	*6*	*10.63, 9.90, 1.45, 3.38, 5.17, 10.35*	*747, 719, 78, 201, 311, 448*
2	*2196*	*25115*	*8*	*14.66, 2.61, 4.35, 9.87 2.42, 5.60, 4.70, 36.75*	*580, 103, 316, 739 196, 235, 212, 1646*
3	*1020*	*18870*	*11*	*11.60, 5.29, 7.94, 7.29, 8.00, 1.20, 11.50, 7.96, 23.15, 1.70, 2.01*	*488, 227, 374, 491, 499, 50, 455, 47, 879, 115, 115*

Estimate the total number of trees in the three districts by various methods and compare their precision.

NOTES

Solution. The calculations are shown in the table given below :

Stratum	W_m	$\left(\frac{1}{n_m}-\frac{1}{N_m}\right)$	$\bar{x}_m$	$\bar{y}_m$	$\hat{R}_m$	$W_m \cdot \bar{x}_m$	$W_m \cdot \bar{y}_m$	$s^2_{x_m}$	$s^2_{y_m}$	$s_{x_m y_m}$
1	0.2345	0.16565	6.81	417.33	61.28	1.60	97.86	15.97	74775.47	1007.05
2	0.5227	0.12454	10.12	503.38	49.74	5.29	263.12	132.66	259113.40	5709.16
3	0.2428	0.08992	7.97	340.00	42.66	1.94	82.55	38.44	65885.60	1404.71

Here $\hat{R} = \sum W_m \bar{y}_m / \sum W_m \bar{x}_m = 443.53/8.80 = 50.40$

(I) **Combined Ratio Estimate.** The estimate of the total number of trees is given by

$$\hat{Y}_{Rc} = \frac{\sum W_m \bar{y}_m}{\sum W_m \bar{x}_m} X = \frac{443.53}{8.80} \times 55238 = 2783995$$

$$v\left(\hat{Y}_{Rc}\right) = \sum N_m^2 \left(\frac{1}{n_m}-\frac{1}{N_m}\right)\left(s^2_{y_m} + \hat{R}s^2_{x_m} - 2\hat{R}s_{xy_m}\right)$$

$$= (985)^2\,(0.16598)\,[74778.80 + (50.40)^2 \times 16.03 - 2 \times 50.40 \times 1008.75] + (2196)^2 \times (0.12454)\,[259107.90 + (50.40)^2 \times 129.64 - 2 \times 50.40 \times 5643.81] + (1020)^2 \times 0.08902\,[65885.60 + (50.40)^2 \times 38.39 - 2 \times 50.40 \times 1403.81]$$

$$= 161057.35 \times 13815.57 + 602802.00 \times 19518.23 + 92595.60 \times 21910.39 = 6019519627.34$$

(II) **Separate Ratio Estimate.** Another estimate of the total number of trees is given by

$$\hat{Y}_{Rs} = \sum \hat{R}_m X_m$$

$$= 61.28 \times 11253 + 49.99 \times 25115 + 42.66 \times 18870$$

$$= 2750076.89 = 2750077$$

Estimated variance of $\hat{Y}_{Rs}$ is given by

$$v\left(\hat{Y}_{Rs}\right) = \sum N_m^2 \left(\frac{1}{n_m}-\frac{1}{N_m}\right)\left(s^2_{y_m} + \hat{R}_m^2 s^2_{x_m} - 2\hat{R}_m s_{xy_m}\right)$$

$$= (985)^2 \times (0.16598)\,[74778.80 + (61.28)^2 \times 16.03 - 2 \times (61.28) \times 1008.75]$$

NOTES

$$+ (2196)^2 \times (0.12454)\ [259107.90 + (49.99)^2 \times 16.03$$
$$- 2 \times 49.99 \times 5643.81] + (1020)^2 \times (0.08902 \times [65885.60$$
$$+ (42.66)^2 \times 38.39 - 2 \times 42.66 \times 1403.69]$$
$$= 161057.35 \times 11342.88 + 602802.00 \times 18810.18$$
$$+ 92595.60 \times 15937.79 = 2441137855.48$$

The efficiency of separate ratio estimate $\left(\hat{Y}_{Rs}\right)$ over the combined ratio estimate $\left(\hat{Y}_{Rc}\right)$ is given by

$$R.P. = \frac{601951962 7.34}{2441137855.48} \times 100$$
$$= 246.58\%.$$

8.11 SUMMARY

- The ratioi estimate of the population mean is given by $\bar{y}_R = R_N \bar{x}_N$ where R_N is the population ratio.
- The relative efficiency of an estimate B compared to that of another estimate A based on a sample of equal size is defined as ratio of the inverse of their variances.
- In large sampling, the ratio estimate is moire efficient than the corresponding sample estimate based on A.M. If the denominator is less than one.

8.12 GLOSSARY

- **Expected value.** A predicted value of a variable, calculated as the sum of all possible values each multiplied by the probability of its occurrence.
- **Variance.** A quantity equal to the square of the standard deviation.
- **Ratio.** The quantitative relation between two amounts showing the number of times one value contains or is contained within the other.

8.13 REVIEW QUESTIONS

1. If y and x are unbiased estimators of the population totals Y and X of the main variate and the auxiliary variate, respectively, based on any sample design, show that the ratio of the exact bias of the ratio estimator,

NOTES

$\frac{y}{x} X$ to its standard error is not greater than the relative standard error of the estimator x. Derive an approximate expression for the bias and the mean square error stating clearly the assumptions involved.

2. If y_i and x_i are unbiased estimators of the population totals Y and X of the main variate and the auxiliary variate, respectively, based on the i^{th} sample selected with probability p_i, show that the ratio estimator $\frac{y_i}{x_i} X$ will be unbiased for the population total Y if the sample i is selected with probability proportional to $x_i p_i$. Derive an expression for its sampling variance.

3. If y and x are unbiased estimators of the population totals Y and X, respectively, based on the same set of sample units drawn according to any sample design, derive the expressions correct to the second degree of approximation for bias and mean square error of the ratio estimator $\frac{y}{x} X$ for estimating the population total Y. State clearly the underlying assumptions and the degree of the approximations taken. Also state under what conditions the above ratio estimator is more efficient than the estimator $\hat{Y}$ for large samples ?

4. (*a*) Define ratio estimator for estimating the population total of a character y and derive an expression for the standard error of the estimator. State the conditions under which the ratio estimator is 'blue' (best linear unbiased estimator).

   (*b*) If the coefficient of variation of the auxiliary variate x is more than twice the coefficient of variation of the character y, show that, in large samples with simple random sampling, the ratio estimator is less precise than the mean per unit estimator. Is the converse true ?

5. In simple random sampling, wor, the ratio estimator of $R = Y/X$ is given by $\bar{y}/\bar{x}$. Obtain an exact expression for the variance of $\bar{y}/\bar{x}$.

6. If y_i, x_i $(i = 1, ..., m)$ are unbiased estimators of Y and X respectively, based on m interpenetrating sub-samples of the same size, prove that $\bar{r}X + m(\bar{y} - \bar{r}\bar{x})/(m-1)$ is an unbiased estimator of Y,

   where $\bar{r} = \frac{1}{m} \sum_{i}^{m} \left(\frac{y_i}{x_i} \right)$

NOTES

7. Values of y and x are measured for each unit in a simple random sample to estimate the population ratio, $R = \bar{Y}/\bar{X}$. Which of the following estimators would you recommend to estimate R ?

(*i*) Always use $\bar{y}/\bar{X}$.

(*ii*) Always use $\bar{y}/\bar{x}$.

(*iii*) Either use $\bar{y}/\bar{X}$ or $\bar{y}/\bar{x}$, depending upon the conditions (Given that x is known).

Give reasons for your choice.

8. Suppose a finite population of N individuals has NP_1 individuals sas agriculturists, of which NP_2 individuals are literates. In order to estimate the population ratio P_2/P_1, a random sample, wor, is drawn where p_1 and p_2 are the sample proportions corresponding to P_1 and P_2, respectively.

(*i*) Show that the estimator p_2/p_1 is more efficient than p_2/P_1, when P_1 is known.

(*ii*) Derive the condition for p_2/p_1 to be more efficient than P_2/p_1 when P_2 is known.

(Elkin, 1953)

9. A sample survey for the study of yield and cultivation practices of guava was conducted in District Allahabad during 1971-72. Out of a total of 146 guava-growing villages in Phulpur-Saran Tehsil, 13 villages were selected by the method of simple random sampling. The data for the total number of guava trees and area under guava orchards for the 13 selected villages are given below :

S. N. of villages	1	2	3	4	5	6	7
Total N. of guava trees	492	1008	714	1265	1889	784	294
Area under guava orchards (in acres)	4.80	5.99	4.27	8.43	14.39	6.53	1.88

S. N. of villages	8	9	10	11	12	13
Total N. of guava trees	798	780	619	403	467	197
Area under guava orchards (in acres)	6.35	6.58	9.18	2.00	2.20	1.00

Given that the total area under guava orchards of 146 villages is 354.78 acres, estimate the total number of guava trees in the tehsil along with its standard error, using the area under guava orchards as the auxiliary variate. Discuss the efficiency of your estimate with the one which does not make any use of the information on the auxiliary variate.

NOTES

10. The number of cows in milk enumerated (y) from a random sample of 20 villages from a tehsil having 84 villages, as also the corresponding census figures (x) in the previous year, are given below :

Villages	y	x
1	237	155
2	1060	583
3	405	205
4	1085	738
5	666	526
6	542	284
7	1337	758
8	1166	681
9	399	143
10	228	111
11	813	616
12	666	576
13	681	540
14	2743	2242
15	1228	940
16	472	387
17	643	675
18	180	220
19	583	654
20	1195	1787

Given that the census estimate of the number of cows in milk in the tehsil was 74488, estimate the number of cows in milk in the current year with and without using the census information and compare the efficiencies of the estimates.

11. A sample of 34 villages was selected from a population of 170 villages, with pps wr for estimating the area under wheat in the region during 1974. The cultivated area was 78000 acres in 1971. Later it was found that the figures for area under wheat (21288 acres) in 1973 were also available for all villages in the region. The relevant data are given below :

NOTES

S.N.	Area under wheat (in acres)		
	(1971) (x_1)	(1973) (x_2)	(1974) (y)
1	401	70	50
2	630	163	149
3	1194	320	284
4	1170	440	381
5	1065	250	278
6	827	125	111
7	1737	558	634
8	1060	254	278
9	360	101	112
10	946	359	355
11	4170	109	99
12	1625	481	498
13	827	125	111
14	96	5	6
15	1304	427	339
16	377	78	80
17	259	75	105
18	186	45	27
19	1767	564	515
20	604	238	249
21	700	92	85
22	524	247	221
23	571	134	133
24	962	131	144
25	407	129	103
26	715	190	175
27	845	363	335
28	1016	235	219
29	184	73	62
30	282	62	79
31	194	71	60
32	439	137	100
33	854	196	141
34	820	255	263

(*i*) Estimate the area under wheat in 1974, by the method of ratio estimation, using the information on wheat area for 1973 and estimate its standard error.

(*ii*) Determine the efficiency of the ratio estimate as compared to that of the usual unbiased estimate.

12. What do you mean by unbiased ratio type estimate.

8.14 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

NOTES

CHAPTER 9 REGRESSION METHOD OF ESTIMATION

OBJECTIVES

After going through this chapter, we should be able to :

- know simple regression estimate.
- compute expected value of simple regression estimate.
- know variance of simple regression estimate.
- know optimum conditions.
- compute regression estimate in stratified sampling.

STRUCTURE

NOTES

9.1 SIMPLE REGRESSION ESTIMATE

The simple regression estimate can be obtained by considering difference estimator defined by

$$\bar{y}_d = \bar{y}_n + \beta(\bar{x}_N - \bar{x}_n) \quad ...(1)$$

where β is fixed constant

Now, $E(\bar{y}_d) = [E(\bar{y}_n) + \beta\bar{x}_N - E(\bar{x}_n)] = \bar{y}_N + \beta(\bar{x}_N - \bar{x}_N) = \bar{y}_N$

$\Rightarrow$ The difference estimator $\bar{y}_d$ is an unbiased estimate of population mean $\bar{y}_N$.

To obtain the optimum value for the constant β, we minimize the variance of the estimate *w.r.t.* β.

Now, $V(\bar{y}_d) = V(\bar{y}_n) + \beta^2 V(\bar{x}_n) - 2\beta\, cov.(\bar{y}_n, \bar{x}_n)$ ...(2)

$$\therefore \quad \frac{d}{d\beta}V(\bar{y}_d) = 0 \quad \Rightarrow \quad \beta = \frac{cov.(\bar{y}_n, \bar{x}_n)}{V(\bar{x}_n)} = \frac{S_{yx}}{S_x^2} \quad ...(3)$$

$\Rightarrow$ the optimum value of β is the regression coefficient of y on x.

Since the regression coefficient β is generally not known, we take consistent estimate of β given by

$$\hat{\beta} = \frac{s_{yx}}{s_x^2}$$

where, $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)$ and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$

Thus, simple regression estimate is given by

$$\boxed{\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)}$$

9.2 EXPECTED VALUE OF SIMPLE REGRESSION ESTIMATE

To find the expected value, we write

$$\left.\begin{aligned} \bar{x}_n &= \bar{x}_N + \varepsilon_1 \\ s_{xy} &= S_{xy} + \varepsilon_2 \\ s_x^2 &= S_x^{\,2} + \varepsilon_3 \end{aligned}\right] \quad ...(1)$$

NOTES

where $E(\varepsilon_1) = E(\varepsilon_2) = E(\varepsilon_3) = 0$ ...(2)

Then,

$$\bar{y}_l = \bar{y}_n + \frac{S_{xy} + \varepsilon_2}{S_x^2 + \varepsilon_3}(-\varepsilon_1)$$

$$= \bar{y}_n - \beta\varepsilon_1\left[1 + \frac{\varepsilon_2}{S_{xy}}\right]\left[1 + \frac{\varepsilon_3}{S_x^{\,2}}\right]^{-1}$$

where β is regression coefficient of y on x. Assuming that $\left|\frac{\varepsilon_3}{S_x^{\,2}}\right| < 1$, expanding and ignoring terms in ε of the order higher than two and taking expectation, we obtain

$$E[\bar{y}_l] \cong \bar{y}_N - \beta\left[\frac{E(\varepsilon_1\varepsilon_2)}{S_{xy}} - \frac{E(\varepsilon_1\varepsilon_3)}{S_x^{\,2}}\right]$$

$$= \bar{y}_N - \beta\left[\frac{cov.(\bar{x}_n, s_{xy})}{S_{xy}} - \frac{cov.(\bar{x}_n, s_x^{\,2})}{S_x^{\,2}}\right] \quad ...(3)$$

Using the method of symmetric functions and neglecting terms of the order $\frac{1}{n^v}$ where $v > 1$, it can be proved that

$$\left.\begin{aligned} cov.(\bar{x}_n, s_{xy}) &\cong \frac{N-n}{Nn}\mu_{21} \\ cov.(\bar{x}_n, s_x^{\,2}) &\cong \frac{N-n}{Nn}\mu_{30} \end{aligned}\right] \quad ...(4)$$

where

$$\mu_{21} = E\left[(x - \bar{x}_N)^2(y - \bar{y}_N)\right] \text{ and } \mu_{30} = E\left[x - \bar{x}_n\right]^3 \quad ...(5)$$

Thus, to the first degree of approximation

$$E[\bar{y}_l] \cong \bar{y}_N - \frac{N-n}{Nn}\beta\left[\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^{\,2}}\right] \quad ...(6)$$

$\Rightarrow$ simple regression estimate $\bar{y}_l$ is a biased estimate. The bias will be negligible if the sample size n is sufficiently large.

Also $\because \bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$

$$\therefore E(\bar{y}_l) = \bar{y}_N + \bar{x}_N E(\hat{\beta}) - E(\hat{\beta}\bar{x}_n)$$

$$= \bar{y}_N - \left[E(\hat{\beta}\bar{x}_n) - E(\hat{\beta})E(\bar{x}_n)\right]$$

$$= \bar{y}_N - cov.\left(\hat{\beta}, \bar{x}_n\right)$$

which again shows that $\bar{y}_l$ is biased by an amount $-cov.\left(\hat{\beta}, \bar{x}_n\right)$.

NOTES

9.3 VARIANCE OF THE SIMPLE REGRESSION ESTIMATE

By definition

$$V(\bar{y}_l) = E\left[\bar{y}_l - E(\bar{y}_l)\right]^2 \qquad ...(1)$$

Now,
$$\bar{y}_l = \bar{y}_n - \beta\varepsilon_1 - \beta\left[\frac{\varepsilon_1\varepsilon_2}{S_{xy}} - \frac{\varepsilon_1\varepsilon_2}{S_x^{\,2}}\right] + \ldots \qquad ...(2)$$

Also
$$E(\bar{y}_l) \cong \bar{y}_N - \frac{N-n}{Nn}\beta\left[\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^{\,2}}\right] \qquad ...(3)$$

Using (3) and (2) in (1), we get

$$V(\bar{y}_l) \cong E\left[(\bar{y}_n - \bar{y}_N) - \beta\varepsilon_1\right]^2$$

$$= E\left(\bar{y}_n - \bar{y}_N\right)^2 + \beta^2 E\left(\varepsilon_1^{\,2}\right) - 2\beta\, E\left[\varepsilon_1(\bar{y}_n - \bar{y}_N)\right]$$

$$= V(\bar{y}_n) + \beta^2\, V(\bar{x}_n) - 2\beta\, COV.(\bar{y}_n, \bar{x}_n) \quad \Big| \quad \bar{x}_n = \bar{x}_N + \varepsilon_1$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)\left[S_y^{\,2} + \beta^2 S_x^{\,2} - 2\beta\, S_{xy}\right]$$

$$\Big| \because V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2\,(1-\rho^2) \qquad ...(4)$$

$$S_y^2 + \beta^2 S_x^2 - 2\beta\, S_{xy} = S_y^2 + \beta^2 S_x^2 - 2\beta^2 S_x^2 \qquad \Big| \because \beta = \frac{S_{yx}}{S_x^{\,2}}$$

$$= S_y^{\,2} - \beta^2 S_x^{\,2}$$

$$= S_y^{\,2} - \rho^2 \frac{S_y^{\,2}}{S_x^{\,2}} S_x^{\,2} \qquad \Big| \because \beta = \frac{\rho S_y S_x}{S_x^{\,2}} = \rho\frac{S_y}{S_x}$$

$$= S_y^{\,2}\left(1-\rho^2\right)$$

where ρ is the coefficient of correlation between y and x in the population.

For large N,

$$V(\bar{y}_l) \cong \sigma_y^2 \frac{(1-\rho^2)}{n} \quad ...(5)$$

NOTES

To have the idea about the accuracy of the large sample formula for the variance of the simple regression estimate, we will now assume that the population is large and the joint distinction of x and y is bivariate normal. Now using the result that in samples from a bivariate normal population, the sample means are distributed independently of the regression coefficients, we have

$$V(\bar{y}_l) = E\left[V\left(\bar{y}_l|\hat{\beta}\right)\right] + V\left[E\left(\bar{y}_l|\hat{\beta}\right)\right]$$

$$= E\left[V\left(\bar{y}_l|\hat{\beta}\right)\right] \quad \left(\because E\left(\bar{y}_l|\hat{\beta}\right) = E(y_l) = \bar{y}_N \ \therefore V\left(\bar{y}_N\right) = 0\right)$$

$$= E\left[\frac{\sigma_y^2}{n} + \hat{\beta}^2 \frac{\sigma_x^2}{n} - 2\hat{\beta}\frac{\sigma_{xy}}{n}\right]$$

$$\left| V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}\right)\right.$$

$$= \frac{\sigma_y^2}{n} + \frac{\sigma_x^2}{n} E\left(\hat{\beta}^2\right) - 2\frac{\sigma_{xy}}{n} E\left(\hat{\beta}\right) \quad ...(6)$$

Now, $E\left(\hat{\beta}\right) = \beta \quad ...(7)$

$\hat{\beta}$ is consistent $\therefore$ it is unbiased. Even if $\hat{\beta}$ is biased, the bias will be zero for large population.

Also variance of consistent estimate $\hat{\beta}$ is given by

$$V\left(\hat{\beta}\right) = \frac{\sigma_y^2 (1-\rho^2)}{\sigma_x^2 (n-3)} \quad ...(8)$$

Using (7) and (8) in (6), we get

$$V(\hat{y}_l) = \frac{\sigma_y^2 (1-\rho^2)}{n}\left[1 + \frac{1}{n-3}\right] \quad ...(9)$$

$$((\because \ V(\hat{y}_l) = \frac{\sigma_y^2}{n} + \frac{\sigma_x^2}{n}\left[V\left(\hat{\beta}\right) + \beta^2\right] - 2\frac{\sigma_{xy}}{n}\beta$$

$$V\left(\hat{\beta}\right) = E\left(\hat{\beta} - \beta\right)^2 = E\left(\hat{\beta}^2\right) + \beta^2 - 2\beta E\left(\hat{\beta}\right) = E\left(\hat{\beta}^2\right) + \beta^2 - 2\beta^2 = E\left(\hat{\beta}^2\right) - \beta^2$$

NOTES

For large population $\beta = \frac{S_{yx}}{S_x^2} = \frac{\sigma_{yx}}{\sigma_x^2}$

$$\therefore \quad V(\bar{y}_l) = \frac{\sigma_y^2}{n} + \frac{\sigma_x^2}{n}\left[\frac{\sigma_y^2(1-\rho^2)}{\sigma_x^2(n-3)} + \frac{\sigma_{yx}^2}{\sigma_x^2}\right] - 2\frac{\sigma_{xy}}{n}\cdot\frac{\sigma_{xy}}{\sigma_x^2}$$

$$= \frac{\sigma_y^2}{n} + \frac{\sigma_y^2}{n}(1-\rho^2)\frac{1}{n-3} + \frac{\sigma_{yx}^2}{n\sigma_x^2} - 2\frac{\rho^2\sigma_y^2\sigma_x^2}{n\sigma_x^2}$$

$$((\because \sigma_{yx} = \rho\sigma_y\sigma_x))$$

$$= \frac{\sigma_y^2}{n} + \frac{\sigma_y^2}{n}(1-\rho^2)\frac{1}{n-3} + \frac{\rho^2\sigma_y^2}{n} - 2\frac{\rho^2\sigma_y^2}{n} = \frac{\sigma_y^2}{n}\left[1 + \frac{1-\rho^2}{n-3} - \rho^2\right]))$$

Thus, we see that unless n is very small, the large sample formula for the variance of the simple regression estimate may be considered adequate.

Again, we have

$$S_y^2 + \beta^2 S_x^2 - 2\beta S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2 + \beta^2\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x}_N)^2$$

$$-2\beta\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)(x_i - \bar{x}_N)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i - \bar{y}_N)^2\right] + \beta^2(x_i - \bar{x}_N)^2$$

$$-2\beta(y_i - \bar{y}_N)(x_i - \bar{x}_N)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i - \bar{y}_N) - \beta(x_i - \bar{x}_N)\right]^2 \quad ...(10)$$

Thus from equation (4), we have

$$V(\bar{y}_l) \cong \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i - \bar{y}_N) - \beta(x_i - \bar{x}_N)\right]^2 \quad ...(11)$$

If the population is regarded as divided into k classes with the N_i units in the ith class having the value x_i each, the variable can be written as

$$V(\bar{y}_l) \cong \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[(y_{ij} - \bar{y}_N) - \beta(x_i - \bar{x}_N)\right]^2 \quad ...(12)$$

Now, when the regression of y on x is linear, we have

$$E(y_{ij}|x_i) = \alpha + \beta x_i \quad ...(13)$$

NOTES

where $\alpha = \bar{y}_N - \beta\bar{x}_N$ ...(14)

$$E(y_{ij}/x_i) = \alpha + \beta x_i$$

$$\therefore E\left[\sum_{i=1}^{N} y_{ij}/x_i\right] = N\alpha + \beta\sum_{i=1}^{N} x_i$$

$$\Rightarrow E\left[\sum \frac{y_{ij}}{N}\right] = \alpha + \beta\bar{x}_N$$

$$\Rightarrow E(\bar{y}_j) = \alpha + \beta\bar{x}_N$$

$$i.e., \bar{y}_N = \alpha + \beta\bar{x}_N$$

Thus, (12) can be written as

$$V(\bar{y}_l) \cong \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{N_i}\left[y_{ij} - E(y_{ij}/x_i)\right]^2$$

$$((\because (y_{ij} - \bar{y}_N) - \beta(x_i - \bar{x}_N) = (y_{ij} - \bar{y}_N) - \left[E(y_{ij}/x_i) - \alpha\right] + (\bar{y}_N - \alpha)$$

$$= y_{ij} - \bar{y}_N - E(y_{ij}/x_i) + \alpha + \bar{y}_N - \alpha))$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{K} N_i V(y_{ij}/x_i) \quad ...(15)$$

In particular if variance of y for fixed x is constant *i.e.*, if $V(y_{ij}/x_i) = \gamma$, then

$$V(\bar{y}_l) \cong \frac{N-n}{N-1}\frac{\gamma}{n}.$$

9.4 ESTIMATE OF THE VARIANCE OF THE SIMPLE REGRESSION ESTIMATE

Since s_y^2, s_x^2 and s_{yx} are unbiased estimates of S_y^2, S_x^2 and S_{yx} respectively, a consistent estimate of the variance of the regression estimate is given by

$$\text{Est. } (\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(s_y^2 + \hat{\beta}^2 s_x^2 - 2\beta s_{xy}\right)$$

or $$\text{Est. } (\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right)s_y^2\left(1 - r^2\right)$$

where $r = \dfrac{s_{xy}}{s_x s_y}$ is the sample correlation coefficient

9.5 CONDITIONS UNDER WHICH THE SIMPLE REGRESSION ESTIMATE IS OPTIMUM

NOTES

We shall assume that

(*i*) The population of size N can be regarded as divided into k-classes with the N_i units in the ith class having the value x_i each with $\sum_{i=1}^{K} N_i = N$

and that in repeated samples of size n, drawn from the population with equal probability and without replacement, the number of units having the value x_i is fixed, say n_i; $i = 1, 2, \ldots k$.

(*ii*) The N_i are sufficiently large so that $(N_i - 1) / N_i \cong 1$ and $\frac{N_i - n_i}{N_i - 1} \cong 1$

(*iii*) $E(y_{ij}/x_i) = \alpha + \beta x_i$ and (*iv*) $V(y_{ij}/x_i) = \gamma_i$

Consider the subclass of linear estimates of the type

$$T = \sum_{i=1}^{k} \lambda_i n_i \bar{y}_{n_i} \qquad \ldots(1)$$

where for all the y observations in the ith class are assumed to have the same weight λ_i which may depend on the x_i but not on the y observations. We shall choose λ_i so that the estimate is optimum in the sense that it is unbiased and has minimum variance.

Since the estimate is to be unbiased, we have

$$E(T/n_1, n_2, \ldots, n_k) = E\left[\sum_{i=1}^{k} \lambda_i n_i \bar{y}_{n_i} | n_1, n_2, \ldots, n_k\right]$$

$$= \sum_{i=1}^{k} \lambda_i n_i (\alpha + \beta x_i) \qquad \left(\left(\because E(\bar{y}_{n_i}) = \bar{y}_{N_i} = E(y_{ij}|x_i)\right)\right)$$

$$= \bar{y}_N = \sum_{i=1}^{k} p_i \bar{y}_{N_i} = \sum_{i=1}^{k} p_i (\alpha + \beta x_i) \quad \left(\left(\because \bar{y}_N = E(\bar{y}_{N_i})\right)\right)$$

where $p_i = \frac{N_i}{N}$.

Thus, $$\sum_{i=1}^{k} (\lambda_i n_i - p_i)(\alpha + \beta x_i) = 0 \qquad \ldots(2)$$

Now, $V(T/n_1, n_2, \ldots, n_k) = \sum_{i=1}^{k} \lambda_i^2 n_i^2 V(\bar{y}_{n_i}|x_i)$

NOTES

$$\cong \sum_{i=1}^{k} \lambda_i n_i^2 \cdot \frac{\gamma_i}{n_i}$$

$$V(\bar{y}_n) = \frac{N-n}{N}\frac{S^2}{n} \quad \therefore \quad V(\bar{y}_{n_i}) = \frac{N_i - n_i}{N_i}\frac{V(y_{ij}|x_i)}{n_i} \cong 1.\frac{\gamma_i}{n_i}$$

$$\therefore \quad V(T/n_1, n_2, \ldots, n_k) \cong \sum_{i=1}^{k} \lambda_i^2 \frac{n_i^2}{w_i} \qquad \ldots(3)$$

where $$w_i = \frac{n_i}{\gamma_i} \qquad \ldots(4)$$

We shall now determine the constants λ_i, α and β so that the variance given by (3) is minimum subject to the condition (2).

Consider the function

$$\phi = \sum_{i=1}^{k} \frac{\lambda_i^2 n_i^2}{w_i} - \mu \sum_{i=1}^{k} (\lambda_i n_i - p_i)(\alpha + \beta x_i) \qquad \ldots(5)$$

where μ is fixed constant.

Differentiate ϕ *w.r.t.* λ_i, α and β and equating to zero, we get

$$\frac{\partial \phi}{\partial \lambda_i} = \frac{2\lambda_i n_i^2}{w_i} - \mu n_i (\alpha + \beta x_i) = 0 \quad ; i = 1, 2, \ldots k \qquad \ldots(6)$$

$$\frac{\partial \phi}{\partial \alpha} = -\mu \sum_{i=1}^{k} (\lambda_i n_i - p_i) = 0 \qquad \ldots(7)$$

$$\frac{\partial \phi}{\partial \beta} = -\mu \sum_{i=1}^{k} x_i(\lambda_i n_i - p_i) = 0 \qquad \ldots(8)$$

From (6), $$\lambda_i = \frac{\mu w_i}{2n_i}(\alpha + \beta x_i) \qquad \ldots(9)$$

Substituting λ_i from (9) in (7) and (8) and putting $\alpha' = \frac{\mu\alpha}{2}$, $\beta' = \frac{\mu\beta}{2}$, we get

$$w\alpha' + w\bar{x}_w \beta' = 1 \qquad \ldots(10)$$

$$w\bar{x}_w \alpha' + \left(\sum_{i=1}^{k} w_i x_i^2\right)\beta' = \bar{x}_N \qquad \ldots(11)$$

where $$w = \sum_{i=1}^{k} w_i \text{ and } \sum_{i=1}^{k} w_i x_i = w\bar{x}_w \qquad \ldots(12)$$

NOTES

Solving (10) and (11) for α' and β', we get

$$\beta' = \frac{\bar{x}_N - \bar{x}_w}{\sum_{i=1}^{k} w_i (x_i - \bar{x}_w)^2} \quad ...(13)$$

and

$$\alpha' = \frac{1}{w} - \frac{\bar{x}_w(\bar{x}_N - \bar{x}_w)}{\sum_{i=1}^{k} w_i (x_i - \bar{x}_w)^2} \quad ...(14)$$

Substituting the optimum values of α' and β' in (9), we get

$$\lambda_i = \frac{w_i}{n_i}\left[\frac{1}{w} + \frac{(\bar{x}_N - \bar{x}_w)(\bar{x}_i - \bar{x}_w)}{\sum_{i=1}^{k} w_i (x_i - \bar{x}_w)^2}\right]^{\bar{y}} \quad ...(15)$$

Using (15) in (1), we see that the estimate T reduces to the weighted regression estimate, namely

$$\bar{y}_{wl} = \sum_{i=1}^{k} w_i \left[\frac{1}{w} + \frac{(\bar{x}_N - \bar{x}_w)(\bar{x}_i - \bar{x}_w)}{\sum_{i=1}^{k} w_i (x_i - \bar{x}_w)^2}\right] \bar{y}_{n_i}$$

$$= \bar{y}_w + \hat{\beta}_w (\bar{x}_N - \bar{x}_w) \quad ...(16)$$

where

$$\bar{y}_w = \frac{1}{w}\sum_{i=1}^{k} w_i \bar{y}_{n_i} \quad ...(17)$$

and

$$\hat{\beta}_w = \frac{\sum_{i=1}^{k} w_i \bar{y}_{n_i}(x_i - \bar{x}_w)}{\sum_{i=1}^{k} w_i (x_i - \bar{x}_w)^2} \quad ...(18)$$

The minimum variance is given by equation (3) *i.e.*,

$$V(\bar{y}_{wl} | n_1, n_2, ..., n_k) = \sum_{i=1}^{k} w_i \left[\frac{1}{w} + \frac{(\bar{x}_N - \bar{x}_w)(\bar{x}_i - \bar{x}_w)}{\sum_{i=1}^{k} w_i (x_i - \bar{x}_w)^2}\right]^2$$

$$= \sum_{i=1}^{k} w_i \left[\frac{1}{w^2} + \frac{(\bar{x}_N - \bar{x}_w)^2 (x_i - \bar{x}_w)^2}{\left[\sum w_i (x_i - \bar{x}_w)^2\right]^2} + \frac{2}{w}(\bar{x}_N - \bar{x}_w)\frac{\sum_{i=1}^{k} w_i (x_i - \bar{x}_w)}{\sum_i w_i (x_i - \bar{x}_w)^2}\right]$$

NOTES

But $\sum_{i=1}^{k} w_i(x_i - \bar{x}_w) = 0$ being algebraic sum of the values from their A.M

$$\therefore V(\bar{y}_{wl}|n_1, n_2, \ldots, n_k) = \frac{1}{w} + (\bar{x}_N - \bar{x}_w)^2 \frac{\sum_{i=1}^{k} w_i(x_i - \bar{x}_w)^2}{\left[\sum_i w_i(x_i - \bar{x}_w)^2\right]^2}$$

$$= \frac{1}{w} + \frac{(x_N - \bar{x}_w)^2}{\sum_{i=1}^{k} w_i(x_i - \bar{x}_w)^2} \quad \ldots(*)$$

Particular Cases

Case I. when $V(y_{ij}|x_i) = \gamma x_i$

$$\therefore \quad w_i = \frac{n_i}{\gamma_i} = \frac{n_i}{V(y_{ij}/x_i)} = \frac{n_i}{\gamma x_i} = \frac{w_i'}{\gamma}$$

where $w_i' = \frac{n_i}{x_i}$

and $w = \sum_{i=1}^{k} w_i = \frac{1}{\gamma} \sum_{i=1}^{k} w_i' = \frac{w'}{\gamma}$.

Using these values in (*), we get

$$V(\bar{y}_{wl}|n_1, n_2, \ldots, n_k) = \gamma \left[\frac{1}{w'} + \frac{(\bar{x}_N - \bar{x}_w)^2}{\sum_{i=1}^{k} w_i'(x_i - \bar{x}_w)^2}\right],$$ which depends upon the unknown numbers w_i and the constant γ only.

Case II. $V(y_{ij}|x_i) = \gamma$ ...(1)

In this case $w_i = \frac{n_i}{\gamma}$ and the weighted regression estimate $\bar{y}_{wl}$ reduces to well known simple regression estimate, namely

$$\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$$

where $\hat{\beta} = \frac{\sum_{i=1}^{k} n_i \bar{y}_{n_i}(x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i(x_i - \bar{x}_n)^2} = \frac{s_{xy}}{s_x^2}$ ...(2)

NOTES

Since $E\left[\bar{y}_l | n_1, n_2, \ldots\ldots, n_k\right] = \bar{y}_N$ it follows that

$$V(\bar{y}_l) = E\left[V(\bar{y}_l | n_1, n_2, \ldots\ldots, n_k)\right] + V\left[E(\bar{y}_l | n_1, n_2, \ldots\ldots, n_k)\right]$$

$$= E\left[V(\bar{y}_l | n_1, n_2, \ldots\ldots, n_k)\right]$$

$$= \gamma\left[\frac{1}{n} + E\left\{\frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_{i=1}^{N}(x_i - \bar{x}_n)^2}\right\}\right] \quad \ldots(3)$$

$$((\because \; V(\bar{y}_{wl}) = \frac{1}{\sum w_i} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_i w_i (x_i - \bar{x}_n)^2}$$

$$V(\bar{y}_i | n_1, \ldots, n_k) = \frac{1}{\sum_i \frac{n_i}{\gamma}} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_{i=1}^{k} \frac{n_i}{\gamma}(x_i - \bar{x}_n)^2} \quad \because w_i = \frac{n_i}{\gamma},\ w = \sum_i w_i$$

$$= \gamma\left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right]$$

$$\therefore E\left[v(\bar{y}_l | n_1, n_2, \ldots\ldots, n_k)\right] = \gamma\left[\frac{1}{n} + E\left\{\frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right\}\right]$$

To find the value of $E\left[\frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right]$, we use the method of symmetric functions. It can be shown that to the terms of order $\frac{1}{n^2}$, we have

$$E\left[\frac{n(\bar{x}_N - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\right] = \left(\frac{1}{n} - \frac{1}{N}\right)\left\{1 + \frac{3}{n} - \frac{6}{N} + \frac{\beta_2}{N} + 2\beta_1\left(\frac{1}{n} - \frac{1}{N}\right)\right\} \quad \ldots(4)$$

where $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$ $\quad \ldots(5)$

Also, $\gamma = V(y_{ij} / x_i) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(y_{ij} - E(y_{ij} / x_i)\right)^2$

NOTES

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[y_{ij}-\alpha-\beta x_i\right]^2$$

$$\rho = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[\left(y_{ij}-\bar{y}_N\right)-\beta\left(x_i-\bar{x}_N\right)\right]^2$$

$$= \sigma_y^2\,(1-\rho^2)$$

$y_i = \alpha + \beta x_i$

$\therefore \frac{1}{N}\sum_{i=1}^{n} y_i = \alpha + \beta \sum_{i=1}^{N} x_i$

$\therefore \bar{y}_N = \alpha + \beta \bar{x}_N$

$$((\because r = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(y_{ij}-\bar{y}_N\right)^2 + \beta^2\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(x_i-\bar{x}_N\right)^2$$

$$-2\beta\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(y_{ij}-\bar{y}_N\right)\left(x_i-\bar{x}_N\right)$$

$= \sigma_y^2 + \beta^2\,\sigma_x^2 - 2\,\beta\,cov.\,(y_{ij}, x_i) = \sigma_y^2 + \beta^2\,\sigma_x^2 - 2\beta\,\sigma_{yx}$ $\because N$ is large

$= \sigma_y^2 + \frac{\sigma_{yx}^2}{\sigma_x^4}\cdot\sigma_x^2 - 2\frac{\sigma_{yx}}{\sigma_x^2}\cdot\sigma_{yx}$ $\quad \because \beta = \frac{S_{yx}}{S_x^2} = \frac{\sigma_{yx}}{\sigma_x^2}$.

$= \sigma_y^2 + \rho^2\,\sigma_y^2 - 2\rho^2\,\sigma_y^2$ $\quad \because \rho^2 = \frac{\sigma_{yx}^2}{\sigma_y^2\sigma_x^2}$

$= \sigma_y^2\,(1-\rho^2)$

Thus, we see that assuming N to be large, the variance to the terms of order $\frac{1}{n^2}$ is

$$V(\bar{y}_l) \cong \frac{\sigma_y^2(1-\rho^2)}{n}\left(1+\frac{1}{n}\right)$$

$$V(\bar{y}_l) = \sigma_y^2(1-\rho^2)\left[\frac{1}{n}+\frac{1}{n}\left(\frac{1}{n}-\frac{1}{N}\right)\left\{1+\frac{3}{n}-\frac{6}{N}+\frac{\beta_2}{N}+2\beta_1\left(\frac{1}{n}-\frac{1}{N}\right)\right\}\right]$$

$$= \sigma_y^2(1-\rho^2)\left[\frac{1}{n}+\frac{1}{n^2}\left(1+\frac{3}{n}+2\beta_1\frac{1}{n}\right)\right] \quad \because N \text{ is large}$$

$$= \frac{\sigma_y^2(1-\rho^2)}{n}\left[1+\frac{1}{n}+\frac{3}{n^2}+\frac{2\beta_1}{n^2}\right]$$

$$= \frac{\sigma_y^2(1-\rho^2)}{n}\left(1+\frac{1}{n}\right) \text{ to the terms of order } \frac{1}{n^2}.$$

NOTES

9.6 COMPARISON OF SIMPLE REGRESSION ESTIMATE WITH THE RATIO ESTIMATE AND THE SIMPLE UNBIASED ESTIMATE

The sampling variance of the simple regression estimate of the population mean is given by

$$V(\bar{y}_l) = \frac{N-n}{Nn} S_y^2 \left(1-\rho^2\right) \quad ...(1)$$

The sampling variance of the ratio estimate of the same order of approximation is given by

$$V(\bar{y}_R) = \frac{N-n}{N_n} \left(S_y^2 - 2R_N \rho S_x S_y + R_N^2 S_x^2\right) \quad ...(2)$$

while that of simple unbiased estimate is

$$V(\bar{y}_n) = \frac{N-n}{N_n} S_y^2 \quad ...(3)$$

Comparing (1) and (3), we see that regression estimate is always better or more efficient than the simple unbiased estimate.

Comparing (1) and (2), we observe that simple regression estimate is more efficient than ratio estimate if

$$V(\bar{y}_l) < V(\bar{y}_R)$$

i.e., if $S_y^2 - 2 R_N \rho S_x S_y + R_N^2 S_x^2 > S_y^2 (1 - \rho^2)$

i.e., if $(\rho S_y - R_N S_x)^2 > 0$, which is always true, unless

$$R_N = \rho \frac{S_y}{S_x} \text{ when equality holds.}$$

Hence the regression estimate is always more efficient than the ratio estimate unless the regression of y on x is a straight line passing through the origin, in which case the two estimates will have equal variance.

9.7 COMPARISON OF SIMPLE REGRESSION WITH STRATIFIED SAMPLING

We know that sampling variance of the estimated mean in stratified sampling is given by

$$V(\bar{y}_w) = \sum_{i=1}^{k} p_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} \quad ...(1)$$

where $p_i = \frac{N_i}{N}$

NOTES

The variance (1) is however, not the appropriate variance to compare with the variance of the simple regression estimate. The appropriate variance for comparison would be the variance of the estimate obtained by classifying the selected random sample by the strata and treating it as if it were a stratified sample. This is obtained by

$$V(\bar{y}_w) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i S_i^2 + \frac{1}{n^2}\sum_{i=1}^{k}(1-p_i)S_i^2$$

$\because N$ is large.

$$V(\bar{y}_w) = \frac{1}{n}\sum_{i=1}^{k} p_i \sigma_i^2 + \frac{1}{n^2}\sum_{i=1}^{k}(1-p_i)\sigma_i^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} p_i \sigma_i^2 + \frac{1}{n}\sum_{i=1}^{k}\sigma_i^2 - \frac{1}{n}\sum_{i=1}^{k} p_i \sigma_i^2\right] \quad ...(2)$$

We have seen that the simple regression estimate is optimum when the regression of y on x is linear and $V(y/x)$ is constant. If the σ_i^2 are approximately constant say equal to σ_w^2, then (2) becomes

$$V(\bar{y}_w) = \frac{1}{n}\sigma_w^2\left[1 + \frac{k}{n} - \frac{1}{n}\right] = \frac{\sigma_w^2}{n}\left[1 + \frac{k-1}{n}\right] \quad ...(3)$$

This is the variance which is comparable with the variance of the simple regression estimate to the same order of approximation *i.e.*,

$$V(\bar{y}_l) = \frac{\sigma_y^2}{n}(1-\rho^2)\left[1 + \frac{1}{n}\right] \quad ...(4)$$

Now, since $\sigma_y^2 (1 - \rho^2)$ represents residual variance about the regression straight line and hence it can never be less than σ_w^2. Thus stratified sample is more efficient than simple regression estimate.

9.8 REGRESSION ESTIMATES IN STRATIFIED SAMPLING

We shall consider two difference estimates, the separate difference estimate and the combined difference estimate.

The separate difference estimate is defined as

$$\bar{y}_{ds} = \sum_{i=1}^{k} p_i\left[\bar{y}_{n_i} + \beta_i\left(\bar{x}_{N_i} - \bar{x}_{n_i}\right)\right] \quad ...(1)$$

where $\beta_i = \dfrac{S_{ixy}}{S_{ix}^2}$, $p_i = \dfrac{N_i}{N}$; $i = 1, 2, ..., k$ are assumed to be known.

NOTES

Now, $$E(\bar{y}_{ds}) = \sum_{i=1}^{k} \left[\bar{y}_{N_i} + \beta_i \bar{x}_{N_i} - \beta_i \bar{x}_{N_i}\right]$$

$$= \sum_{i=1}^{k} p_i \, \bar{y}_{N_i} = E\left(\bar{y}_{N_i}\right) = \bar{y}_N$$

Thus $\bar{y}_{ds}$ is an unbiased estimate of the population mean $\bar{y}_N$. Its variance is given by

$$V(\bar{y}_{ds}) = \sum_{i=1}^{k} p_i^{\,2}\left[V\left(\bar{y}_{n_i}\right) + p_i^{\,2}\, V\left(\bar{x}_{n_i}\right) - 2\beta_i \, cov.\left(\bar{y}_{n_i}, \bar{x}_{n_i}\right)\right]$$

$$= \sum_{i=1}^{k} p_i^{\,2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left[S_{iy}^{\,2} + \beta_i^{\,2} S_{ix}^{\,2} - 2\beta_i S_{ixy}\right]$$

$$= \sum_{i=1}^{k} p_i^{\,2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left[S_{iy}^{\,2} - \beta_i^{\,2} S_{ix}^{\,2}\right] \qquad ...(2)$$

$$\left(\left(\because \beta_i = \frac{S_{ixy}}{S_{ix}^{\,2}}\right)\right)$$

Now, we construct a combined difference estimate.

Consider

$$\bar{y}_{dc} = \sum_{i=1}^{k} p_i \bar{y}_{n_i} + \beta\left(\bar{x}_N - \sum_{i=1}^{k} p_i \, \bar{x}_{n_i}\right) \qquad ...(3)$$

Now, $$E(\bar{y}_{dc}) = \sum_{i=1}^{k} p_i E\left(\bar{y}_{n_i}\right) + \beta\bar{x}_N - \beta \sum_{i=1}^{k} p_i E\left(\bar{x}_{n_i}\right)$$

$$= \bar{y}_N + \beta\,\bar{x}_N - \beta \sum_{i=1}^{k} p_i \bar{x}_{N_i}$$

$$= \bar{y}_N + \beta\,\bar{x}_N - \beta\, E\left(\bar{x}_{N_i}\right) = \bar{y}_N + \beta\,\bar{x}_N - \beta\,\bar{x}_N = \bar{y}_N$$

Hence $\bar{y}_{dc}$ is unbiased estimate of $\bar{y}_N$.

To determine the optimum value of β, we minimize the variance of the estimate $\bar{y}_{dc}$ *w.r.t.* β

Now,

$$V(\bar{y}_{dc}) = V\left(\sum_{i=1}^{k} p_i \bar{y}_{n_i}\right) + \beta^2 V\left(\sum_{i=1}^{k} p_i \bar{x}_{n_i}\right) - 2\beta \, cov.\left(\sum_{i=1}^{k} p_i \bar{y}_{n_i}, \sum_{i=1}^{k} p_i \bar{x}_{n_i}\right) \qquad ...(4)$$

NOTES

$$\therefore \quad \frac{d}{d\beta}V(\bar{y}_{dc}) = 0 \Rightarrow \beta = \frac{cov.\left(\sum_{i=1}^{k} p_i \bar{y}_{n_i}, \sum_{i=1}^{k} p_i \bar{x}_{n_i}\right)}{V\left(\sum_{i=1}^{k} p_i \bar{x}_{n_i}\right)} \qquad ...(5)$$

Thus, $$V(\bar{y}_{dc}) = V\left(\sum_{i=1}^{k} p_i \bar{y}_{n_i}\right) - \beta^2 V\left(\sum_{i=1}^{k} p_i \bar{x}_{n_i}\right) \text{ from (4)}$$

$$= \sum_{i=1}^{k} p_i^2 V\left(\bar{y}_{n_i}\right) - \beta^2 \sum_{i=1}^{k} p_i^2 V\left(\bar{x}_{n_i}\right)$$

$$= \sum_{i=1}^{k} p_i^{\,2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left[S_{iy}^{\,2} - \beta^2 S_{ix}^{\,2}\right] \qquad ...(6)$$

Now, from (5)

$$\beta = \frac{\sum_{i=1}^{k} p_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{ixy}}{\sum_{i=1}^{k} p_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{ix}^2} = \frac{\sum_{i=1}^{k} w_i \beta_i}{\sum_{i=1}^{k} w_i} \qquad ...(7)$$

where $$w_i = p_i^{\,2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{ix}^{\,2} \qquad ...(8)$$

Then from (2) and (6), we get

$$V(\bar{y}_{dc}) - V(\bar{y}_{ds}) = \sum_{i=1}^{k} w_i (\beta_i - \beta)^2 \qquad ...(9)$$

which is always *+ve*. Thus unless, the regression coefficient is the same from stratum to stratum, the separate difference estimate will be more efficient than the combined difference estimate.

When the regressions coefficient β_i are not known, we estimate them from the sample and we obtain the separate regression estimate

$$\bar{y}_{ls} = \sum_{i=1}^{k} p_i\left[\bar{y}_{n_i} + \hat{\beta}_i\left(\bar{x}_{N_i} - \bar{x}_{n_i}\right)\right] \qquad ...(10)$$

where $$\hat{\beta}_i = \frac{S_{ixy}}{S_{ix}^{\,2}}.$$

Clearly the estimate $\bar{y}_{ls}$ is biased but the bias vanishes when the sample size within each stratum is sufficiently large.

((∵ In that case $\hat{\beta}_i \to \beta_i$))

Now, variance of the estimate to the 1st degree of approximation is

$$V(\bar{y}_{ls}) \cong \sum_{i=1}^{k} p_i^{\,2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left(S_{iy}^{\,2} + B_i^{\,2} S_{ix}^{\,2} - 2\beta_i S_{ixy}\right)$$

NOTES

((∴ In case of simple regression estimate $\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$, variance is

$$V(\bar{y}_l) \cong \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}\right))$$

$$\therefore \quad V(\bar{y}_{ls}) = \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{iy}^2 \left(1 - \rho_i^2\right) \qquad ...(11)$$

where ρ_i is the coefficient of correlation between y and x for the ith stratum. The estimate of the variance is given by

$$\text{Est. } V(\bar{y}_{ls}) = \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left(s_{iy}^2 + \hat{\beta}_i^2 s_{ix}^2 - 2\hat{\beta}_i s_{ixy}\right)$$

$$= \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left[\left(y_{ij} - \bar{y}_{n_i}\right) - \hat{\beta}_i\left(x_{ij} - \bar{x}_{n_i}\right)\right]^2 \qquad ...(12)$$

In case of the combined difference estimate, when β the pooled regression coefficient is not known, we replace it by its estimate $\hat{\beta}$, giving the combined regression estimate

$$\bar{y}_{lc} = \sum_{i=1}^{k} p_i \bar{y}_{n_i} + \hat{\beta}\left(\bar{x}_N - \sum_{i=1}^{k} p_i \bar{x}_{n_i}\right) \qquad ...(13)$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{ixy}}{\sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{ix}^2} \qquad ...(14)$$

If the n_i are large and proportional to N_i and the finite correction factors can be ignored, $\hat{\beta}$ reduces to the usual pooled estimate of the regression coefficient, namely,

$$\hat{\beta} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{n_i}\right)\left(y_{ij} - \bar{y}_{n_i}\right)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{n_i}\right)^2} \qquad ...(15)$$

It can be shown that the variation of $\bar{y}_{lc}$ to the 1st degree of approximate is

$$V(\bar{y}_{lc}) \cong \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left(S_{iy}^2 + \beta^2 S_{ix}^2 - 2\beta S_{ixy}\right) \qquad ...(16)$$

and its estimate is

$$\text{Est. } V(\bar{y}_{lc}) = \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left[s_{iy}^2 + \hat{\beta}^2 s_{ix}^2 - 2\hat{\beta} s_{ixy}\right]$$

$$= \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \frac{1}{n_i - 1} \sum_{j=1}^{ni} \left[\left(y_{ij} - \bar{y}_{n_i}\right) - \hat{\beta}\left(x_{ij} - \bar{x}_{n_i}\right)\right]^2 \qquad ...(17)$$

NOTES

9.9 ILLUSTRATIVE EXAMPLES

Example 1. *Using the data given in Table I, estimate the total number of milch animals in 117 villages of zone A by the method of regression estimation. Also, compare its precision with the ratio estimate and mean per unit estimate.*

Table I

S. No. of village	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Number of milch animals in survey (y)	*1129*	*1144*	*1125*	*1138*	*1137*	*1127*	*1163*
Number of milch animals in census (x)	*1141*	*1144*	*1127*	*1153*	*1117*	*1140*	*1153*
S. No. of village	*8*	*9*	*10*	*11*	*12*	*13*	*14*
Number of milch animals in survey (y)	*1153*	*1164*	*1130*	*1153*	*1125*	*1116*	*1115*
Number of milch animals in census (x)	*1146*	*1189*	*1137*	*1170*	*1115*	*1130*	*1118*
S.No. of village	*15*	*16*	*17*				
Number of milch animals in survey (y)	*1112*	*1112*	*1123*				
Number of milch animals in ceusus (x)	*1122*	*1113*	*1166*				

Solution.

Here, $N = 117, \quad n = 17, \quad X = 143968$

$$\sum x_i = 19381, \quad \sum y_i = 19266$$

$$\bar{x} = 1140.06, \quad \bar{y} = 1133.29$$

$$s_x^2 = 458.56, \quad s_y^2 = 287.85, \quad s_{xy} = 262.86$$

Hence, $b = s_{xy}/s_x^2 = 0.5732, \quad r = s_{xy}/s_x s_y = 0.7235$

Regression estimate of the total number of milch animals is obtained as

$$\hat{Y}_l = N\bar{y}_l = N\left[\bar{y} + b\left(\bar{X} - \bar{x}\right)\right]$$

$$= 117\ [1130.29 - 0.5732\ (1140.06 - 1230.49)]$$

$$= 139263$$

and an estimate of the variance of $\hat{Y}_l$ is given as

$$v\left(\hat{Y}_l\right) = \frac{(1-f)N^2}{n}\left(1 - r^2\right)s_y^2$$

NOTES

$$= \frac{100 \times 117}{17} \; 150.6757 \; = 1,30,700$$

From Example 6.1,

$$v\left(\hat{Y}_R\right) = 118,890 \quad \text{and} \quad v\left(\hat{Y}\right) = 129,285$$

Hence, the relative precision of the regression estimate over the mean per unit estimate is

$$R.P. = \left\{\frac{v\left(\hat{Y}\right)}{v\left(\hat{Y}_l\right)}\right\} \times 100 = 114.64\%$$

Similarly, the relative precision of the regression estimate over the ratio estimate is given by

$$R.P. = \left\{\frac{v\left(\hat{Y}_R\right)}{v\left(\hat{Y}_l\right)}\right\} \times 100 = 124.67\%$$

Example 2. *Consider the data given in Table. II, estimate the total number of trees in the districts by the regression method of estimation and compare its precision.*

Table II

Stratum number	*Total no. of villages* (N_m)	*Total area (in hect.) under orchard* (X_m)	*No. of villages in sample* (n_m)	*Area under orchards in ha.* (x_m)	*Total no. of trees* (y_m)
1	*985*	*11253*	*6*	*10.63, 9.90, 1.45, 3.38, 5.17, 10.35*	*747, 719, 78, 201, 311, 448*
2	*2196*	*25115*	*8*	*14.66, 2.61, 4.35, 9.87, 2.42, 5.60, 4.70, 36.75*	*580, 103, 316, 739, 196, 235, 212, 1646*
3	*1020*	*18870*	*11*	*11.60, 5.29, 7.94, 7.29, 8.00, 1.20, 11.50, 7.96, 23.15, 1.70, 2.01*	*488, 227, 374, 491, 499, 50, 455, 47, 879, 115, 115*

Estimate the total number of trees in the three districts by various methods and compare their precision. The calculations have been shown in the table given below :

NOTES

Stratum	W_m	$\left(\frac{1}{n_m}-\frac{1}{N_m}\right)$	$\bar{x}_m$	$\bar{y}_m$	$\hat{R}_m$	$W_m \cdot \bar{x}_m$	$W_m \cdot \bar{y}_m$	$s^2_{x_m}$	$s^2_{y_m}$	$s_{x_m y_m}$
1	0.2345	0.16565	6.81	417.33	61.28	1.60	97.86	15.97	74775.47	1007.05
2	0.5227	0.12454	10.12	503.38	49.74	5.29	263.12	132.66	259113.40	5709.16
3	0.2428	0.08992	7.97	340.00	42.66	1.94	82.55	38.44	65885.60	1404.71

Solution.

$$b_1 = 62.93, \quad b_2 = 43.53, \quad b_3 = 36.56$$

(*i*) *Separate Regression Estimate.* The estimate of the total number of trees is given by

$$\hat{Y}_{ls} = \sum_m^k N_m \left[\bar{y}_m + b_m \left(\bar{X}_m - \bar{x}_m\right)\right] = 2{,}672{,}911$$

and
$$v\left(\hat{Y}_{ls}\right) = \sum_m^k N_m^2 \frac{(1-f_m)}{n_m}\left(s^2_{ym} - b_m^2\, s^2_{xm}\right)$$

$$= \sum_m^k N_m^2 \left(\frac{1}{n_m} - \frac{1}{N_m}\right)\left(s^2_{ym} - b_m^2\, s^2_{xm}\right)$$

$$= 1{,}870{,}633{,}332$$

Therefore, efficiency over ratio estimate is given by

$$\left(\frac{2{,}441{,}137{,}835}{1{,}870{,}633{,}332}\right) \times 100 = 130.50$$

(*ii*) *Combined Regression Estimate.* An estimate of the total number of trees is given by

$$\hat{Y}_{lc} = N\,\bar{y}_{lc} = N\left[\bar{y}_{st} + b\left(\bar{X} - \bar{x}_{st}\right)\right] = 2{,}643{,}949$$

and
$$v\left(\hat{Y}_{lc}\right) = \sum_m^k \frac{W_m^2\,(1-f_m)}{n_m\,(n_m-1)} \sum^{n_m} \left[\left(y_{mj} - \bar{y}_m\right) - b_c\left(x_{mj} - \bar{x}_m\right)\right]^2$$

$$= 2{,}020{,}917{,}640$$

Therefore, efficiency over ratio estimate

$$= \frac{6{,}019{,}519{,}627}{2{,}020{,}917{,}640} \times 100 = 297.86$$

Similarly, the efficiency of the separate regression estimate $\left(\hat{Y}_{ls}\right)$ over the combined regression estimate $\left(\hat{Y}_{lc}\right)$ is

$$RE = \frac{2{,}020{,}917{,}640}{1{,}870{,}633{,}332} \times 100 = 108.03$$

NOTES

9.10 SUMMARY

- Simple regression estimate is $\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$.
- $\hat{\beta}$ is consistent $\therefore$ it is unbiased. Even if $\hat{\beta}$ is biased, the bias will be zero for large population.
- Variance of consistent estimate $\hat{\beta}$ is $V(\hat{\beta}) = \dfrac{\sigma_y^2(1-\rho^2)}{\sigma_x^2(n-3)}$.
- (*i*) The population of size N can be regarded as divided into k-classes with the N_i units in the ith class having the value x_i each with $\sum_{i=1}^{K} N_i = N$ and that in repeated samples of size n, drawn from the population with equal probability and without replacement, the number of units having the value x_i is fixed, say n_i; $i = 1, 2, \ldots k$.

  (*ii*) The N_i are sufficiently large so that $(N_i - 1) / N_i \equiv 1$ and $\dfrac{N_i - n_i}{N_i - 1} \equiv 1$

  (*iii*) $E(y_{ij}/x_i) = \alpha + \beta x_i$ and (*iv*) $V(y_{ij}/x_i) = \gamma_i$
- Estimated mean in stratified sampling is $V(\bar{y}_w) = \sum_{i=1}^{k} p_i^2 \dfrac{N_i - n_i}{N_i} \dfrac{S_i^2}{n_i}$.

9.11 GLOSSARY

- **Regression.** A measure of the relation between the mean value of one variable and corresponding values of other variables.
- **Estimate.** Roughly calculate or judge the value.

9.12 REVIEW QUESTIONS

1. An eye estimate of the fruit weights (x_i) on each tree in an orchard having 100 trees was made. The total weight (X) was found to be 12,500 kg. A random sample of 10 trees was taken and the actual weights of fruits (y_i) along with the eye estimates were as below :

Actual weight (y_i)	51	42	46	39	71	61	58	57	58	67
Eye est weight (x_i)	56	47	48	40	78	59	52	58	55	67

NOTES

(*i*) Estimate the total actual fruit weight Y by taking the estimator

$$\hat{Y}_d = N\left[\bar{y} + \left(\bar{X} - \bar{x}\right)\right]$$

and find its sampling variance.

(*ii*) Estimate the total actual fruit weight Y by taking the linear regression estimator

$$\hat{Y}_l = N\left[\bar{y} + b\left(\bar{X} - \bar{x}\right)\right]$$

Also find the sampling variance and compare the results.

2. For estimating the total cattle population, a random sample, wr, of 24 villages was selected from the total 1238 villages. The number of cattle obtained in the survey is given below for each sample village, together with the corresponding census figure relating to a previous period :

S. N. of village	*Number of cattle*		*S. N. of village*	*Number of cattle*		*S. N. of village*	*Number of cattle*	
	Census	*survey*		*Census*	*survey*		*Census*	*survey*
1	623	654	9	161	210	17	330	375
2	690	696	10	298	555	18	218	212
3	534	530	11	2045	2110	19	160	147
4	293	315	12	1069	592	20	210	297
5	69	78	13	706	707	21	262	401
6	842	640	14	1795	1890	22	204	252
7	475	692	15	1406	1123	23	185	199
8	371	292	16	118	115	24	574	564

Compare the efficiency of the regression estimator with the ratio estimator. It is given that the number of cattle for the previous period of 1238 villages is 680,900.

9.13 FURTHER READINGS

- *Sampling Techniques,* William G. Cochran, Wiley India Pvt. Ltd.
- *Elements of Sampling Theory and Methods,* Z. Govindarajulu, Prentice Hall.
- *Probability & Statistics for Scientists and Engineers,* Pearson Education.

❑❑❑

APPENDIX : CASE STUDIES

NOTES

CHAPTER 1

1. Draw a schedule for enquiry into the state of employment, under-employment in the rural sectors of India. Justify the definitions and concepts and also elaborate a set of instructions to the field workers.
2. Plan a sample survey to study the problem of indebtedness among the rural agricultural population in India. Suggest a suitable survey plan on the following points:

   (*i*) Sampling units (*ii*) Sampling frame

   (*iii*) Method of sampling (*iv*) Method of data collection.

   Draft a suitable questionnaire that may be used in this regard.

CHAPTER 2

1. For a realistic comparison between wtr and wr sampling schemes in which the effective sample size or total cort is same is both cases, study the following example:

   Example 1. In order to estimate the mean of a finite population a sample of size x is selected with replacement and the number u of distinct units determined. Let the estimator used be

   $$y' = \frac{1}{x_0} f(u) \frac{s_y}{u}, \quad x_0 = E f(u)$$

   Show that y' is unbaised with variance given by

   $$V(y') = E\left[\frac{f^2(u)/u}{n_0^2} - \frac{1}{N}\right] s^2 + \left(\hat{y}^2 - \frac{S^2}{N}\right) \frac{V f(u)}{n_0^2}$$

   Compare it with the without replacement sampling scheme in which a sample of size $E(u)$ is selected and the estimator used is

   $$\hat{y} = Sy/E(u)$$

   Also show that the variance $v(y)$ is always smaller than $v(y')$ provided $s^2/\hat{y}^2 < N$.

NOTES

Example 2. If sampling with replacement is continued till the sample contains n distinct units, two estimators may be formed, one based on the distinct units only and the other based on all selections. For the comparison of two estimators, study the following excercise:

In order to estimate the mean of a finite population, sampling with replacement with equal probabilities is continued till the sample contains n distinct units. Let v be the total number of selections made, k_r ($\Sigma k_r = v$) being the frequency of appearance of the r^{th} distinct unit in the sample. Defining $\hat{y}_v = Sk_r y_r/v$ and $y_n = Sy_r/n$, prove that

(*i*) $\hat{y}_v$ and y_n are unbiased

(*ii*) $V(y_v) = E\frac{1}{v}\frac{2}{y}$

(*iii*) $E(v) = N\left(\frac{1}{N} + \frac{1}{N-1} + ... + \frac{1}{N-n+1}\right)$

(*iv*) $E\frac{1}{v} > \frac{1}{E(v)} > (N-n)\,[n(N-1)]^{-1}$

Hence or otherwise show that $v(y_v) \geq v(y_n)$.

CHAPTER 3

1. *For an extension of stratification problems to two dimensions, study the following problem:*

   Example 1. There is a symmetric continuous bivariate distribution with frequency function

   $$f(x, y),\ a \leq x \leq b,\ c \leq y \leq d.$$

   The problem is to divide the population into four strata by drawing lines parallel to the axes through the point (x_0, y_0). If sample allocation is proportional, show that the double dichotomy point (x_0, y_0), for which the generalized variance $\begin{vmatrix} V(\bar{x}) & \text{Cov}(\bar{x}, \bar{y}) \\ \text{Cov}(\bar{x}, \bar{y}) & V(\bar{y}) \end{vmatrix}$ is a minimum, is the center of gravity of the distribution.

2. *The study of the following problem shows that moderate departures of the actual allocation from the optimum do not lead to any appreciable increase in the variance:*

Example 2. A population is divided into two strata with $\frac{N_1}{N_2} = d$, $\frac{S_1}{S_2} = d$. Let M denote $[n_1/n_2]/[n_1'/n_2']$, where n_1, n_2 is a general allocation of the total sample size n, and n_1', n_2' is the optimum allocation for purposes of estimating the population mean in stratified simple random sampling. Show that the relative precision α of a general allocation to the optimum allocation is given by

$$\alpha = \mu (\lambda d + 1)^2 (\lambda \mu d + 1)^{-1} (\lambda d + \mu)^{-1}$$

By tabulating α against M for different values of λd, show that the optimum is flat in the sense that there is no appreciable loss of precision for $\frac{1}{2} \leq M \leq 2$.

NOTES

CHAPTER 4

Exmple 1. Let a systematic sample of every hundredth household be taken with random start j between 1 and 100 from a population containing 100 h + k households, where h and k are integers with $0 \leq k \leq 99$. The number of households in the sample will be h, with a hance of $1 - \frac{k}{100}$ and $h + 1$, with a chance of $k/100$. Show that the expected sample size will be $h + k/100$, the variance of sample size being $(1 - k/100)\ k/100$.

If the distribution of k can be taken to be uniform over the range 0 to 99, show that the average variance would be very nearly 1/6.

Example 2. A population contains $N = nk$ units where k is odd. Denote the mean of systematic sample based on the random start is taken between 1 and k by $\bar{y}_i$. The centered systematic sample estimate will be obtained by taking the mean of the central units (numbered $\frac{k+1}{2}$, $k + (k + 1)/2$, ...) from each of n strata formed when a random start systematic sample is selected. If the population is monotonic increasing, the centered systematic sample mean $\bar{y}_c$ will be the median of the k random-start systematic sample means $\bar{y}_1 < \bar{y}_2 < .... < \bar{y}_k$. The mean square

NOTES

error of y_c will be $(\bar{y}_c - \hat{y})^2$ while the variance of y_i would be $\sum(\bar{y}_i - \hat{y})^2/k$. Use the result,

$$(\text{mean} - \text{median})^2 < \text{variance}$$

to prove that centered systematic sampling is more efficient than random start systematic sampling in the case of monotone populations.

CHAPTER 5

1. The study of following exercise will give the optimum size of the cluster with specified cort and variance.

   A population consists of N clusters each containing M elements. A simple random sample of n clusters is selected to estimate the population mean per element. An unbiased estimator would be Sy_i/n with a variance of $S_b^2/(nm)$ where $S_b^2 = \sum M (y_i - \bar{y})^2/(N-1)$, and the finite population correction is ignored. Assuming S_w^2, the variance within clusters to be given by $a\ M^g$, $g > o$, and the cort function to be $c = c_1\ Mn + c_2\ \sqrt{n}$, find the optimum size of the cluster for which the variance of the estimator is a minimum, given the total cort of the survey. Show that the optimum value of M will be smaller if c_1 increases or decreases.

2. A useful way of forming clusters of two elements each is expressed in the following exercise:

   Consider a population of four units with the variate values y_1, y_2, y_3 and y_4 arranged in ascending order of magnitude. If the population is to be divided into clusters of two units each, there will be three ways of forming the clusters, one of the clusters being (y_1, y_2) or (y_1, y_3) or (y_1, y_4) in the three situations. The object is to estimate the population total by selecting one cluster at random. If the absolute errors are denoted by e_1, e_2, e_3, show that $e_3 \leq e_2 \leq e_1$. Hence show that, for loss functions which are monotonically increasing with e, the risk is a minimum for the third method of forming clusters. Generalize this result to populations containing an even number of units by proving that the best way of forming clusters of two units is to take units equidistant from each end.

NOTES

CHAPTER 6

1. The comparison of unstratified pps sampling with stratified simple random sampling is well explained in the following example:

Example. A population is divided into < strate, the size of a unit in the ith stratum being x_i. Assume that

$$y_{i\partial} = \alpha + \beta x_i + e_{i\partial}$$

where $E(e_{ir}/x_i) = 0, \quad V(e_{ir}/x_i) = a\, x_i^g$.

Denoting by V_p, V_o and V_{pps} for variances of stratified proportionate, stratified optimum and unstratified with replacement pps estimates of the population mean, show that

$$E(V_p) = \frac{a}{nN}\sum x_i^g - \frac{a}{N^2}\sum x_i^g$$

$$E(V_o) = \frac{a}{nN^2}\left(\sum x_i^{g/2}\right)^2 - \frac{a}{N^2}\sum x_i^g$$

$$E(V_{pps}) = \frac{a}{nN^2}\left[X\sum x_i^{g-1} - \sum x_i^g\right]\alpha = 0$$

Hence prove that the stratified optimum estimator is superior to the pps estimator. Further prove that the condition for the pps estimator to be superior to stratified proportionate estimator is $\rho(x, x^{g-1}) > 0$, provided that $(n-1)/N$ is negligible relative to unity.

CHAPTER 7

For the understanding of better estimator based on distinct subunits, solve the following exercise:

Example 1. A population contains N psu's (primary sampling units), M_i being the number of subunits in the ith psu. A sample of n psu's is selected with replacement with pubabilities p_i, $\sum_i p_i = 1$. If the ith psu occurs in the sample λ_i times, λ_i independent subsamples of m_i subunits each are selected from it, sampling being wtr simple random for each subsample. Consider the following estimators of the population total Y:

$$\hat{y} = \frac{1}{n}\cdot\sum_{i=1}^{n}\frac{M_i\,\bar{y}_i}{p_i}, \qquad \hat{y}' = \frac{1}{n}\cdot\sum_{i=1}^{N}\frac{\lambda_i \cdot M_i \cdot \bar{y}_i^1}{p_i}$$

NOTES

In the first case $\bar{y}_i$ is the mean per subunit based on a sample of m_i subunits. In the latter case the mean $\bar{y}_i^1$ is based on the distinct subunits in the sample of m_i λ_i subunits. Show that $\hat{y}'$ gives a smaller variance than $\hat{y}$.

Example 2. In a multistage design the psu's are selected without replacement and sampling is done independently in the psu's. Let $E(z_i/i) = Z_i$, where z_i is an estimate made from the subsample drawn in the ith psu. Then show that $t = s\, z_i$ is an unbiased estimate of $\sum_i \pi_i Z_i$ with a variance of

$$V(t) = V\left(\hat{Z}\right) + \sum_i \pi_i \sigma_i^2 \text{ where}$$

$$\hat{Z} = S\, Z_i, \quad \sigma_i^2 = E\left[(z_i - Z_i)^2/i\right]$$

And an unbiased estimate of $V(t)$ is provided by

$$\hat{V}(t) = \left[\dot{v}\left(\hat{Z}\right)\right]_{Z_i - z_i} + S\pi_i\, \hat{\sigma}_i^2$$

Where $v\left(\hat{Z}\right)$ is an unbiased estimated of $V\left(\hat{Z}\right)$ in one-stage sampling and $\hat{\sigma}_i^2$ is an unbiased estimate of σ_i^2 based on sampling at the second and subsequent stages.

CHAPTER 8

1. The sufficient condition that the first approximation to the true variance be an understatement is referred in the following example:

Example 1. In wtr simple random sampling it is usual to use $\bar{y}/\bar{x}$ as an estimate of $R = Y/X$. Obtaining the exact expression for the variance of $\bar{y}/\bar{x}$, show that a sufficient condition for the usual approximation $V(\bar{y} - R\,\bar{x})/\bar{x}^2$ to be an understatement for $V(\bar{y}/\bar{x})$ is that $\rho\left[\frac{1}{x}, (\bar{y} - R\,\bar{x})^2\right] \geq 0$ where ρ stands for the correlation coefficient.

NOTES

Example 2. The following exercise gives an alternative method of reducing the bias of the ratio estimator:

For estimating the population ratio $R = E(y)/E(x)$, a random sample of size n is split into two halves to give $r_1 = y_1/x_1$ and $r_2 = y_2/x_2$, while the estimator based on the complete sample is $r = y/x = \frac{1}{2}(y_1 + y_2) \Big/ \left[\frac{1}{2}(x_1 + x_2)\right]$. (The y_i and x_i are the actual sample means).

We shall assume that the x_i are normally distributed with variance $2h$, which $0(n^{-1})$ and that

$$y_i = a + bx_i + u_i \text{ where } E(u_i/x_i) = 0,\ E(u_i^2/x_i) = 2\delta.$$

Then $y = a + bx + u$, where $u = (u_1 + u_2)/2$.

We shall choose the units of measurement such that $E(x) = 1$ and let $x = 1 - z$. Neglecting terms of order x^{-4} or lower, show that

$$E(x^{-1}) = 1 + h + 3h^2 + 15h^3,$$

$$E(x^{-2}) = 1 + 3h + 15h^2 + 105h^3.$$

Use these results to prove that the bias of r for estimating R is $a(h + 3h^2 + 15h^3)$, which is $0(n^{-1})$ and

$$V(r) = a^2(h + 8h^2 + 68h^3) + 8(1 + 3h + 15h^2 + 105h^3).$$

On the other hand, if the estimator

$t = 2r - \frac{1}{2}(r_1 + r_2)$ is used, then its bias is $a(6h^2 + 90h^3)$, which is $0(n^{-2})$.

Further

$V(t) = a^2(h + 4h^2 + 12h^3) + \delta(1 + 2h + 8h^2 + 108h^3)$, which is smaller than $V(r)$.

CHAPTER 9

1. The following excercise studies that when regression of y on x is linear, then the usual ratio estimator is better than the ratio type estimator.

Example 1. If the regression of y on x is linear, that is, $E(y/x) = ax + b$, show that in simple random sampling the estimator $\bar{y}/\bar{x}$ will give a smaller large sample variance than the ratio-type estimator given by

$$\bar{r} + (N-1)\,n\,(\bar{y} - \bar{r}\,\bar{x})\left[N(n-1)\,\bar{x}\right]^{-1}$$

where $\bar{r} = n^{-1} S(y_i/x_i)$

NOTES

Example 2. The following exercise gives an alternative derivation of the variance of the regression estimator:

Define the difference estimator as

$$t_1 = \bar{y} + P\left(\bar{X} - \bar{x}\right) \text{ and the regression estimator as}$$

$$t_2 = t_1 + (b - \beta)\left(\bar{X} - \bar{x}\right) \text{ and hence}$$

$$MSE\,(t_2) = V\,(t_1) + E\,(b - \beta)^2\left(\bar{X} - \bar{x}\right)^2 +$$

$$2E\left[(b - \beta)\left(\bar{X} - \bar{x}\right)(t_1 - \bar{y})\right].$$

Show that

$$E\left[(b - \beta)^2\left(\bar{X} - \bar{x}\right)^2\right] = 0\,(n^{-2}) \text{ and}$$

$E\left[(b - \beta)\left(\bar{X} - \bar{x}\right)(t_1 - \bar{y})\right]$ is of order lower than n^{-1}, whereas $V\,(t_1)$ is of order n^{-1}. Hence prove that for sufficiently large n,

$$MSE\,(t_2) = V\,(t_1) = \sigma_y^2\,(1 - \rho^2)/n.$$

❑❑❑